Cryptic genetic structure and copy-number variation in the ubiquitous forest symbiotic fungus *Cenococcum geophilum*

Authors:

Benjamin Dauphin[†], Maíra de Freitas Pereira[†,¶], Annegret Kohler[¶], Igor V. Grigoriev[§,#], Kerrie Barry[#], Hyunsoo Na[#], Mojgan Amirebrahimi[#], Anna Lipzen[#], Francis Martin[¶], Martina Peter[†,*], Daniel Croll[‡,*]

Affiliations:

[†]Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

[¶]INRAE, UMR 1136 INRAE-University of Lorraine, Interactions Arbres/Microorganismes, Laboratory of Excellence ARBRE, INRAE-Grand Est, Champenoux, France

[§]Department of Plant and Microbial Biology, University of California, Berkeley, USA

[#]U.S. Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, USA

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

[*]Corresponding authors: daniel.croll@unine.ch, martina.peter@wsl.ch

[*]Co-last authors contributed equally to this study

ORCID:

24 B.D. 0000-0003-0982-4252; M.F. 0000-0002-7659-957X; A.K. 0000-0002-9575-9567; I.G.
25 0000-0002-3136-8903; K.B. 0000-0002-8999-6785; C.V. 0000-0001-9778-2659; F.M. 0000-
26 0002-4737-3715; M.P. 0000-0002-6365-6889; D.C. 0000-0002-2072-380X
27
28 Running title: Comparative genomics of *Cenococcum geophilum*

Originality-Significance Statement

We provide the first report on the genetic structure and copy-number variation of the globally ubiquitous and key forest symbiotic fungus *Cenococcum geophilum* using whole-genome sequencing data. We found divergent lineages within sampling sites, while closely related lineages appear over large geographic distances on a continental scale. Even though no sexual spore forming structures have been reported to date, we provide evidence of recombination in a specific lineage suggesting mating activity. Our findings help explain the high genetic diversity occurring within populations and their resilience to changing and adverse environmental conditions. Furthermore, we identify a single *MAT1* idiomorph per genome, confirming heterothallism, and discover that major genomic rearrangements are found in their flanking regions based on chromosomal synteny analysis. Intriguingly, a pseudogene of the opposite functional idiomorph has been characterised in each genome, suggesting a common homothallic ancestor to the species. As *Cenococcum geophilum* is a pivotal mycorrhizal associate of a broad range of trees and shrubs providing nutrition and water supply to their hosts, we highlight and discuss the potential role of the large genome-wide structural variations in environmental selection.

Summary

Ectomycorrhizal (ECM) fungi associated with plants constitute one of the most successful symbiotic interactions in forest ecosystems. ECM support trophic exchanges with host plants and are important factors for the survival and stress resilience of trees. However, ECM clades often harbour morpho-species and cryptic lineages, with weak morphological differentiation. How this relates to intraspecific genome variability and ecological functioning is poorly known. Here, we analysed 16 European isolates of the ascomycete *Cenococcum geophilum*, an extremely ubiquitous forest symbiotic fungus with no known sexual or asexual spore forming structures but with a massively enlarged genome. We carried out whole-genome sequencing to

identify single-nucleotide polymorphisms. We found no geographic structure at the European scale but divergent lineages within sampling sites. Evidence for recombination was restricted to specific cryptic lineages. Lineage differentiation was supported by extensive copy-number variation. Finally, we confirmed heterothallism with a single *MAT1* idiomorph per genome. Synteny analyses of the *MAT1* locus revealed substantial rearrangements and a pseudogene of the opposite *MAT1* idiomorph. Our study provides the first evidence for substantial genome-wide structural variation, lineage-specific recombination and low continent-wide genetic differentiation in *C. geophilum.* Our study provides a foundation for targeted analyses of intra-specific functional variation in this major symbiosis.

Keywords: Ascomycete, copy-number variation, Dothideomycetes, ectomycorrhizal fungi, genomics, phylogenetics, recombination, symbiosis

66 **Introduction**

67 Ectomycorrhizal (ECM) fungi associated with plants constitute one of the most successful
68 symbiotic interactions in forest ecosystems (Martin *et al.*, 2016). Abundant ECM partners
69 support the nutrition of many host tree species by facilitating the acquisition of essential
70 nutrients, which is a key determinant of plant survival and productivity under abiotic stress or
71 changing environmental conditions. With exceptional species richness (~20,000; Martin *et al.*,
72 2016), ECM fungi carry high functional diversity among species that is a key factor of
73 ecosystem resilience (Larsen *et al.*, 2016). The environment, through biotic and abiotic factors,
74 plays a major role in shaping ECM communities and their host range. Environmental variation
75 supports ECM species diversity at both small (e.g. host plants) and large spatial scales, and
76 underlies phenotypic plasticity of functional importance for host trees (van der Linde *et al.*,
77 2018). Although intraspecific diversity of important symbiotic traits was highlighted in a
78 handful of studies (Cairney, 1999; Rineau & Courty, 2011), the extent to which such variation
79 is genetically driven and contributes to functional diversity within ECM communities remains
80 largely unknown (Johnson *et al.*, 2012).

81 *Cenococcum geophilum* Fr. (Dothideomycetes, Ascomycota; Spatafora *et al.*, 2012) is
82 one of the most ubiquitous and abundant ECM fungi in the world (LoBuglio, 1999). *C.*
83 *geophilum* colonises a large variety of hosts and occupies a wide ecological niche, from alpine
84 to tropical regions (Trappe, 1962). The species tolerates adverse environmental conditions, as
85 it is commonly found in soils with limited water availability, contaminated by heavy metals,
86 or in sites with stressful climatic conditions such as at the timberline (LoBuglio, 1999;
87 Hasselquist *et al.*, 2005; Gonçalves *et al.*, 2009; Herzog *et al.*, 2013). Increased drought
88 tolerance and resistance to desiccation have been shown for *C. geophilum* mycelia and ECM
89 root-tips as compared to other ECM species (Coleman *et al.*, 1989; di Pietro *et al.*, 2007;
90 Fernandez *et al.*, 2013). These studies also revealed high intraspecific variability and

91 demonstrated that melanin content is one of the key functional traits conferring an advantage
92 under water stress. In nature, *C. geophilum* isolates have shown patterns of local adaptation to
93 serpentine soils with a significant effect of nickel concentrations on fitness-related traits
94 (Gonçalves *et al.*, 2009; Bazzicalupo *et al.*, 2020). Collectively, these results suggest that this
95 cosmopolitan ECM fungus has undergone selection pressures imposed by various
96 environmental factors.

97 *Cenococcum geophilum* is typically recognised based on morphology, i.e., a morpho-
98 species. It forms black melanised mycorrhizae with thick, darkly pigmented hyphae spreading
99 from the root tips into the surrounding soil. Although rare in other ECM species (Smith *et al.*,
100 2015), *C. geophilum* produces sclerotia that act as a "sclerotial bank" in the soil, which are
101 resistant propagules that are melanised and can withstand desiccation for several decades and
102 sequester carbon as dead sclerotia for centuries (Watanabe *et al.*, 2007; Obase *et al.*, 2014,
103 2018). Although being common, the biology and life cycle remains poorly understood.
104 Phylogenetic analyses have revealed multiple divergent *C. geophilum* lineages on small spatial
105 scales, even at the level of soil core samples (Douhan and Rizzo, 2005; Obase *et al.*, 2016).
106 Combined with the high genetic diversity within populations (ECM root tips and sclerotia),
107 this suggests that *C. geophilum* has undergone cryptic speciation (Obase *et al.*, 2017; Vélez *et*
108 *al.*, 2021).

109 Our understanding of *C. geophilum* reproduction is limited. Whilst no spore-producing
110 structures were observed, indicating strictly asexual reproduction through vegetative mycelial
111 propagation, including germination of sclerotia, population genetic studies suggested
112 recombination events that would have involved sexual mating (LoBuglio and Taylor, 2002;
113 Douhan *et al.*, 2007). However, mating may well be restricted to certain environmental
114 conditions or to a few specific lineages with distinct gene pathways involved in sexuality, as
115 suggested in arbuscular mycorrhizal fungi (Riley and Corradi, 2013). This is supported by

116 phylogenetic studies that showed little gene flow within *C. geophilum* populations. Subtle
117 geographic structure with long-distance disjunction suggests (Obase *et al.*, 2016) a complex
118 alternation of sexual and asexual reproduction over space and time (Obase *et al.*, 2017).
119 Besides further sampling efforts, analyses at a whole genome level are essential to determine
120 the prevalence of outcrossing and to better understand patterns of cryptic speciation.

121 The genome of *C. geophilum,* the only known ECM member within Dothideomycetes,
122 is uniquely large for this class (Peter *et al.*, 2016). A high-quality reference genome (strain
123 1.58) revealed that about 75% of the genome is composed of transposable elements (TEs) and
124 this high proportion is unevenly distributed in the genome, with TE-rich and TE-poor regions
125 (de Freitas Pereira *et al.*, 2018). Its closest sequenced relatives within Gloniaceae,
126 Mytilinidiales, the saprotrophic species *Glonium stellatum* and *Lepidopterella palustris,* for
127 example, showed a fourfold smaller genome size (Peter *et al.*, 2016). TE proliferation seems
128 to be a hallmark for the transition from saprotrophy to symbiosis and in fact, *C. geophilum* is
129 the ECM species with the highest proportion of TEs known to date among mycorrhizal fungi
130 with one of the largest genomes, i.e., 178 Mbp; Miyauchi *et al.*, 2020). Nevertheless, the gene
131 content (14,748) is similar to the average number of genes found in fungi (Miyauchi *et al.*,
132 2020). Intraspecific genome size variation suggests important structural rearrangements or
133 gene duplications and deletions (Bourne *et al.*, 2014). Investigation of the reference genome
134 sequence highlighted the presence of sex-related genes and an intact *MAT1* locus containing
135 one of the two idiomorphs (here *MAT1-1*) typical for a heterothallic bipolar mating system of
136 ascomycetes (Peter *et al.*, 2016). The mating-type locus carrying the opposite idiomorph, i.e.,
137 *MAT1-2,* has not been analysed in *C. geophilum* so far.

138 In this study, we used whole-genome sequencing to determine the population genetic
139 structure of 16 *C. geophilum* isolates across Europe and test for recombination events. We
140 reconstructed genome-wide phylogenetic relationships within the genus and characterised gene

141 duplication and deletion events to investigate structural patterns of genetic diversity. Then, we
142 analysed patterns of selection at the gene and genome level using nucleotide diversity, Tajima's
143 *D*, and the ratio of non-synonymous to synonymous polymorphisms. Lastly, we analysed the
144 genetic architecture of the mating type locus and its recent sequence evolution.
145

146 **Results**

147 *Read mapping, variant calling, and polymorphism filtering*

148 Illumina sequencing of 16 *C. geophilum* strains resulted in 348.9 million read pairs (Fig. 1A;
149 Table S1). Of those, 293.6 million read pairs (84.16%) were successfully mapped on the
150 reference strain 1.58. The highest and lowest number of raw read pairs was available for the
151 strains 1.101 and 1.094, respectively, and the highest (95.54%) and lowest (64.37%) overall
152 alignment with the reference strain was found in strains 1.061 and 1.096, respectively (Table
153 S1). An average of 11.3 million read pairs were used for mapping (Table S1). We used the
154 Genome Analysis Toolkit (GATK) to detect genetic variants from the 16 sequenced strains and
155 found, after filtering, a total of 345,358 single nucleotide polymorphisms (SNPs), including
156 284,479 bi-allelic SNPs, of which 67,186 were synonymous variant sites based on the SnpEff
157 analysis.
158 [PASTE TABLE 1]

159 *Phylogenetic tree reconstructions*

160 We inferred phylogenetic trees based on the GAPDH and ITS loci using maximum-likelihood
161 (ML) analyses to identify the main lineages of *C. geophilum* strains. Our two-locus nuclear
162 dataset contained alignment lengths of 597 and 949 bp for GAPDH and ITS, respectively, with
163 a total of 1,546 aligned bp for the 40 strains (including outgroups; Table S2). The GAPDH

164 locus had the most variable sites. The inferred GAPDH and ITS ML tree topologies revealed
165 several incongruencies, but the concatenated partitions generated a fairly well-resolved
166 phylogenetic tree, with a majority of branches with BS values > 80; Figs. 1B, S1–2). With
167 minor exceptions, we found strong support for several monophyletic clades but those were not
168 congruent with the geographical origin of the samples (Fig. 1B). Overall, the clades that
169 emerged from the sequenced European strains were depicted in the worldwide clades 5 and 6
170 described by Obase et al. (2016).

171 To improve resolution, we used genome-wide data and found highly congruent tree
172 topologies compared to the GAPDH and ITS phylogeny. The phylogenomic tree was robust to
173 the use of two near neutral datasets of either synonymous SNPs or randomly-selected genome-
174 wide SNPs (Fig. 1C). Both genome-based ML trees were well-resolved with most of branches
175 being strongly supported (BS > 90; Fig. 1C). This indicates that neutral divergence in the
176 coding regions matches the phylogenetic relationships given by the divergence in non-coding
177 or coding regions.

178 [PASTE FIGURE 1]

179 *Recombination events and genetic structure*

180 We analysed patterns of recombination using a Neighbour-Net network in SplitsTree and found
181 evidence for significant events among the genomes of *C. geophilum* strains (Fig. 2A). We
182 identified two major genetic groups connected by two strains (1.005 and 1.076). The Φ-test
183 supported recombination across the genome when all *C. geophilum* were included ($P < 0.001$;
184 Table S3). However, applying the Φ test to only the cluster not containing the reference genome
185 (non-ref. cluster) yielded no significant evidence for recombination ($P = 0.641$), suggesting
186 one clonal and one recombinant genetic group.

187 As for the phylogenetic analyses, we found no clear geographic structure based on
188 either principal components (PC) or Bayesian clustering analysis. For example, isolates from

189 the Pfynwald (CHE) sampling site (1.057, 1.58, 1.060, 1.061, and 1.064) clustered into the two
190 major groups, reporting a new case of highly divergent lineages present in the same sampling
191 site (Fig. 2A). In addition, isolates from the Blizyn (POL) sampling site (1.094, 1.096, and
192 1.097) also showed divergent lineages, but these were only represented in the clonal group with
193 both mating types (Fig. 2A). No fungal lineages exclusive to certain host tree species were
194 observed in this dataset. The reference genome 1.58 appeared at the edge of the recombinant
195 genetic group. The first axes of the PCA captured 61.9% and 9.6% of the genetic variance (PC1
196 and PC2, respectively; Fig. 2B). In total, the first three axes explained 76.8% of the overall
197 genotypic variation (Fig. S3). Consistent with the Neighbour-Net network, PC1 and PC2
198 separated three to four genetic groups with substantial differentiation between the Pfynwald
199 strains. The intermediate genetic group consisting of the 1.005 and 1.076 strains also clustered
200 together independently. In addition, Bayesian clustering analyses were congruent between
201 replicates (Fig. S4) and we found the optimal number of clusters to be $K = 2$ (Fig. 2C). At $K =$
202 2, two major groups were clearly identified and the intermediate group (strains 1.005 and
203 1.076), displayed in the PCA and network analysis, belonged to the Pfynwald group containing
204 the sequenced strain 1.58. At $K = 3$, the intermediate group still exhibited patterns of admixture
205 between the two major groups, and two Polish genotypes (strains 1.094 and 1.096 belonging
206 to the phylogenetic Clade 6 in Obase *et al.*, 2016) indicate contributions from a third genetic
207 group. At higher K values, no significant changes were observed in genotype ancestry
208 memberships.
209 [PASTE FIGURE 2]

210 *Detection of copy-number variations*

211 We conducted copy-number variation (CNV) analyses comparing read depth of *C. geophilum*
212 strain sequencing sets against the reference genome. We improved the robustness of CNV calls
213 by removing any CNV calls in regions where self-alignments of strain 1.58 reads against the

214 1.58 reference genome produced candidate CNVs. Overall, we found that gene duplications
215 occurred ~3.5 fold more frequently than deletions (Fig. 3A). This predominance of gene
216 duplications was only detected at the highest rates of overlap between genes and CNV regions,
217 i.e., >95% similarity. In total, we identified 217 and 72 gene duplications and deletions,
218 respectively, using a threshold of >50% similarity. By inferring the overall tree topology of
219 based on CNV genotypes, we found patterns consistent with the outcomes of the genetic
220 clustering and phylogenetic analyses (Fig. 3B). First, five strains (1.022, 1.094, 1.096, 1.097,
221 and 1.101) formed a monophyletic clade that was paraphyletic in either the two-locus or
222 genome-based phylogenies (Figs. 1C, 3B). Second, we identified a monophyletic clade with
223 six strains (1.060, 1.075, 1.100, 1.102, 1.104, and 1.58) that constituted a paraphyletic clade in
224 the previous analyses. Furthermore, we tested whether the distance between genes affected by
225 CNVs and the closest TEs are smaller than those for all genes as indicated for other species
226 (e.g. Hartmann and Croll, 2017). Interestingly, CNV genes were at a slightly larger distance
227 compared to the background (P = 0.028, 719 and 448 median base pairs, respectively; Fig. 4).
228 [PASTE FIGURE 3]
229 [PASTE FIGURE 4]

230 *Within-species genetic diversity and signatures of selection*

231 We assessed within-species genetic diversity across the 16 *C. geophilum* genomes and
232 identified genomic regions with elevated diversity rates along the two longest scaffolds, i.e.,
233 #1 and #2; Fig. 5A,C, on which the density of coding sequences and transposable elements
234 (TEs) were negatively correlated (Fig. 5B,D). At the gene level, we estimated the nucleotide
235 diversity per site (π) to be 0.000–0.020, 0.000–0.028, and 0.000–0.013 for the "all", "ref.
236 cluster", and "non-ref. cluster" datasets, respectively (Fig. 6A). The genome-wide average π
237 value was 3.31×10^{-3}, 1.31×10^{-3}, and 1.01×10^{-3}, respectively (Fig. 6A). These results support
238 that the ref. cluster is genetically more diverse than the non-ref. cluster, although the sample

239 size is nearly double in the latter. At the gene-level, we calculated Tajima's *D* values ranging
240 from -1.90–2.86, -0.87–2.32, and -1.97–2.42 for the "all", "ref. cluster", and "non-ref. cluster"
241 dataset, respectively (Fig. 6B). At the genome-wide level, the average Tajima's *D* estimates
242 were 0.14 (ref. cluster) and 0.31 (non-ref. cluster), with a value of 1.64 for the "all" dataset
243 (Fig. 6B). This strongly skewed Tajima's *D* distribution when all strains are included towards
244 positive values indicates strong population substructure (Fig. 6B). However, the "ref. cluster"
245 dataset showed a non-normal distribution of Tajima's *D* with an excess of negative values,
246 which is indicative of a population expansion. The non-ref. cluster showed a distribution
247 largely centred around 0 under neutrality and constant population size. A slight deviation
248 towards positive values suggests a historic decrease in population size or some population
249 substructure. The distribution of the *p*N/*p*S ratio assessed from the all dataset suggests
250 purifying selection (Fig. 6C).
251 [PASTE FIGURE 5]
252 [PASTE TABLE 2]
253 Then, we characterised the top-candidate genes putatively under selection by
254 identifying those with the lowest Tajima's *D* estimates in either the "ref. cluster" or "non-ref.
255 cluster" (see Experimental procedures). Based on functional annotation, we found that those
256 top-candidate genes differed to some extent between the ref. and non-ref. clusters (Table 2). In
257 the recombining "ref. cluster", increased numbers of genes were detected that are putatively
258 involved in self/non-self recognition containing HET, NACHT, WD40- and Ankyrin-repeat
259 domains, as well as genes that are members of multigene-families that are expanded in *C.*
260 *geophilum* or unique in this species and its closest relatives *G. stellatum* and *L. palustris* as
261 compared to other Dothideomycetes (Peter *et al.*, 2016). In the "non-ref. cluster", we found
262 increased number of genes involved in the secondary metabolism (SM; e.g. Lofgren *et al.*,
263 2021) including two genes found in SM clusters for prenyltransferase synthesis (DMATs;

264 Khaldi *et al.*, 2010). In both ref. and non-ref. clusters, we found one gene coding for a
265 carbohydrate-active enzyme (CAZyme; Cantarel *et al.*, 2009). In the "ref. cluster", it codes for
266 a secreted β-galactosidase (Zeiner *et al.*, 2016), a member of the glycoside hydrolyse family
267 35 (GH35) implicated in hemicellulose degradation (van den Brink and de Vries, 2011) that is
268 contracted in the *C. geophilum* genome as compared to other Dothideomycetes. In the "non-
269 ref. cluster", it is a member of the lytic polysaccharide monooxigenases (LPMO) family A11
270 that are almost exclusively present in ascomycetes and the one functionally characterised so
271 far involved in chitin degradation (Hemsworth *et al.*, 2014). Other genes found in both groups
272 encode for proteins involved in information storage and processing as well as cellular processes
273 and signalling (Table S4). All these genes were expressed in free-living mycelium and/or
274 functioning ECM root-tips, but most of them not differentially regulated in these tissues and if
275 so, then mostly down-regulated in ECMs (Table S4).
276 [PASTE FIGURE 6]

277 *Chromosomal synteny analysis of the MAT1-1 and MAT1-2 idiomorphs*

278 Our *de novo* draft genome assemblies of other strains than 1.58 generated 151,48 million base
279 pairs (Mbp) on average, with a mean number of 26,963 contigs per assembly and a mean largest
280 contig size of 165,938 bp. The value range for N50 and L50 were 5,720–24,382 and 2,247–
281 4,759, respectively (Table S5). The GC content varied from 33 to 40%, with an average of
282 37%. Mismatches were lowest and highest with 61 and 260 per 100 kbp for the 1.101 and 1.094
283 strains, respectively.

284 We identified two idiomorphs at the *MAT1* locus in *C. geophilum* isolates and found
285 genetic rearrangements in its flanking coding and non-coding regions based on the synteny
286 analysis (Fig. 7). A single functional *MAT1* idiomorph was detected in each strain, supporting
287 the heterothallism of the species (Figs. 1, 7; Table 1). In both *MAT1* loci types, we found a
288 pseudogene of the opposite *MAT1* idiomorph close to the functional one showing conserved

289 strand orientation as compared to strains carrying the functional opposite idiomorphs with
290 unchanged relative position of their open reading frames (ORF). The same was observed for
291 the closely related *G. stellatum (MAT1-2*: Glost2|444166), whereas *L. palustris*, a homothallic
292 species and also member of the Mytilinidiales, carries both functional *MAT1* idiomorphs with
293 conserved orientation but rearranged position (*MAT1-1*: Leppa1|291743; *MAT1-2*:
294 Leppa1|291655). The *MAT1-1* and *MAT1-2* idiomorphs were surrounded by a conserved gene
295 encoding a MFS transporter (555485) and a gene coding for a protein (72340) of unknown
296 function, both of which showing evidence for an insertion polymorphism generating a non-
297 syntenic region among the *MAT1-1* and *MAT1-2* strains. Furthermore, the flanking *APN2* gene
298 (686539) coding for a putative DNA lyase, and the conserved protein TEX2 (634502) coding
299 for a putative PH domain-containing integral membrane protein were also present in the *MAT1*
300 loci of the closely related *G. stellatum* and *L. palustris* and are often found adjacent to *MAT1*
301 idiomorphs in ascomycetes (Debuchy *et al.*, 2014; Wang *et al.*, 2016; APN2: Glost2|386243,
302 Leppa1|414878; TEX2: Glost2|444168, Leppa1|322867). We further found a gene (555074)
303 encoding a cysteine-type peptidase and a protein of unknown function (72410).
304 [PASTE FIGURE 7]

305 **Discussion**

306 *Cryptic population genetic structure and evidence for recombination*

307 Our genomic analyses of European *C. geophilum* strains, which showed that divergent lineages
308 co-occur in single sampling sites without major geographic structuring, support previous
309 reports from North America that were based on a restricted number of markers (Douhan and
310 Rizzo, 2005; Vélez *et al.*, 2021). Interestingly, our network analysis indicated that the two main
311 genetic groups had possibly different levels of recombination, with most of the sampled strains
312 being clustered into the clonal group with little gene flow. It is unclear whether this restricted
313 gene flow is due to incompatible mating systems or clonal proliferation. Since neither sexual
314 nor asexual spore forming structures have been reported for *C. geophilum*, this suggests a
315 vegetative mycelial propagation in the soil with limited migration ability of the species. In the
316 recombinant group, cryptic sexual events must have occurred at some point in the near or
317 distant past. In case of a asexually reproducing population, i.e., mating incompatibility, low
318 genetic similarities are expected between strains sampled in the same site, which could explain
319 the observed pattern within the Pfynwald site, where one strain (1.060) clearly differs and did
320 not group in the recombinant cluster (Fig. 2A). Our synteny analyses of the *MAT1* loci showed
321 important rearrangements in close vicinity to the *MAT1* genes, but whether this affects mating
322 compatibility is unclear. Apart from mating incompatibility or sexual versus vegetative
323 propagation as well as isolation by geographic distance, other evolutionary processes could
324 explain the current genetic structure, such as demographic history, past fluctuations in effective
325 population size (e.g. population expansion or contraction), or processes of adaptation to
326 environmental conditions during glacial and interglacial periods, as reported in other ECM
327 fungi (Geml *et al.*, 2006; Sánchez-Ramírez *et al.*, 2015).

328 Alternatively, recent adaptation of *C. geophilum* to new hosts or environmental
329 conditions may have led to the emergence of cryptic lineages with strong genetic structure over
330 short geographical distances. While *C. geophilum* is known to have a broad host tree spectrum,
331 some lineages may nevertheless show host preferences, as reported in other ECM species
332 (Rochet *et al.*, 2011; Lofgren *et al.*, 2021). In this respect, the tree-fungus interaction might
333 have shaped differentiation and genetic isolation of these cryptic lineages, in combination with
334 recent selection. The present data do not support such a mechanism because no lineage
335 grouping according to the dominant host tree species was observed. Also, plant-fungus
336 synthesis experiments showed that functional ectomycorrhizas are formed with diverse tree
337 species, both broadleaves and conifers, for the reference strain 1.58 but also for others (e.g. de
338 Freitas Pereira *et al.*, 2018). This shows that the symbiosis can be established with diverse
339 hosts, even if there are possibly host preferences under natural conditions. To shed light on
340 host specificity as a potential driver of genetic structuring within and among sites, detailed
341 information on *C. geophilum* populations associated with roots of different host tree species is
342 needed. Regarding the environment, adaptation to soil characteristics may prevent wide
343 distribution, foster differentiation through geographical isolation and, ultimately, lead to
344 independent and clonal lineages. Note that edaphic specialisation can be coupled with, or even
345 supplanted by, the competition avoidance strategy of ECM species within soil communities
346 (Mujic *et al.*, 2016). Clearly, our dataset with 16 strains is limited and further population
347 genomic studies, with more isolates per populations, are required to be conclusive about the
348 prevalence of sexual reproduction under varying ecological conditions throughout the species
349 range.

350 *Massive gene duplications and functional regions under positive selection*

351 Our CNV analysis revealed broad genomic alterations among *C. geophilum* lineages with about
352 3.5-fold more gene duplications than deletions. Interestingly, the genome-wide pattern of CNV

353 genotypes is congruent with the tree topology from phylogenetic reconstructions suggesting
354 that CNVs have accumulated gradually since the origin of the species. However, the genomic
355 distances between the CNV genes and their closest TEs were in the same range as those for all
356 genes. This is unexpected since previous studies showed that genes affected by CNVs are often
357 close to TEs due to their mutagenic potential (e.g. Hartmann and Croll, 2017). A reason for
358 this finding might be the high abundance of TEs throughout the *C. geophilum* genome.
359 Although de Freitas *et al.* (2018) found islands of TE-rich and TE-poor regions, these were
360 rather small and scattered throughout the scaffolds so that most genes have TEs in close
361 vicinity. With about 75% of its genome comprising TEs, *C. geophilum* exhibits the highest
362 proportion of transposons and retrotransposons found in ECM fungi (18–75%; Miyauchi *et al.*,
363 2020) and other Dothideomycetes (1–75%; Peter *et al.*, 2016). Our findings corroborate the
364 recent proliferation of species-specific TEs and its evolutionary significance in affecting
365 functional rearrangements of the genome (Daboussi and Capy, 2003; Muszewska *et al.*, 2019).
366 However, other mechanisms such as aneuploidy, chromosomal duplication, unequal crossing
367 over may also be responsible for these gene duplications (Bourne *et al.*, 2014). Such high
368 intraspecific genomic diversity has lately been observed for other long-considered asexual
369 fungi, such as the ancient arbuscular mycorrhizal glomeromycetes, and TE activity was one of
370 the explanatory mechanism (Chen *et al.*, 2018). Likewise, genomic studies indicated the
371 presence of cryptic sex in these species, which, as in our case, might explain the high
372 adaptability of these fungi to a wide range of hosts and habitats and likely has functional
373 significance in host-environment interactions (Chen *et al.*, 2018; Reinhardt *et al.*, 2021).

374 As in another ascomycete ECM fungus, *Tuber melanosporum* (Payen *et al.*, 2015), we
375 found a large proportion of genes that exhibited an excess of low frequency polymorphisms
376 relative to neutral expectation (Tajima's $D < -1$, non-ref. cluster; Fig. 6B). Candidate genes
377 identified as putatively under selection need to be interpreted with caution given our limited

378 sample size and heterogeneous genetic structure. Nevertheless, a large number of genes
379 identified in the recombining ref. cluster are putatively implicated in self/non-self recognition
380 containing HET, NACHT, WD40- and Ankyrin-repeat domains and are members of expanded
381 gene families. Such HET domain proteins were also identified as putative candidates under
382 selection for the outcrossing ECM ascomycete *T. melanosporum* in a similar analysis of six
383 isolates covering its geographical distribution (Payen *et al.*, 2015). Fungal genomes contain a
384 large number of genes encoding HET domains of unknown function, but resolved functions
385 often relate to vegetative incompatibility (e.g. conspecific recognition; Paoletti, 2016), hetero-
386 specific, or even inter-kingdom recognition as seen for characterised NOD-like receptors
387 (Dyrka *et al.*, 2014). These genes were not up- or down-regulated in ECM root tips as compared
388 with free-living mycelium in a synthesis experiment (Peter *et al.*, 2016) and, hence, are likely
389 not involved in specific symbiotic functions in *C. geophilum.* The repeat domains have been
390 shown to evolve under positive diversifying selection (Dyrka et al 2014) as were some groups
391 of HET domain containing proteins that displayed higher allelic diversity and a restricted
392 phylogenetic distribution (Paoletti et al 2016). In the more clonal non-ref. cluster, for which
393 Tajima's *D* values were generally more negative and centred around 0, we found increased
394 number of candidate genes putatively under selection, which are involved in the secondary
395 metabolism. This matches findings from fungal pathogens (e.g. Hartmann *et al.*, 2018).
396 Interestingly, clearer signatures of positive selection were found for an outcrossing fungal
397 pathogen species as compared to a more clonally reproducing generalist fungal pathogen
398 (Derbyshire, 2020).

399 *Genomic rearrangements of the MAT1 locus*

400 The bipolar mating-type (*MAT*) system governs sexual reproduction in ascomycetes and has a
401 key role in establishing cell identity and determining the sexual cycle. Two fungal partners can
402 mate only if they carry different *MAT* idiomorphs. Based on our chromosomal synteny

403 analysis, we showed that both mating-types exist in *C. geophilum* isolates with the *MAT1-1* or
404 the *MAT1-2* idiomorph. In fact, after screening the sixteen genome sequences, a single *MAT1*
405 idiomorph was consistently found in each strain, thus confirming the heterothallism of the
406 species by excluding autogamy. Since the first in-depth genomic analysis of the *MAT1* locus
407 of the *C. geophilum* reference genome (strain 1.58; Peter *et al.*, 2016), the genomic architecture
408 of the intact *MAT1* locus was assumed to be stable and no intraspecific reorganisation was
409 reported in the species. Here, we highlight that both *MAT1-1* and *MAT1-2* loci have undergone
410 important rearrangements with gene expansions in surrounding regions. Surprisingly, we
411 identified a *MAT1-1* pseudogene close to the *MAT1-2* idiomorph, i.e., about 3 kbp, and
412 reciprocally, a *MAT1-2* pseudogene close to the *MAT1-1* idiomorph, i.e., about 2 kbp. This
413 raises the question of whether a homothallic ancestor may have existed and been the precursor
414 of independent heterothallic lineages. Interestingly, this *MAT1-1* pseudogene, also found in *G.*
415 *stellatum*, is relatively similar to the functional *MAT1-1* idiomorph characterised in the
416 homothallic outgroup *L. palustris* (Fig. 7), with the same orientation and an equivalent gene
417 length. Moreover, the gene sequence of the *MAT1-2* idiomorph appears to be enlarged in both
418 *C. geophilum* and *G. stellatum* compared to *L. palustris*, indicating important protein
419 modification. Overall, the genetic makeup of the *MAT1-1* and *MAT1-2* pseudogenes most
420 likely originates from a common ancestor shared with *L. palustris*, from which distinct mating
421 types evolved. Such mating shifts, from heterothallism to homothallism, were highlighted in
422 the *Neurospora* genus, with up to six independent transition events assessed from 43 taxa
423 (Nygren *et al.*, 2011). Distinct genetic mechanisms enabled these transitions (e.g. translocation
424 and unequal crossover) with the mediation of specific transposons that reshaped the genetic
425 architecture of the *MAT* locus (Gioti *et al.*, 2012). The high mobility of genetic material in the
426 surrounding *MAT* locus is intriguing and points to the role of TE activities and associated
427 silencing mechanisms such as RIP in shaping the gene architecture and diversity that may

428 orchestrate sexual reproduction in *C. geophilum* (Serrato-Capuchina and Matute, 2018). In
429 several fungal species, TEs were described within or in flanking regions of the *MAT* locus, e.g.
430 in *Blastomyces dermatitidis* (Li *et al.*, 2013), *Ceratocystis* sp. (Simpson *et al.*, 2018), or *T.*
431 *melanosporum* (Martin *et al.*, 2010). Together, these studies support the idea that TEs are
432 frequently inserted at *MAT* loci. They are therefore key evolutionary forces for genome
433 dynamics likely important not only in pathogenic fungal host and environment interactions as
434 earlier stated (Raffaele and Kamoun, 2012), but also for symbiotic interactions, as indicated by
435 increased proliferation of these elements in ECM fungal species, above all in *C. geophilum*
436 (Miyauchi *et al.*, 2020). Further population genomic studies including genomic architecture
437 and prevalence of *MAT* loci within populations are needed to fully understand the success of
438 this widespread and abundant ECM species and its role in forest ecosystems.

439 *Conclusions*

440 Our study provides one of the first intraspecific genomic comparison of an ECM fungal species
441 sampled on a regional and continental scale. Using whole-genome sequencing from 16
442 European isolates of the ubiquitous forest symbiotic fungus *Cenococcum geophilum*, we found
443 divergent lineages in geographically restricted located sampling sites and no strong geographic
444 structuring. Our genome-wide polymorphism analyses support subdivisions within the species
445 and suggest two distinct main genetic groups, one more clonal and one recombinant. The
446 phylogeny of lineages and groupings are largely supported by the abundant copy-number
447 variations found among genomes. Our gene diversity analyses indicate higher genetic diversity
448 in the recombinant group even though the clonal cluster contained almost double as many
449 strains. Although these results need to be interpreted with caution, because of the limited
450 number of strains studied and an uneven sampling, the top-candidate genes putatively under
451 positive selection based on Tajima's *D* statistics differed between the two groups. More genes
452 from lineage-specific, expanded gene families involved in self/non-self recognition were found

453 in the recombinant cluster, whereas genes involved in the secondary metabolism detected in
454 the more clonal cluster. Lastly, we confirmed the heterothallism of *C. geophilum* based on the
455 chromosomal synteny analysis of the *MAT1-1* and *MAT1-2* idiomorphs and uncovered
456 important genetic rearrangements in their surrounding coding and non-coding regions for both,
457 strains carrying the same or opposite *MAT1* idiomorphs. Collectively, these findings
458 demonstrate complex genome architectures in this important ECM symbiont that maybe due
459 to cryptic sex and/or TE-based mechanisms, as similarly suggested for the long-believed
460 asexual arbuscular mycorrhizal fungi. Future studies will elucidate the role that this genetic
461 diversity plays in symbiotic function and the potential for adaptation to changing environments.

462 ## Experimental Procedures

463 *Study range and selection of isolates*

464 In this study, we selected 16 *Cenococcum geophilum* isolates to cover as broad as possible the
465 European range of the species and the divergent phylogenetic lineages previously reported
466 (Obase *et al.*, 2016). All strains were isolated from sclerotia collected below diverse host tree
467 species, i.e., broadleaves and conifers, to maximise intra-specific genetic diversity in close and
468 distant sampling sites. After screening, both mating types were included to investigate sexual
469 reproduction from natural populations (Fig. 1; Table 1). Pure cultures of *C. geophilum* were
470 grown in liquid culture containing a Cg-Medium, i.e., a modified Melin Norkran's (MMN)
471 medium comprising casein (Kerner *et al.*, 2012), and incubated at room temperature in the
472 dark. We cultivated isolates for a month and harvested mycelia that were then pulverised in
473 liquid nitrogen for storage at –80°C until processing.

474 *DNA extraction and barcoding*

475 Genomic DNA was extracted using the PowerMax Soil DNA extraction kit (MoBio, now
476 DNeasy PowerMax Soil kit of Qiagen) according to the manufacturer's instructions. DNA was
477 quantified by spectrophotometry (Qubit, Thermo Fisher Scientific, Waltham, USA), and the
478 quality was inspected on agarose gel (1.5%). A minimum of 1 µg of DNA was sent to the US
479 Department of Energy Joint Genome Institute (JGI) for genome resequencing.

480 To identify the lineage of each isolate, we amplified and sequenced two
481 phylogenetically informative loci at the intraspecific level, the glyceraldehyde 3-phosphate
482 dehydrogenase (GAPDH) gene and the internal transcribed spacer (ITS) region (Obase *et al.*,
483 2016), using Sanger dideoxy sequencing. The optimised polymerase chain reaction (PCR)
484 programs for ITS and the GAPDH gene were conducted following Obase *et al*. (2014). For the
485 GAPDH gene, we used universal primers, as reported in Berbee *et al.* (1999), and for ITS, we

486 used ITS1F and ITS4 (White *et al.*, 1990; Gardes and Bruns, 1993). PCR products were cleaned
487 up using the illustra™ ExoProStar™ enzymatic PCR and sequence reaction clean kit according
488 to the manufacturer's instructions (Merck KGaA, Darmstadt, Germany). Cycle sequencing was
489 performed using the BigDye™ Terminator v3.1 Cycle Sequencing kit and products were
490 purified with the BigDye XTerminator™ purification kit (Thermo Fisher Scientific) before to
491 analyse them on an Applied Biosystems ABI3130 genetic analyser. Loci were sequenced with
492 both forward and reverse primers to ensure unambiguous chromatogram profiles and all
493 sequences were deposited to NCBI (MZ570241-MZ570257; MZ577329-MZ577345).

494 *Genome sequencing and reference genome sequences*

495 Plate-based DNA library preparation for Illumina sequencing was performed on the Sciclone
496 NGS robotic liquid handling system (PerkinElmer, Waltham, USA) using KAPA Biosystems
497 library preparation kit (Roche, Basel, Switzerland). For each sample, 200 ng of DNA was
498 sheared to 600 base pairs (bp) using a Covaris LE220 focused-ultrasonicator (Woburn,
499 Massachusetts, USA). The sheared DNA fragments were size selected by double-SPRI and
500 then the selected fragments were end-repaired, A-tailed, and ligated with Illumina compatible
501 sequencing adaptors from IDT containing a unique molecular index barcode for each sample
502 library. The prepared library was quantified using KAPA Biosystem's next-generation
503 sequencing library qPCR kit and run on a Roche LightCycler 480 real-time PCR instrument.
504 The quantified library was then multiplexed with other libraries, and the pool of libraries was
505 then prepared for sequencing on the Illumina HiSeq sequencing platform utilising a TruSeq
506 paired-end cluster kit, v4, and Illumina's cBot instrument to generate a clustered flow cell for
507 sequencing (San Diego, California, USA). Sequencing of the flow cell was performed on the
508 Illumina HiSeq 2500 sequencer using HiSeq TruSeq SBS sequencing kits, v4, following a
509 2×100 indexed run recipe at the Joint Genome Institute (JGI; Walnut Creek, California, USA).

510 We accessed the *C. geophilum* raw reads data from the NCBI Short Read Archive
511 (Table S1) and used the *Glonium stellatum* genome as an outgroup species (Peter *et al.*, 2016;
512 https://genome.jgi.doe.gov/Glost2) in downstream analyses. The *C. geophilum* reference
513 genome sequence and gene annotation data (filtered models) for the strain 1.58 were
514 downloaded from the JGI (Peter *et al.*, 2016; https://genome.jgi.doe.gov/Cenge3). The
515 sequenced strains included the reference genome strain 1.58 to control for the quality of SNP
516 calling (Peter *et al.*, 2016).

517 *Read mapping, variant calling, and polymorphism filtering*

518 We downloaded sequencing data using fastq-dump, as implemented in the SRA-Toolkit v.
519 2.9.0 (http://ncbi.github.io/sra-tools/), and inspected read quality with FastQC v. 0.11
520 (Andrews, 2017). We performed quality trimming of raw reads with Trimmomatic v. 0.36
521 (Bolger *et al.*, 2014) based on the following quality filters: ILLUMINACLIP:TruSeq3-
522 PE.fa:2:30:10, SLIDINGWINDOW:5:10, LEADING:10, TRAILING:10, and MINLEN:50.
523 Trimmed reads were mapped on the reference genome 1.58 using Bowtie v. 2.3 with the -very-
524 sensitive and -local settings (Langmead *et al.*, 2009). The resulting BAM files were sorted by
525 position and indexed with SAMtools v. 1.6 (Li *et al.*, 2009).

526 We applied the Genome Analysis Toolkit (GATK) v. 3.7 and used the HaplotypeCaller
527 to perform the SNP and indel discovery (McKenna *et al.*, 2010). Individual SNP calls were
528 merged and jointly genotyped using the GATK GenotypeGVCFs tool. Then, we filtered the
529 joint Variant Call Format (VCF) file using hard filtering parameters according to GATK best
530 practices recommendations (Van der Auwera *et al.*, 2013). The following thresholds were
531 applied for filtering: AN ≥ 12, i.e., at least 12 isolates had a callable genotype, QUAL ≥ 5,000,
532 QD ≥ 5.0, MQ ≥ 20.0, ReadPosRankSum, MQRankSum and BaseQRankSum each comprised
533 between –2.0 and 2.0. We further classified SNP loci as synonymous or non-synonymous
534 variant sites using SnpEff v. 4.3p (Cingolani *et al.*, 2012).

535 *Phylogenetic tree reconstructions*

536 To infer the phylogenetic relationships of the 16 *C. geophilum* strains within the seven main
537 clades described in Obase *et al.* (2016), we assembled a broader dataset (GAPDH and ITS) and
538 reconstructed a two-loci phylogenetic tree based on maximum likelihood (ML) analysis. The
539 additional strains (22 *C. geophilum*, one *Pseudocenococcum floridanum*, and one *G. stellatum*)
540 covered most of the phylogenetic diversity and intercontinental distribution of the species. This
541 includes important disjunctions in North America and potential cryptic taxa in European Clades
542 5 and 6 (Obase *et al.*, 2016), which were investigated using a coalescent-based species
543 delimitation approach (Zhang *et al.*, 2013). For each locus, we aligned sequences using
544 MAFFT v. 7 with default parameters (Katoh *et al.*, 2019) and excluded ambiguous regions
545 using Gblocks v.0.91b with the less stringent settings (Castresana, 2000). For each locus, we
546 implemented the optimal substitution model, based on AICc scores calculated with the program
547 ModelTest-NG v. 0.1.7 (Darriba *et al.*, 2019; Table S2). We ran ML tree searches with a
548 randomised maximum parsimony starting tree and gamma-distributed site rates using RAxML-
549 NG v. 1.0.0 (Kozlov *et al.*, 2019; Edler *et al.*, 2021). We performed 20 runs and set 1,000
550 bootstrap replicates for each run, computed on the best-scoring ML tree with a bootstrap
551 random-number seed. We fixed *G. stellatum* as the outgroup to root phylograms.

552 At the genome-wide level, we generated two SNP datasets to infer independent
553 phylogenetic tree reconstructions. First, from the overall SNP data genotyped in all 17 strains
554 (16 *C. geophilum* and one *G. stellatum*), we randomly selected a single SNP per 1 kilo base-
555 pair (kbp) of the reference genome sequence using VCFtools v. 0.1.15 (Danecek *et al.*, 2011).
556 Second, from the overall SNP dataset, we retained only synonymous variants identified in
557 coding regions (see below SnpEff). We used each SNP dataset for an independent ML tree
558 reconstruction with RAxML-NG (Kozlov *et al.*, 2019; Edler *et al.*, 2021). We ran the best ML
559 tree by analysing 20 different starting trees and performed 1,000 bootstrap replicates on the

560 best-scoring ML tree with a bootstrap random-number seed. We implemented a General Time
561 Reversible (GTR) substitution model and allowed for rate heterogeneity (GTR+G).

562 *Recombination events and genetic structure*

563 To gain insights into recombination events and genetic structure, we generated a SNP dataset
564 for intraspecific genetic analyses of *C. geophilum* using the following criteria: a single SNP
565 randomly selected per 1 kbp of the reference genome and genotyped in a least 90% of the
566 strains, i.e., genotyped in at least 15 out of 16 strains. We used SplitsTree v. 4.14.6 to build up
567 a Neighbour-Net network from all *C. geophilum* strains (Huson and Bryant, 2006), and
568 uncorrected p distances calculated from the SNP matrix, which is the proportion of positions
569 at which two sequences differ, to generate the network. The network was drawn based on equal
570 angle splits. Tests of recombination within genetic groups of strains were carried out using the
571 Φ test implemented in SplitsTree (Bruen *et al.*, 2006). The test identifies evidence of
572 recombination based on phylogenetic compatibilities of neighbouring SNPs in the matrix. We
573 ran the test with the default settings of a 100-characters window size on the SNP matrix. Three
574 genotype datasets were analysed to assess lineage-specific patterns; comprising all strains
575 (referred as "all"), a reference cluster which contained the reference genome 1.58 (referred as
576 "ref. cluster" including 1.005, 1.057, 1.061, 1.064, 1.076, and 1.58), and a non-reference cluster
577 which did not contain the reference genome (referred as "non-ref. cluster" including 1.022,
578 1.060, 1.075, 1.094, 1.096, 1.097, 1.100, 1.101, 1.102, and 1.104).

579 We conducted Bayesian unsupervised genetic clustering using STRUCTURE v. 2.3.4
580 (Pritchard *et al.*, 2000). We used an admixture model with correlated frequencies and no prior
581 regarding the population of origin (Falush *et al.*, 2003). The *K* parameter was tested for values
582 ranging from 1 to 5, with 10 repetitions for each tested *K* value. We used 50,000 iterations as
583 a burn-in period and 100,000 iterations per run for the Monte Carlo Markov Chain (MCMC)
584 replicates. We visually inspected the convergence of parameters and diagnosed the optimal *K*

585 value based on the stabilisation of –Ln(*K*) values. We performed a principal component
586 analysis (PCA) to explore genotypic variability based on the SNP matrix using the R package
587 SNPRelate v. 1.20.1 (Zheng *et al.*, 2012).

588 *Detection of copy-number variations*

589 We identified segmental deletions and duplications in sequenced isolates compared to the
590 reference genome 1.58 using CNVcaller (Wang *et al.*, 2017). CNVcaller uses normalised read-
591 depth of short read alignments to identify high-confidence copy-number variation (CNV)
592 regions. The window size for the CNV calling was set to 500 bp according to the
593 recommendations based on sequencing coverage. The lower and upper GC content limits were
594 set to 0.2 and 0.7, respectively, and the upper limit of gap content was set to 0.5. We required
595 a minimum of Pearson's correlation coefficient between two adjacent non-overlapping
596 windows to be >0.5, as recommended for less than 30 samples. We also specified that only a
597 single individual displayed a potential gain/loss event to define a candidate CNV window (-h
598 1 and -f 0.05) given the heterogeneity in genotypes and small sample size. CNV regions were
599 further filtered based on the normalised read depth. Deletions and duplications were retained
600 if the normalised read depth was <0.4 and >1.8, respectively. CNV regions called for the
601 reference genome 1.58 were aligned against the reference genome for validation, which
602 showed possible issues in read alignment and/or the local assembly quality. Hence, gene
603 deletions and duplications among *C. geophilum* strains were only retained if the reference strain
604 showed a normalised read depth between 0.7–1.3 for the same region. A gene was considered
605 affected by a duplication or deletion if more than 50% of the coding sequence length was
606 overlapping with a CNV region using the R package ggplot2 v. 3.3.0 (Wickham, 2011). One
607 CNV region may contain several duplicated or deleted genes. Then, genomic distances between
608 all genes and their closest transposable elements (TEs), and between CNV genes and their

609 closest TEs were calculated with the *closest* function in BEDTools v 2.28.0 (Quinlan and Hall,
610 2010). Significance between median values was calculated using a two-sided Wilcoxon test.

611 *Within-species genetic diversity and signatures of selection*

612 To improve our understanding of genomic signatures of selection, we analysed three species-
613 specific estimates of gene diversity and their deviations from genome-wide average; nucleotide
614 diversity per site (π), Tajima's *D*, and *p*N/*p*S ratios. We assessed these population genetic
615 statistics for individual genes using the R package PopGenome v. 2.6.1 (Pfeifer *et al.*, 2014)
616 and included all SNPs falling within coding sequence boundaries that were genotyped in at
617 least 15 out of 16 *C. geophilum* strains. At the genome-wide level, a negative Tajima's *D* is
618 indicative of a population size expansion (e.g., after a bottleneck), whereas a positive Tajima's
619 *D* is supporting a decrease in population size. At the gene-specific level, negative Tajima's *D*
620 values are indicating of recent positive selection, while positive values are imputable to
621 balancing selection (Hohenlohe *et al.*, 2010). For π estimates, a decrease in gene-specific
622 values in comparison to the average genome-wide level suggests a signature of positive
623 selection (Hohenlohe *et al.*, 2010). Regarding *p*N/*p*S, a ratio greater than one implies positive
624 selection, whereas a value less than one assumes purifying selection (Tanaka and Nei, 1989).
625 The three genotype datasets (see above) were analysed to evaluate the robustness of the results.
626 To further examine the genes putatively under selection, we identified top-candidate genes
627 based on the 20 lowest Tajima's *D* values in each of the two datasets, the ref. and non-ref.
628 cluster. Functional annotation of the top-candidate genes was carried as outlined in Peter *et al.*,
629 (2016) and according to the JGI annotations (mycocosm.jgi.doe.gov), including predicted gene
630 ontology (GO), EuKaryotic Orthologous Groups (KOG), protein family clarification using the
631 InterPro database (Blum *et al.*, 2021), Carbohydrate-Active enZYme (CAZY; Cantarel *et al.*,
632 2009) and secondary metabolism genes and cluster prediction (Khaldi *et al.*, 2010). Gene
633 expression data from free-living mycelium and ECM root-tips were taken from the study of

634 Peter *et al*. 2016 and presented here. In addition, we examined the affiliation to multigene
635 families, their PFAM domains and family size changes according to expansion/depletion
636 analysis as detailed in Ohm *et al*. (2012) and CAFE analyses (De Bie *et al.*, 2006) using the
637 species set outlined in Peter *et al*. 2016 and available on the JGI *C. geophilum* 1.58 portal under
638 the MCL Clusters Run "Cenococcum + 50 genomes 03/21/14.399".

639 *Chromosomal synteny analysis of the MAT1 locus*

640 We analysed chromosomal synteny of the *MAT1* locus containing the *MAT1-1* and *MAT1-2*
641 idiomorphs based on our *de novo* genome assemblies of individual *C. geophilum* isolates. We
642 used SPAdes v. 3.11.1 to assemble genomic scaffolds based on the paired-end Illumina
643 sequencing data described above (Bankevich *et al.*, 2012). The SPAdes pipeline includes the
644 BayesHammer read error correction module and builds contigs in a stepwise procedure based
645 on increasing kmer lengths. We defined a kmer range of 21, 33, 45, 67 and 89 and enabled the
646 --careful option to reduce mismatches and indel errors in the assembly. Assembled scaffolds
647 were error-corrected based on read realignments. We used the tool QUAST v. 4.6.0 to retrieve
648 assembly statistics for each strain (Gurevich *et al.*, 2013) and performed a blastn to identify
649 assembled scaffolds that spanned the *MAT1* loci of the reference genome strain 1.58 (Camacho
650 *et al.*, 2009; Peter *et al.*, 2016). For this, three genes in each direction of the *MAT1* locus were
651 inspected. Once matching scaffolds were identified in individual genome assemblies, we
652 performed pairwise blastn analyses between these scaffolds. Segmental blast hits between pairs
653 of scaffolds known to contain the *MAT1* locus were leniently filtered for a minimum identity
654 of 50% and a minimum alignment length of 50 bp to assess synteny patterns. For the
655 comparison of *G. stellatum*, *L. palustris* and *C. geophilum*, we performed a tblastx search
656 translated nucleotide and filtered additionally for a minimum bitscore of 100. We visualised
657 pairwise synteny among scaffolds using the R package GenoPlotR v. 0.8.7 (Guy *et al.*, 2011).

Acknowledgements

We thank Stephanie Pfister and Barbara Meier for help in the laboratory; the staff at the Joint Genome Initiative for sequencing. We thank Claire Veneault-Fourrey for helpful comments and discussions. This project was supported within the framework of the ARBRE/WSL project "Blacksecret" by grants from the WSL and the French National Agency of Research (ANR), and as part of the "Investissement d'Avenir" program (ANR-11_LABX-0002-01) of Labex ARBRE (CFP15; to AK and MP) and the EC-supported Network of Excellence Evoltree (GOCE-016322 to MP).

666 ## References

667 Andrews, S. (2017) FastQC: a quality control tool for high throughput sequence data.

668 Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine,
669 A., et al. (2013) From FastQ data to high-confidence variant calls: The genome analysis
670 toolkit best practices pipeline. *Curr Protoc Bioinforma* **43**: 1–33.

671 Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., et al. (2012)
672 SPAdes: A new genome assembly algorithm and its applications to single-cell
673 sequencing. *J Comput Biol* **19**: 455–477.

674 Bazzicalupo, A.L., Ruytinx, J., Ke, Y.H., Coninx, L., Colpaert, J. V., Nguyen, N.H., et al.
675 (2020) Fungal heavy metal adaptation through single nucleotide polymorphisms and
676 copy-number variation. *Mol Ecol* **29**: 4157–4169.

677 Berbee, M.L., Pirseyedi, M., and Hubbard, S. (1999) Cochliobolus phylogenetics and the origin
678 of known, highly virulent pathogens, inferred from ITS and glyceraldehyde-3-phosphate
679 dehydrogenase gene sequences. *Mycologia* **91**: 964–977.

680 De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006) CAFE: A computational tool
681 for the study of gene family evolution. *Bioinformatics* **22**: 1269–1271.

682 Blum, M., Chang, H.Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., et al.
683 (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids*
684 *Res* **49**: D344–D354.

685 Bolger, A.M., Lohse, M., and Usadel, B. (2014) Trimmomatic: A flexible trimmer for Illumina
686 sequence data. *Bioinformatics* **30**: 2114–2120.

687 Bourne, E.C., Mina, D., Gonçalves, S.C., Loureiro, J., Freitas, H., and Muller, L.A. (2014)
688 Large and variable genome size unrelated to serpentine adaptation but supportive of
689 cryptic sexuality in *Cenococcum geophilum*. *Mycorrhiza* **24**: 13–20.

690 van den Brink, J. and de Vries, R.P. (2011) Fungal enzyme sets for plant polysaccharide
691 degradation. *Appl Microbiol Biotechnol* **91**: 1477–1492.

692 Bruen, T.C., Philippe, H., and Bryant, D. (2006) A simple and robust statistical test for
693 detecting the presence of recombination. *Genetics* **172**: 2665–2681.

694 Cairney, J.W. (1999) Intraspecific physiological variation: Implications for understanding
695 functional diversity in ectomycorrhizal fungi. *Mycorrhiza* **9**: 125–135.

696 Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden,
697 T.L. (2009) BLAST+: Architecture and applications. *BMC Bioinformatics* **10**: 1–9.

698 Cantarel, B.I., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B.
699 (2009) The Carbohydrate-Active EnZymes database (CAZy): An expert resource for
700 glycogenomics. *Nucleic Acids Res* **37**: 233–238.

701 Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in
702 phylogenetic analysis. *Mol Biol Evol* **17**: 540–552.

703 Charlesworth, D. and Willis, J.H. (2009) The genetics of inbreeding depression. *Nat Rev Genet*
704 **10**: 783–796.

705 Chen, E.C.H., Morin, E., Beaudet, D., Noel, J., Yildirir, G., Ndikumana, S., et al. (2018) High
706 intraspecific genome diversity in the model arbuscular mycorrhizal symbiont
707 *Rhizophagus irregularis*. *New Phytol* **220**: 1161–1171.

708 Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., et al. (2012) A program
709 for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff:
710 SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**: 80–92.

711 Coleman, M.D., Bledsoe, C.S., and Lopushinsky, W. (1989) Pure culture response of
712 ectomycorrhizal fungi to imposed water stress. *Can J Bot* **67**: 29–39.

713 Daboussi, M.J. and Capy, P. (2003) Transposable elements in filamentous fungi. *Annu Rev*
714 *Microbiol* **57**: 275–299.

715 Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., et al. (2011)
716 The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

717 Darriba, D., Posada, D., Kozlov, A.M., Stamatakis, A., Morel, B., and Flouri, T. (2020)
718 ModelTest-NG: A new and scalable tool for the selection of DNA and protein
719 evolutionary models. *Mol Biol Evol* **37**: 291–294.

720 Debuchy, R., Berteaux-Lecellier, V., and Silar, P. (2014) Mating systems and sexual
721 morphogenesis in ascomycetes. In *Cellular and molecular biology of filamentous fungi*.
722 Borkovich, K.A. and Ebbole, D.J. (eds). Washington, U.S.A.: ASM Press, pp. 499–535.

723 Derbyshire, M.C. (2020) Bioinformatic detection of positive selection pressure in plant
724 pathogens: The neutral theory of molecular sequence evolution in action. *Front Microbiol*
725 **11**: 1–14.

726 Douhan, G.W., Martin, D.P., and Rizzo, D.M. (2007) Using the putative asexual fungus
727 *Cenococcum geophilum* as a model to test how species concepts influence recombination
728 analyses using sequence data from multiple loci. *Curr Genet* **52**: 191–201.

729 Douhan, G.W. and Rizzo, D.M. (2005) Phylogenetic divergence in a local population of the
730 ectomycorrhizal fungus *Cenococcum geophilum*. *New Phytol* **166**: 263–271.

731 Dyrka, W., Lamacchia, M., Durrens, P., Kobe, B., Daskalov, A., Paoletti, M., et al. (2014)
732 Diversity and variability of NOD-like receptors in fungi. *Genome Biol Evol* **6**: 3137–3158.

733 Edler, D., Klein, J., Antonelli, A., and Silvestro, D. (2021) raxmlGUI 2.0: A graphical interface
734 and toolkit for phylogenetic analyses using RAxML. *Methods Ecol Evol* **12**: 373–377.

735 Falush, D., Stephens, M., and Pritchard, J.K. (2003) Inference of population structure using
736 multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164**:
737 1567–1587.

738 Fernandez, C.W., McCormack, M.L., Hill, J.M., Pritchard, S.G., and Koide, R.T. (2013) On
739 the persistence of *Cenococcum geophilum* ectomycorrhizas and its implications for forest
740 carbon and nutrient cycles. *Soil Biol Biochem* **65**: 141–143.

741 de Freitas Pereira, M., Veneault-Fourrey, C., Vion, P., Guinet, F., Morin, E., Barry, K.W., et
742 al. (2018) Secretome analysis from the ectomycorrhizal ascomycete *Cenococcum*
743 *geophilum*. *Front Microbiol* **9**: 1–17.

744 Gardes, M. and Bruns, T.D. (1993) ITS primers with enhanced specificity for basidiomycetes
745 - application to the identification of mycorrhizae and rusts. *Mol Ecol* **2**: 113–118.

746 Geml, J., Laursen, G.A., O'Neill, K., Nusbaum, H.C., and Taylor, D.L. (2006) Beringian
747 origins and cryptic speciation events in the fly agaric (*Amanita muscaria*). *Mol Ecol* **15**:
748 225–239.

749 Gioti, A., Mushegian, A.A., Strandberg, R., Stajich, J.E., and Johannesson, H. (2012)
750 Unidirectional evolutionary transitions in fungal mating systems and the role of

751 transposable elements. *Mol Biol Evol* **29**: 3215–3226.

752 Gonçalves, S.C., Martins-Loução, M.A., and Freitas, H. (2009) Evidence of adaptive tolerance
753 to nickel in isolates of *Cenococcum geophilum* from serpentine soils. *Mycorrhiza* **19**: 221–
754 230.

755 Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013) QUAST: Quality assessment tool
756 for genome assemblies. *Bioinformatics* **29**: 1072–1075.

757 Guy, L., Kultima, J.R., Andersson, S.G.E., and Quackenbush, J. (2011) genoPlotR:
758 comparative gene and genome visualization in R. *Bioinformatics* **27**: 2334–2335.

759 Hartmann, F.E. and Croll, D. (2017) Distinct trajectories of massive recent gene gains and
760 losses in populations of a microbial eukaryotic pathogen. *Mol Biol Evol* **34**: 2808–2822.

761 Hartmann, F.E., McDonald, B.A., and Croll, D. (2018) Genome-wide evidence for divergent
762 selection between populations of a major agricultural pathogen. *Mol Ecol* **27**: 2725–2741.

763 Hasselquist, N., Germino, M.J., McGonigle, T., and Smith, W.K. (2005) Variability of
764 *Cenococcum* colonization and its ecophysiological significance for young conifers at
765 alpine-treeline. *New Phytol* **165**: 867–873.

766 Hemsworth, G.R., Henrissat, B., Davies, G.J., and Walton, P.H. (2014) Discovery and
767 characterization of a new family of lytic polysaccharide monooxygenases. *Nat Chem Biol*
768 **10**: 122–126.

769 Herzog, C., Peter, M., Pritsch, K., Günthardt-Goerg, M.S., and Egli, S. (2013) Drought and air
770 warming affects abundance and exoenzyme profiles of *Cenococcum geophilum* associated
771 with *Quercus robur*, *Q. petraea* and *Q. pubescens*. *Plant Biol* **15**: 230–237.

772 Hohenlohe, P.A., Phillips, P.C., and Cresko, W.A. (2010) Using population genomics to detect
773 selection in natural populations: Key concepts and methodological considerations. *Int J*
774 *Plant Sci* **171**: 1059–1071.

775 Huson, D.H. and Bryant, D. (2006) Application of phylogenetic networks in evolutionary
776 studies. *Mol Biol Evol* **23**: 254–267.

777 Johnson, D., Martin, F., Cairney, J.W., and Anderson, I.C. (2012) The importance of
778 individuals: Intraspecific diversity of mycorrhizal plants and fungi in ecosystems. *New*
779 *Phytol* **194**: 614–628.

780 Katoh, K., Rozewicki, J., and Yamada, K.D. (2019) MAFFT online service: Multiple sequence
781 alignment, interactive sequence choice and visualization. *Brief Bioinform* **20**: 1160–1166.

782 Kerner, R., Delgado-Eckert, E., Del Castillo, E., Müller-Starck, G., Peter, M., Kuster, B., et al.
783 (2012) Comprehensive proteome analysis in *Cenococcum geophilum* Fr. as a tool to
784 discover drought-related proteins. *J Proteomics* **75**: 3707–3719.

785 Khaldi, N., Seifuddin, F.T., Turner, G., Haft, D., Nierman, W.C., Wolfe, K.H., and Fedorova,
786 N.D. (2010) SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal*
787 *Genet Biol* **47**: 736–741.

788 Kozlov, A.M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019) RAxML-NG: A
789 fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference.
790 *Bioinformatics* **35**: 4453–4455.

791 Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009) Ultrafast and memory-efficient
792 alignment of short DNA sequences to the human genome. *Genome Biol* **10**: 1–10.

793 Larsen, P.E., Sreedasyam, A., Trivedi, G., Desai, S., Dai, Y., Cseke, L.J., and Collart, F.R.

794 (2016) Multi-omics approach identifies molecular mechanisms of plant-fungus
795 mycorrhizal interaction. *Front Plant Sci* **6**: 1–17.

796 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009) The sequence
797 alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

798 Li, W., Sullivan, T.D., Walton, E., Averette, A.F., Sakthikumar, S., Cuomo, C.A., et al. (2013)
799 Identification of the mating-type (*MAT*) locus that controls sexual reproduction of
800 blastomyces dermatitidis. *Eukaryot Cell* **12**: 109–117.

801 van der Linde, S., Suz, L.M., Orme, D. l., Cox, F., Andreae, H., Asi, E., et al. (2018)
802 Environment and host as large-scale controls of ectomycorrhizal fungi. *Nature* **558**: 243–
803 248.

804 LoBuglio, K.F. (1999) *Cenococcum*. In *Ectomycorrhizal Fungi: Key Genera in Profile*.
805 Cairney, J.W.G. and Chambers, S.M. (eds). Heidelberg, Germany: Springer, pp. 287–309.

806 LoBuglio, K.F. and Taylor, J.W. (2002) Recombination and genetic differentiation in the
807 mycorrhizal fungus *Cenococcum geophilum* Fr. *Mycologia* **94**: 772–780.

808 Lofgren, L.A., Nguyen, N.H., Vilgalys, R., Ruytinx, J., Liao, H.L., Branco, S., et al. (2021)
809 Comparative genomics reveals dynamic genome evolution in host specialist
810 ectomycorrhizal fungi. *New Phytol* **230**: 774–792.

811 Martin, F., Kohler, A., Murat, C., Balestrini, R., Coutinho, P.M., Jaillon, O., et al. (2010)
812 Périgord black truffle genome uncovers evolutionary origins and mechanisms of
813 symbiosis. *Nature* **464**: 1033–1038.

814 Martin, F., Kohler, A., Murat, C., Veneault-Fourrey, C., and Hibbett, D.S. (2016) Unearthing
815 the roots of ectomycorrhizal symbioses. *Nat Rev Microbiol* **14**: 760–773.

816 McKenna, A., Hanna, M., Banks, E., Sivachenko, A.Y., Cibulskis, K., Kernytsky, A.M., et al.
817 (2010) The genome analysis toolkit: A MapReduce framework for analyzing next-
818 generation DNA sequencing data. *Genome Res* **20**: 1297–1303.

819 Miyauchi, S., Kiss, E., Kuo, A., Drula, E., Kohler, A., Sánchez-García, M., et al. (2020) Large-
820 scale genome sequencing of mycorrhizal fungi provides insights into the early evolution
821 of symbiotic traits. *Nat Commun* **11**: 1–17.

822 Mujic, A.B., Durall, D.M., Spatafora, J.W., and Kennedy, P.G. (2016) Competitive avoidance
823 not edaphic specialization drives vertical niche partitioning among sister species of
824 ectomycorrhizal fungi. *New Phytol* **209**: 1174–1183.

825 Muszewska, A., Steczkiewicz, K., Stepniewska-Dziubinska, M., and Ginalski, K. (2019)
826 Transposable elements contribute to fungal genes and impact fungal lifestyle. *Sci Rep* **9**:
827 1–10.

828 Nygren, K., Strandberg, R., Wallberg, A., Nabholz, B., Gustafsson, T., García, D., et al. (2011)
829 A comprehensive phylogeny of *Neurospora* reveals a link between reproductive mode
830 and molecular evolution in fungi. *Mol Phylogenet Evol* **59**: 649–663.

831 Obase, K., Douhan, G.W., Matsuda, Y., and Smith, M.E. (2014) Culturable fungal assemblages
832 growing within *Cenococcum* sclerotia in forest soils. *FEMS Microbiol Ecol* **90**: 708–717.

833 Obase, K., Douhan, G.W., Matsuda, Y., and Smith, M.E. (2018) Isolation source matters:
834 Sclerotia and ectomycorrhizal roots provide different views of genetic diversity in
835 *Cenococcum geophilum*. *Mycologia* **110**: 473–481.

836 Obase, K., Douhan, G.W., Matsuda, Y., and Smith, M.E. (2017) Progress and challenges in

837 understanding the biology, diversity, and biogeography of *Cenococcum geophilum*. In
838 *Biogeography of mycorrhizal symbiosis*. Tedersoo, L. (ed)., pp. 299–317.

839 Obase, K., Douhan, G.W., Matsuda, Y., and Smith, M.E. (2016) Revisiting phylogenetic
840 diversity and cryptic species of *Cenococcum geophilum* sensu lato. *Mycorrhiza* **26**: 529–
841 540.

842 Ohm, R.A., Feau, N., Henrissat, B., Schoch, C.L., Horwitz, B.A., Barry, K.W., et al. (2012)
843 Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen
844 Dothideomycetes fungi. *PLoS Pathog* **8**:.

845 Paoletti, M. (2016) Vegetative incompatibility in fungi: From recognition to cell death,
846 whatever does the trick. *Fungal Biol Rev* **30**: 152–162.

847 Payen, T., Murat, C., Gigant, A., Morin, E., De Mita, S., and Martin, F. (2015) A survey of
848 genome-wide single nucleotide polymorphisms through genome resequencing in the
849 Périgord black truffle (*Tuber melanosporum* Vittad.). *Mol Ecol Resour* **15**: 1243–1255.

850 Peter, M., Kohler, A., Ohm, R.A., Kuo, A., Krützmann, J., Morin, E., et al. (2016)
851 Ectomycorrhizal ecology is imprinted in the genome of the dominant symbiotic fungus
852 *Cenococcum geophilum*. *Nat Commun* **7**: 12662.

853 Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S.E., and Lercher, M.J. (2014) PopGenome: An
854 efficient Swiss army knife for population genomic analyses in R. *Mol Biol Evol* **31**: 1929–
855 1936.

856 di Pietro, M., Churin, J.L., and Garbaye, J. (2007) Differential ability of ectomycorrhizas to
857 survive drying. *Mycorrhiza* **17**: 547–550.

858 Pritchard, J.K., Stephens, M., and Donnelly, P. (2000) Inference of population structure using
859 multilocus genotype data. *Genetics* **155**: 945–959.

860 Quinlan, A.R. and Hall, I.M. (2010) BEDTools: A flexible suite of utilities for comparing
861 genomic features. *Bioinformatics* **26**: 841–842.

862 Raffaele, S. and Kamoun, S. (2012) Genome evolution in filamentous plant pathogens: Why
863 bigger can be better. *Nat Rev Microbiol* **10**: 417–430.

864 Reinhardt, D., Roux, C., Corradi, N., and Di Pietro, A. (2021) Lineage-specific genes and
865 cryptic sex: Parallels and differences between arbuscular mycorrhizal fungi and fungal
866 pathogens. *Trends Plant Sci* **26**: 111–123.

867 Riley, R. and Corradi, N. (2013) Searching for clues of sexual reproduction in the genomes of
868 arbuscular mycorrhizal fungi. *Fungal Ecol* **6**: 44–49.

869 Rineau, F. and Courty, P.E. (2011) Secreted enzymatic activities of ectomycorrhizal fungi as a
870 case study of functional diversity and functional redundancy. *Ann For Sci* **68**: 69–80.

871 Rochet, J., Moreau, P.A., Manzi, S., and Gardes, M. (2011) Comparative phylogenies and host
872 specialization in the alder ectomycorrhizal fungi Alnicola, Alpova and Lactarius
873 (Basidiomycota) in Europe. *BMC Evol Biol* **11**: 1–14.

874 Sánchez-Ramírez, S., Tulloss, R.E., Guzmán-Dávalos, L., Cifuentes-Blanco, J., Valenzuela,
875 R., Estrada-Torres, A., et al. (2015) In and out of refugia: Historical patterns of diversity
876 and demography in the North American Caesar's mushroom species complex. *Mol Ecol*
877 **24**: 5938–5956.

878 Serrato-Capuchina, A. and Matute, D.R. (2018) The role of transposable elements in
879 speciation. *Genes* **9**: 1–29.

880 Simpson, M.C., Coetzee, M.P.A., van der Nest, M.A., Wingfield, M.J., and Wingfield, B.D.
881 (2018) Ceratocystidaceae exhibit high levels of recombination at the mating-type (*MAT*)
882 locus. *Fungal Biol* **122**: 1184–1191.

883 Smith, M.E., Henkel, T.W., and Rollins, J.A. (2015) How many fungi make sclerotia? *Fungal*
884 *Ecol* **13**: 211–220.

885 Spatafora, J.W., Owensby, C.A., Douhan, G.W., Boehm, E.W.A., and Schoch, C.L. (2012)
886 Phylogenetic placement of the ectomycorrhizal genus *Cenococcum* in Gloniaceae
887 (Dothideomycetes). *Mycologia* **104**: 758–765.

888 Tanaka, T. and Nei, M. (1989) Positive Darwinian selection observed at the variable-region
889 genes of immunoglobulins. *Mol Biol Evol* **6**: 447–459.

890 Trappe, J.M. (1962) *Cenococcum graniforme* - Its distribution, ecology, mycorrhiza formation,
891 and inherent variation.

892 Vélez, J., Morris, R., Vilgalys, R., Labbé, J., and Schadt, C. (2021) Phylogenetic diversity of
893 200+ isolates of the ectomycorrhizal fungus *Cenococcum geophilum* associated with
894 *Populus trichocarpa* soils in the Pacific Northwest, USA and comparison to globally
895 distributed representatives. *PLoS One* **16**: e0231367.

896 Wang, N.Y., Zhang, K., Huguet-Tapia, J.C., Rollins, J.A., and Dewdney, M.M. (2016) Mating
897 type and simple sequence repeat markers indicate a clonal population of *Phyllosticta*
898 *citricarpa* in Florida. *Phytopathology* **106**: 1300–1310.

899 Wang, X., Zheng, Z., Cai, Y., Chen, T., Li, C., Fu, W., and Jiang, Y. (2017) CNVcaller: Highly
900 efficient and widely applicable software for detecting copy number variations in large
901 populations. *Gigascience* **6**: 1–12.

902 Watanabe, M., Sato, H., Matsuzaki, H., Kobayashi, T., Sakagami, N., Maejima, Y., et al.
903 (2007) 14C ages and δ13C of sclerotium grains found in forest soils. *Soil Sci Plant Nutr*
904 **53**: 125–131.

905 White, T., Bruns, T., Lee, S., and Taylor, J. (1990) Amplification and direct sequencing of
906 fungal ribosomal RNA genes for phylogenetics. In *PCR protocols: a guide to methods*
907 *and applications*. pp. 315–322.

908 Wickham, H. (2011) ggplot2. *Wiley Interdiscip Rev Comput Stat* **3**: 180–185.

909 Zeiner, C.A., Purvine, S.O., Zink, E.M., Paša-Tolić, L., Chaput, D.L., Haridas, S., et al. (2016)
910 Comparative analysis of secretome profiles of manganese(II)-oxidizing Ascomycete
911 fungi. *PLoS One* **11**:.

912 Zhang, J., Kapli, P., Pavlidis, P., and Stamatakis, A. (2013) A general species delimitation
913 method with applications to phylogenetic placements. *Bioinformatics* **29**: 2869–2876.

914 Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., and Weir, B.S. (2012) A high-
915 performance computing toolset for relatedness and principal component analysis of SNP
916 data. *Bioinformatics* **28**: 3326–3328.

917

Data Sharing and Data Availability

Raw sequence data used in this study are available in NCBI under the Sequence Read Archive identifiers listed in Table S1.

Competing Interests

The authors declare no competing interests.

923 # Tables

924 **Table 1: Details of the *Cenococcum geophilum* individuals analysed in this study.**
925

Strain	*MAT1* locus	Isolate origin	Code	Site	Forest	Isolation year	Latitude [° N]	Longitude [° E]
1.104	*MAT1-2*	Finland	FIN	Punkaharju	*Pinus sylvestris*	20.08.08	61.800000	29.316667
1.076	*MAT1-1*	France	FRA	La Crête	*Fagus sylvatica*	1999	48.111111	5.356944
1.075	*MAT1-1*	France	FRA	La Crête	*Fagus sylvatica*	1999	48.111111	5.356944
1.100	*MAT1-1*	France	FRA	Ventoux	Mixed	30.07.09	44.179497	5.241850
1.094	*MAT1-1*	Poland	POL	Blizyn	Mixed	01.03.09	51.037222	20.869083
1.096	*MAT1-2*	Poland	POL	Blizyn	Mixed	13.08.09	51.037111	20.868694
1.097	*MAT1-1*	Poland	POL	Blizyn	Mixed	13.08.09	51.037222	20.870250
1.022	*MAT1-2*	Switzerland	CHE	Nessental	*Picea abies*	13.06.80	46.722500	8.296667
1.101	*MAT1-2*	Switzerland	CHE	Parabock	*Picea abies*	15.05.08	46.690960	7.236940
1.102	*MAT1-2*	Switzerland	CHE	Parabock	*Picea abies*	10.10.08	46.690960	7.236940
1.057	*MAT1-1*	Switzerland	CHE	Pfynwald	*Pinus sylvestris*	23.04.08	46.301389	7.611944
1.58	*MAT1-1*	Switzerland	CHE	Pfynwald	*Pinus sylvestris*	30.04.08	46.301389	7.611944
1.061	*MAT1-1*	Switzerland	CHE	Pfynwald	*Pinus sylvestris*	30.04.08	46.301389	7.611944
1.064	*MAT1-1*	Switzerland	CHE	Pfynwald	*Pinus sylvestris*	04.09.08	46.301389	7.611944
1.060	*MAT1-1*	Switzerland	CHE	Pfynwald	*Pinus sylvestris*	30.04.08	46.301389	7.611944
1.005	*MAT1-2*	Switzerland	CHE	Strada	*Picea abies*	03.11.77	46.767500	9.178889

926

927 **Table 2: Characteristics of the 20 top-candidate genes putatively under selection in the**
928 **ref. and non-ref. cluster based on Tajima's *D* statistics.** Estimation of the nucleotide
929 diversity per site is shown for each candidate gene.
930

Dataset	Protein ID	Predicted function	Multi-gene family			*D*	*π*
			Size (No. proteins in Cenge)	Size evolution (CAFE)	Conservation (in Cenge, Glost, Leppa)		
Ref. cluster	166793	Signal transduction mechanisms	4			-0.87	0.00851
	738677	Self/non-self recognition (NACHT domain)	66	expanded		-0.869	0.00765
	701966	Cytoskeleton	1			-0.869	0.00736
	486968	Self/non-self recognition (HET, Ankyrin)	97		over-represented	-0.868	0.01591
	700681	Translation, ribosomal structure and biogenesis	1		unique	-0.868	0.00427
	665335	Chromatin structure and dynamics	2		unique	-0.867	0.00925
	552403	Self/non-self recognition (NACHT domain)	56	expanded		-0.867	0.01231
	608632	Self/non-self recognition (NACHT domain)	14			-0.867	0.00503
	353610	Nucleotide transport and metabolism	2			-0.866	0.00933
	222924	-	1		unique	-0.866	0.00612
	653800	Cytoskeleton	6			-0.865	0.01017
	702114	Transcription	2			-0.865	0.00546
	605605	CAZyme (Glycoside Hydrolase 35)	2	contracted		-0.865	0.00269
	597082	Energy production and conversion	1			-0.864	0.00768
	672908	Chromatin structure and dynamics	1		unique	-0.864	0.00595
	487620	-	4			-0.864	0.00901
	674647	Transcription	1			-0.864	0.00227
	661861	RNA processing and modification	1			-0.863	0.00408
	637860	Replication, recombination and repair	1			-0.863	0.00517
	237852	-	1			-0.863	0.00686
Non-ref. cluster	675988	-	1			-1.971	0.00055
	144717	Secondary metabolism (DMAT cluster, decorating ABC transporter)	6	contracted		-1.955	0.00019
	682552	Methyltransferase	1			-1.884	0.00072
	343761	Protein binding	5			-1.853	0.00128
	650216	Secondary metabolism	5			-1.852	0.00049
	650873	Transcription	1			-1.849	0.00042
	674343	Secondary metabolism	9			-1.844	0.0003
	676801	Secondary metabolism	1			-1.832	0.00064
	698845	Cell wall/membrane/envelope biogenesis	1			-1.831	0.00064
	202303	Secondary metabolism (DMAT cluster, DMAT backbone gene)	3			-1.804	0.00098
	672500	CAZyme (LPMO AA11)	6			-1.803	0.0005
	691644	Amino acid transporter	33			-1.777	0.00094
	650097	-	1			-1.768	0.00048

671949	Posttranslational modification, protein turnover, chaperones	15	-1.767	0.00077
707138	Zinc finger, mRNA export	1	-1.766	0.00029
366609	RNA processing and modification	2	-1.763	0.00076
563055	RNA processing and modification	1	-1.734	0.00082
653324	-	1	-1.731	0.00032
706589	Secondary metabolism (Beta-lactamase, response to antibiotic)	4	-1.714	0.0006
663021	-	2	-1.709	0.00234

Note: Protein Id: JGI Cenge3 portal; Predicted function: KOG class (EuKaryotic Orthologous Groups) and other: detailed prediction see Table S4; Multi-gene family: JGI MCL Cluster Run *Cenococcum* + 50 genomes 03/21/14.399 (https://mycocosm.jgi.doe.gov/clm/run/cenococcum-and-50-genomes.399; Ohm *et al.*, 2012; Peter *et al.*, 2016) Conservation: protein orthologues only present (unique) or over-represented ($P \leq 0.05$; Wilcoxon test, two-sided, Benjamini-Hochberg multiple testing correction) in group Cenge (*Cenococcum geophilum*), Glost (*Glonium stellatum*), Leppa (*Lepidopterella palustris*) as compared to other Dothideomycetes (outgroup); Tajima's *D* (*D*); nucleotide diversity per site (π)

Figures

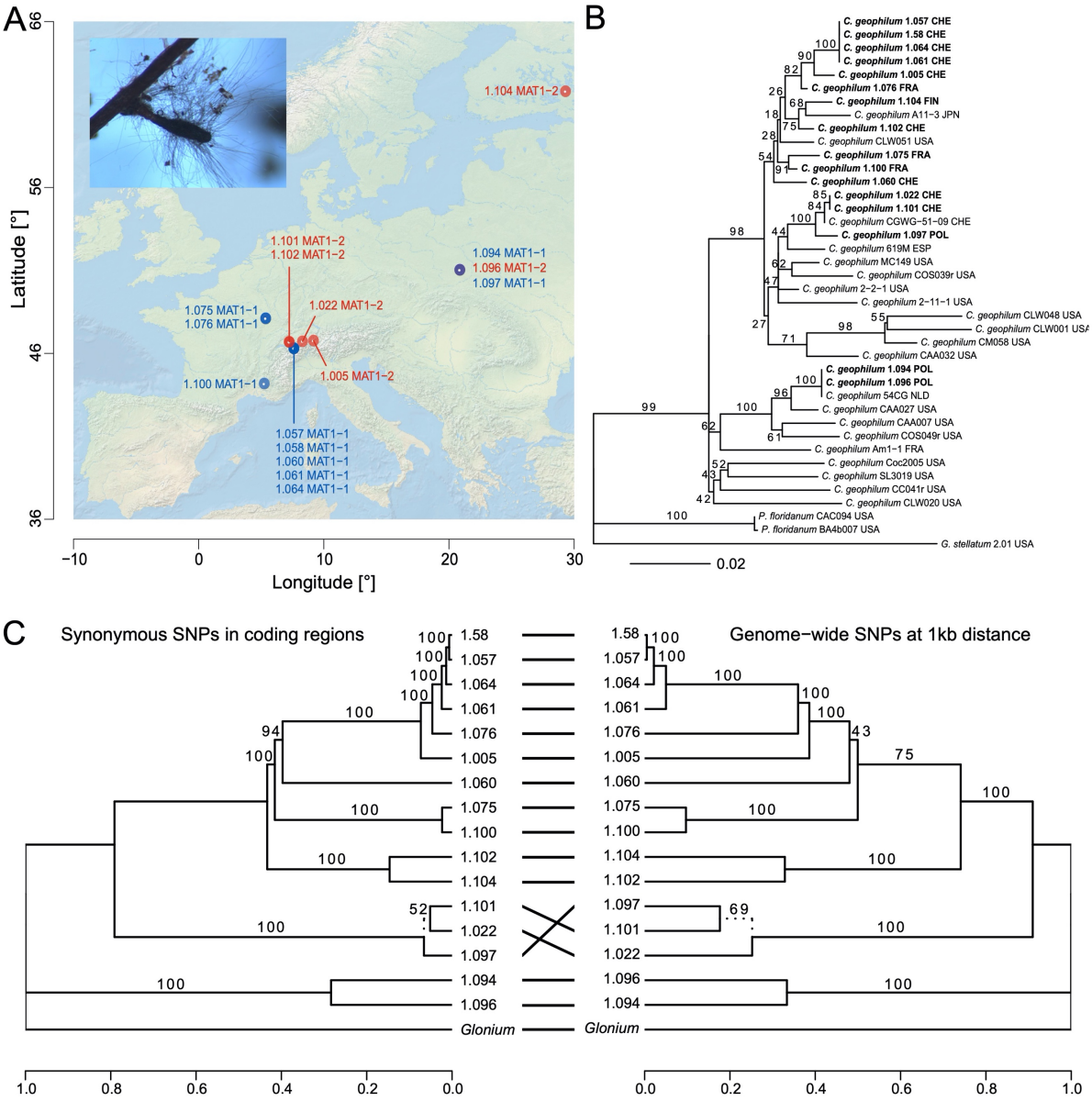
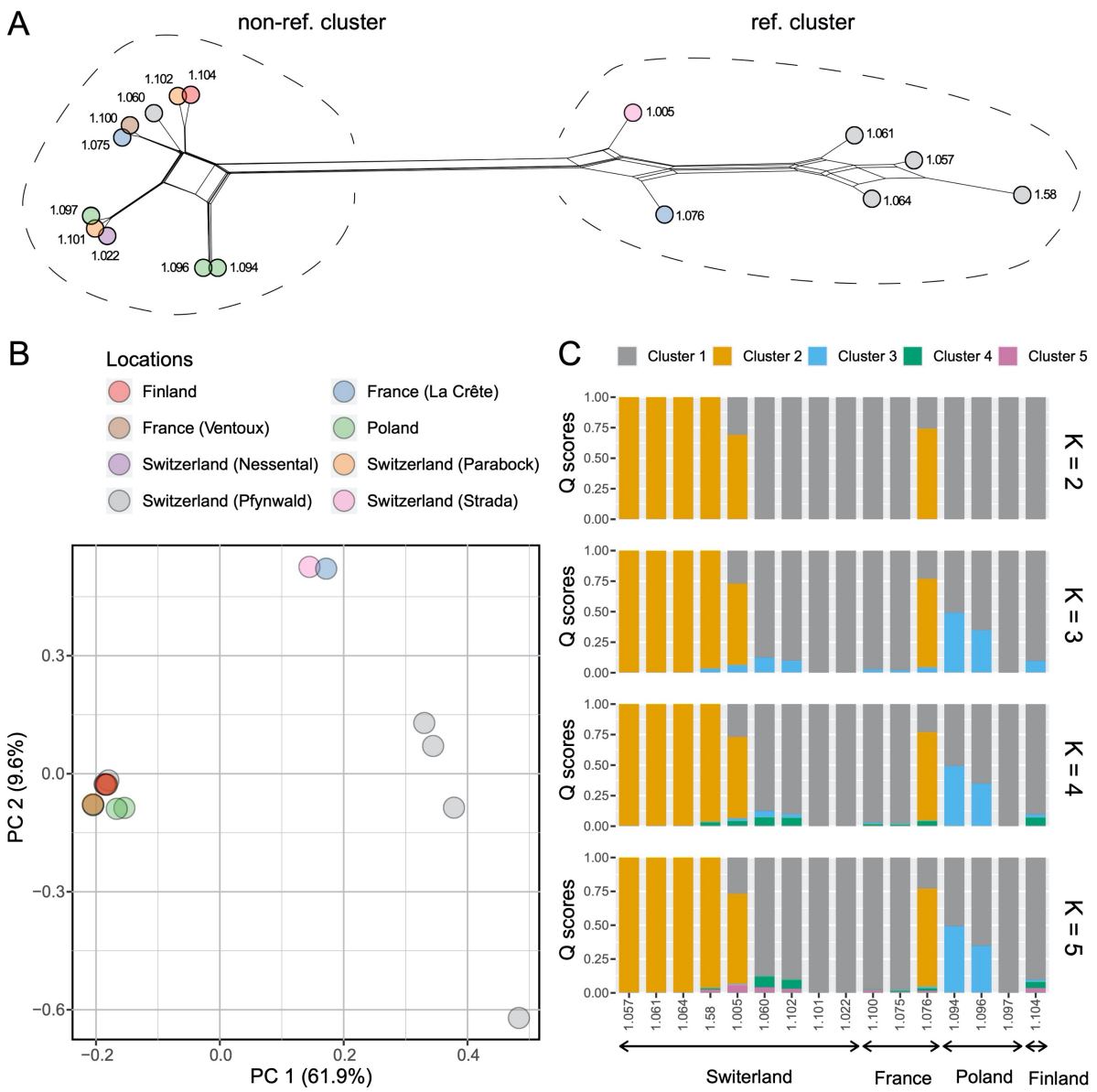


Figure 1: Study range, phylogenetic relationships, and genomic relatedness of the *Cenococcum geophilum* strains sampled. A) Sampling sites of the *C. geophilum* isolates with an illustration of the ecto-mycorrhiza morphology (Photo: M. Peter). Geographic distribution of MAT loci is given for each sampled location. **B**) Phylogenetic relationships within the genus *Cenococcum* based on two nuclear loci, i.e., ITS and GAPDH. One outgroup species, i.e., *Glonium stellatum,* was used to root the phylogenetic tree. The scale bar shows the substitution rate. Tips in bold are the 16 samples analysed in this study using whole-genome sequencing data. **C**) Genomic relatedness among the *C. geophilum* sampled, including the reference genome (strain 1.58). Whole-genome SNP data capture both repetitive and coding regions while synonymous SNPs only reflect neutral divergence in coding regions. Topologies are rooted with *Glonium stellatum* as outgroup. Branch supports represent bootstrap values from the maximum likelihood tree reconstruction. Strain codes are described in Table S1.



953

954 **Figure 2: Recombination analysis and population genetic differentiation among the**
955 ***Cenococcum geophilum* strains sampled. A**) Network-based phylogenetic reconstruction
956 supporting recombination events in one of the two main lineages. **B**) Principal component (PC)
957 analysis showing the genomic variation without clear congruence with geography. **C**) Ancestry
958 inferences for $K = 2$–5 indicating some admixed genotypes between the major genetic groups.
959 Strain codes are described in Table S1. Colours refer to the geographic origin and genetic group
960 of each strain for A–B and C, respectively.

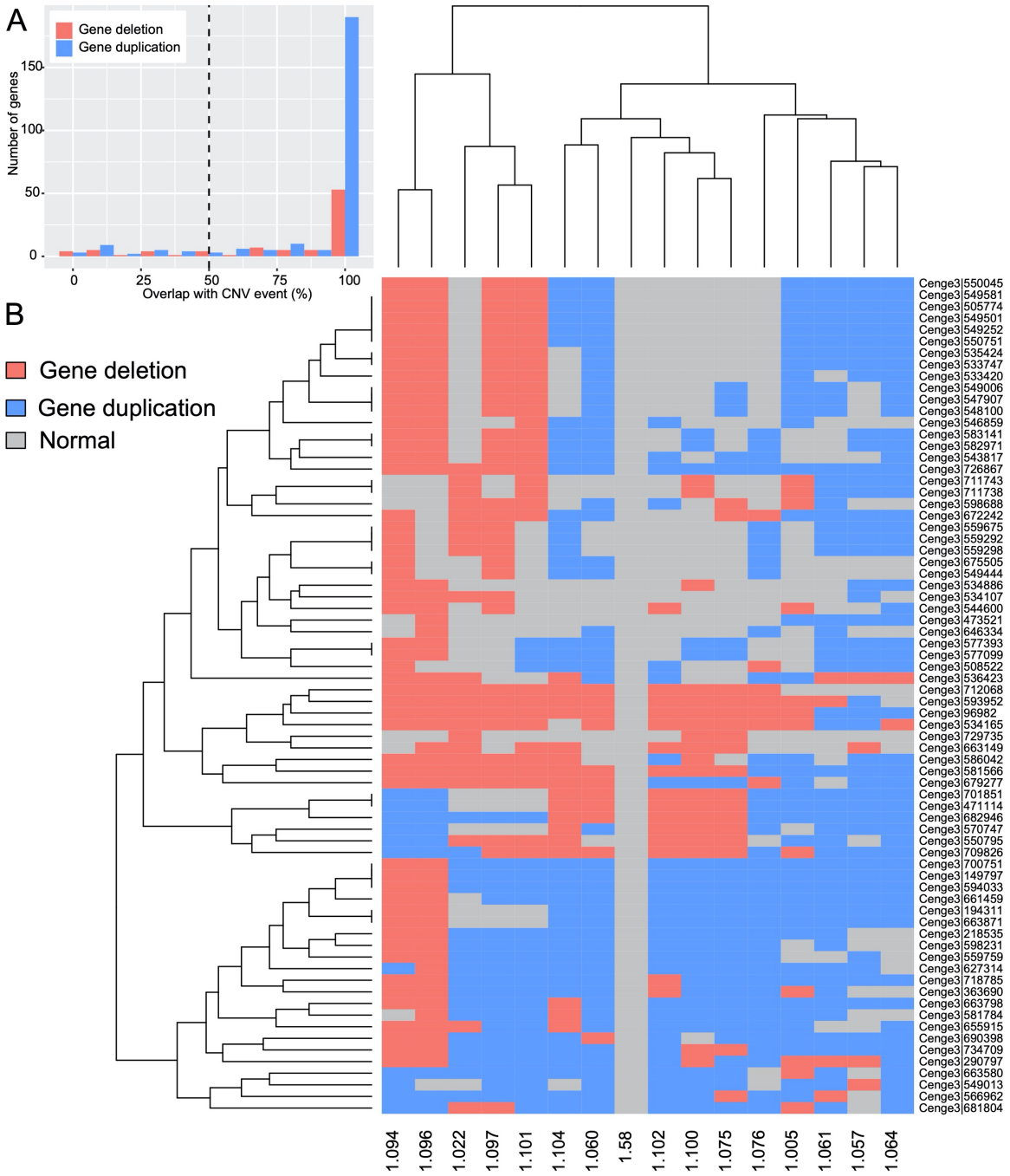


Figure 3: Analysis of copy-number variation (CNV) regions among the sampled *Cenococcum geophilum* strains. A) Gene content including CNV with the percentage of length covered by deletion or duplication >50%. Deletions refer to any gene locus showing at least one deletion in a strain while duplications include any gene locus showing only duplication or normal copy-numbers without deletion. **B**) Gene comparison affected by deletion or duplication among *C. geophilum* strains. Tree topologies on the left and top show the grouping according to gene CNVs. The y-axis shows gene identifiers of CNV-affected genes. Note that the strain 1.58 displays only single-copy genes because the CNV calls were filtered to contain only those genes that were confirmed as single-copy from the reference genome. Strain codes are described in Table S1.

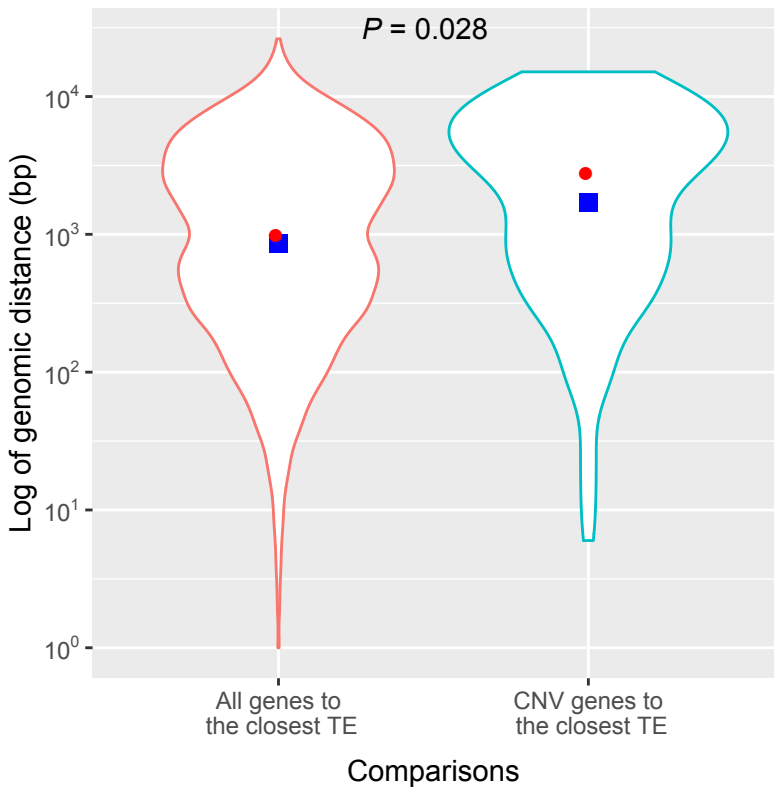


Figure 4: Distribution of genomic distances between all genes and their closest transposable elements (TEs), and between copy-number variation (CNV) genes and their closest TEs. The significance between the median values was calculated using the Wilcoxon test. The red dots and blue squares represent mean and median values, respectively.

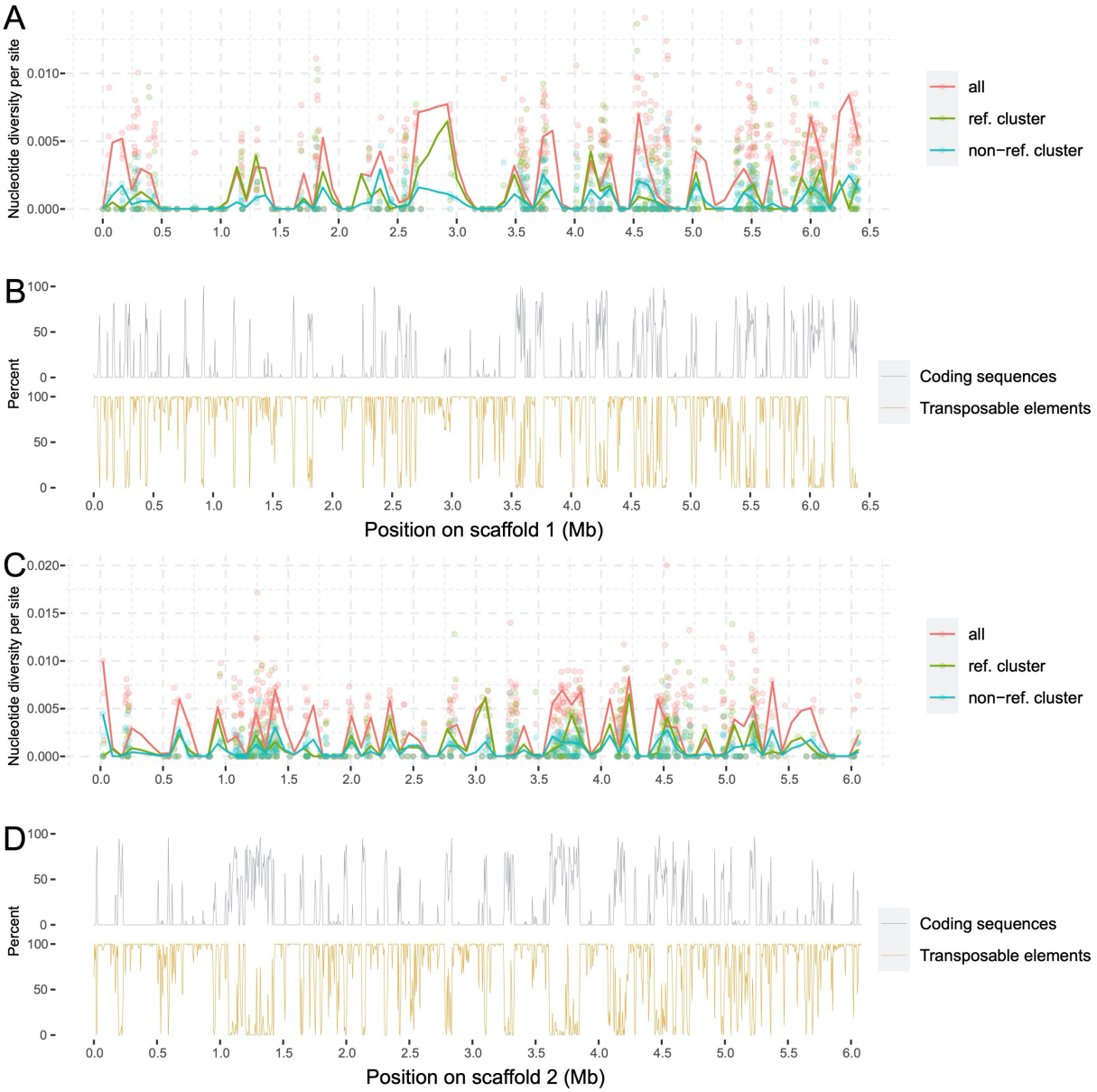
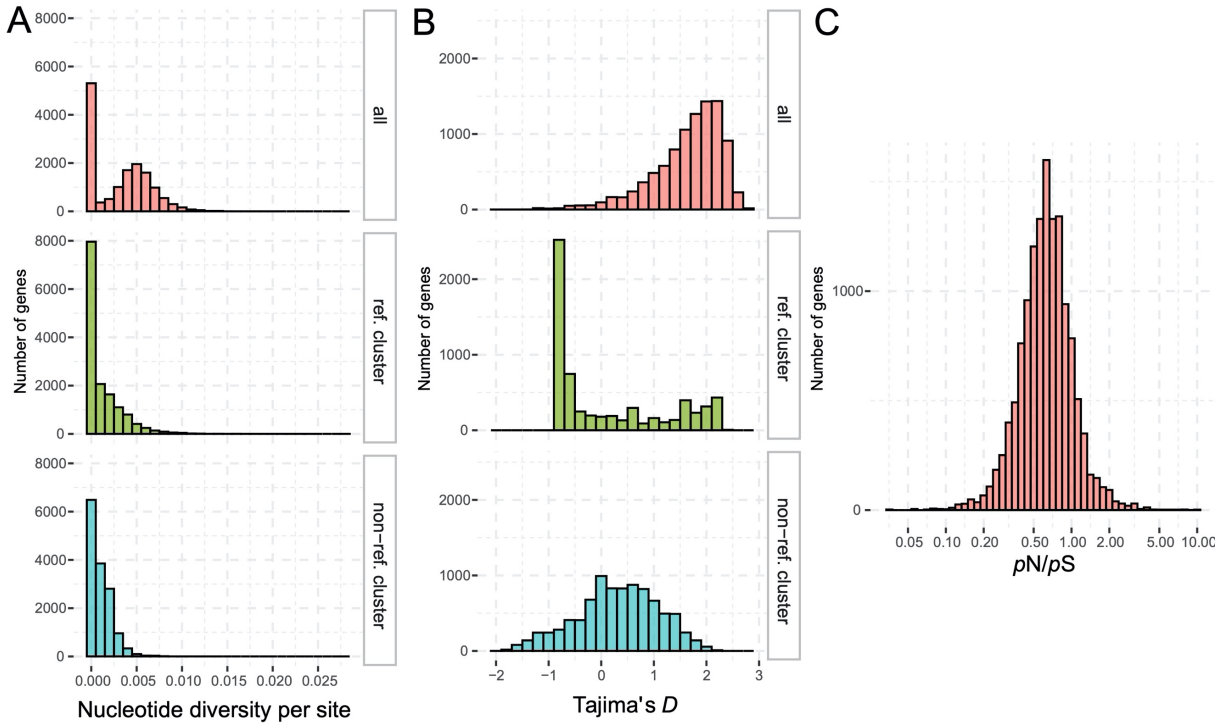


Figure 5: Genetic diversity analysis and coding sequence variation along scaffolds #1 and #2 in *Cenococcum geophilum* strains used in this study. A) and **C**) Nucleotide diversity per site is given for each gene identified in the 1.58 genome assembly. **B**) and **D**) Density of gene content and transposable element is presented. Results are shown for the complete dataset ("all") and subsets ("ref. cluster" and "non-ref. cluster") of *Cenococcum geophilum* strains used in this study (see section Experimental Procedures).



984

985 **Figure 6: Within-species genetic diversity and signatures of selection. A**) Distribution of
986 nucleotide diversity per site for each gene. **B**) Distribution of gene-specific Tajima's *D*. **C**)
987 Distribution of the gene-specific ratio of non-synonymous to synonymous polymorphisms
988 (*p*N/*p*S). Results are shown for the complete dataset ("all") and subsets ("ref. cluster" and "non-
989 ref. cluster") of *Cenococcum geophilum* strains used in this study (see section Experimental
990 Procedures). Note that the reference cluster is genetically more diverse than the non-ref. cluster.

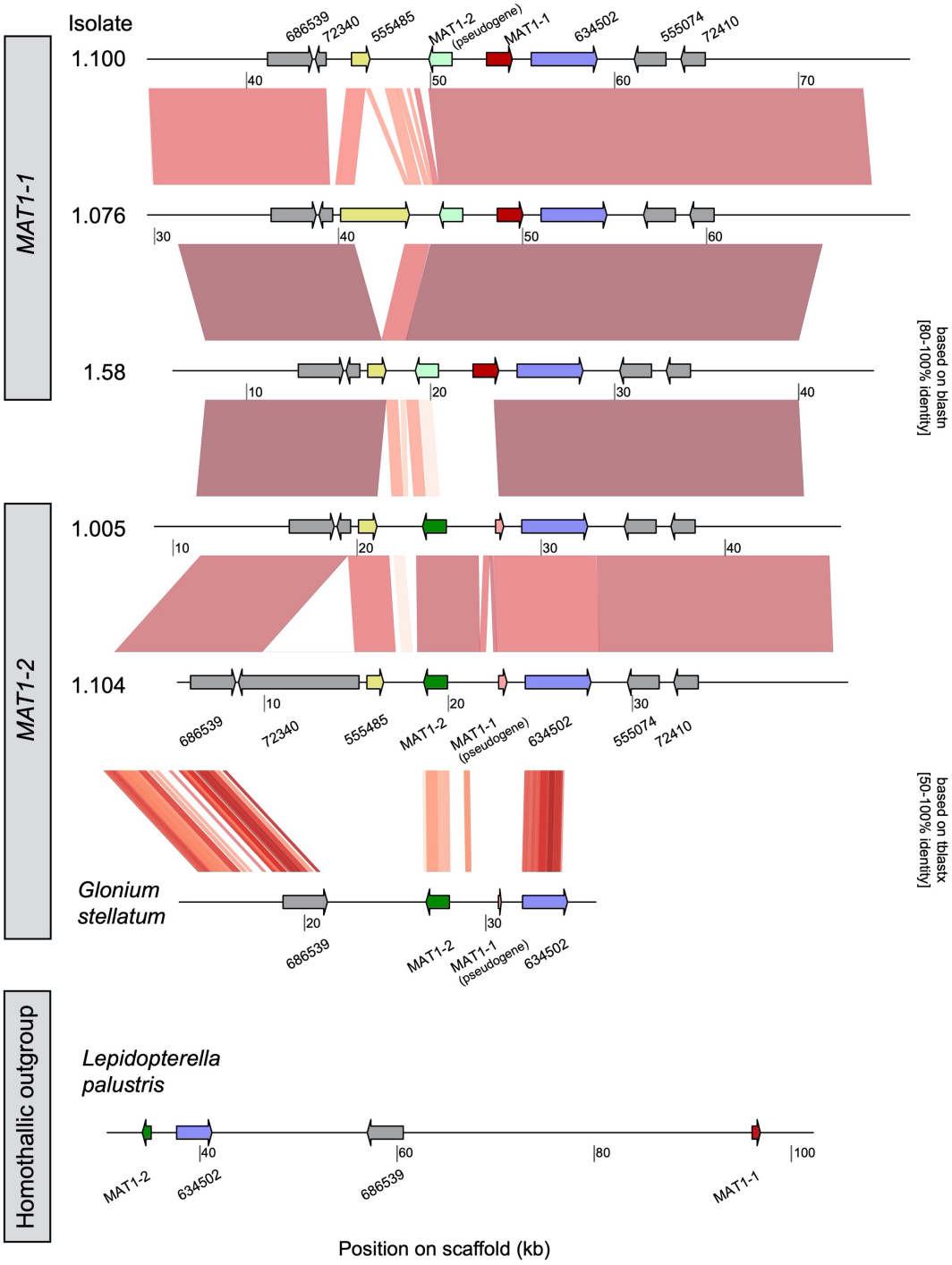
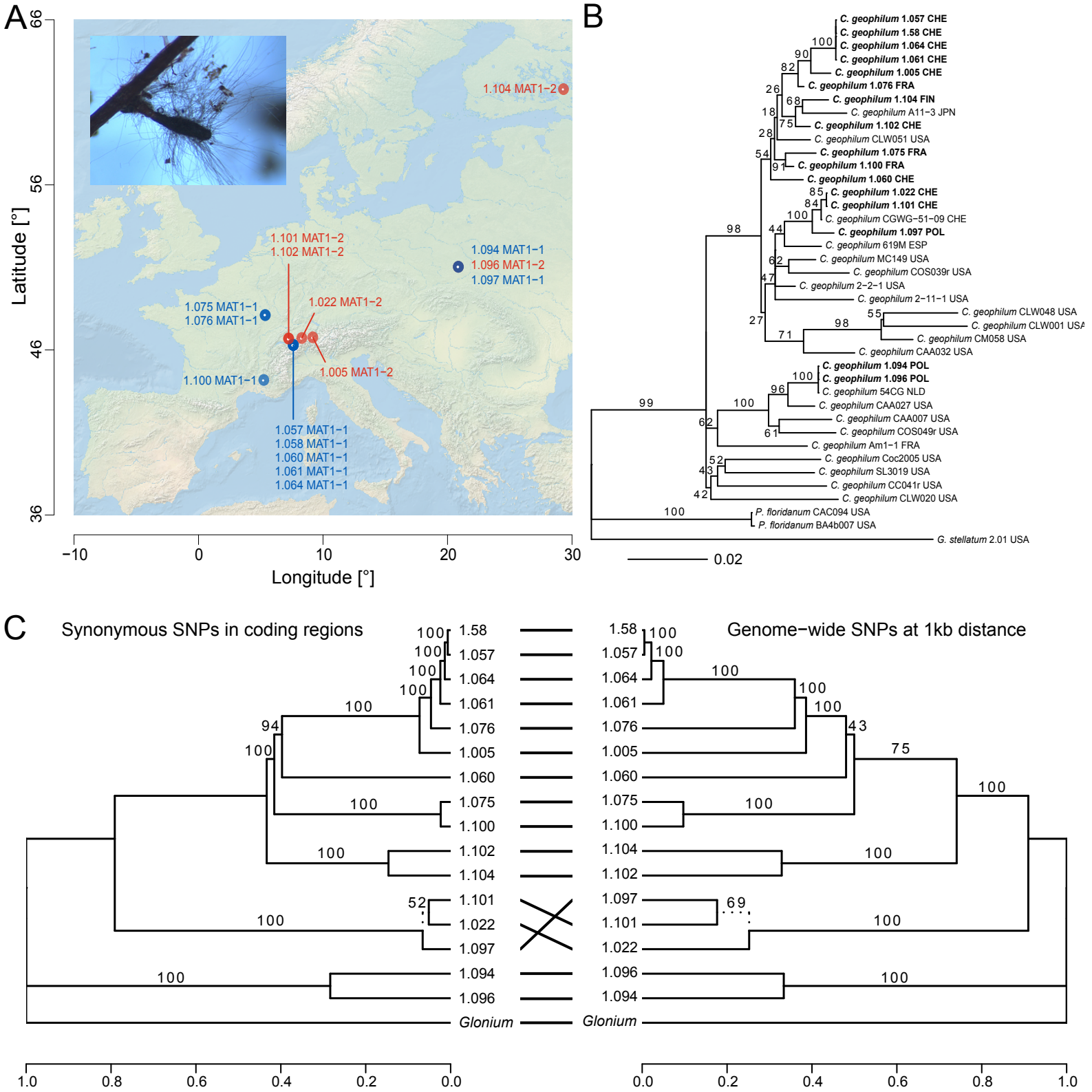
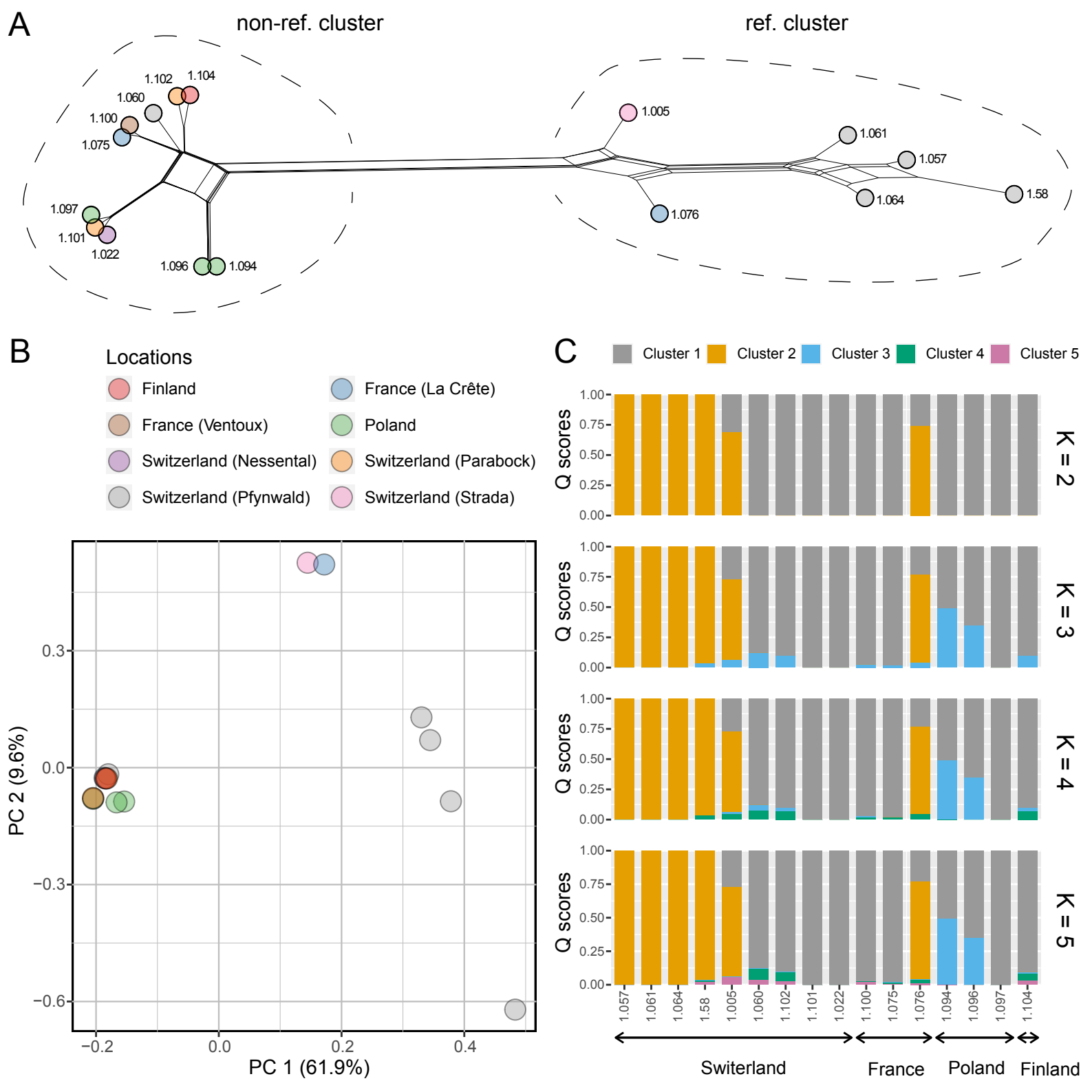
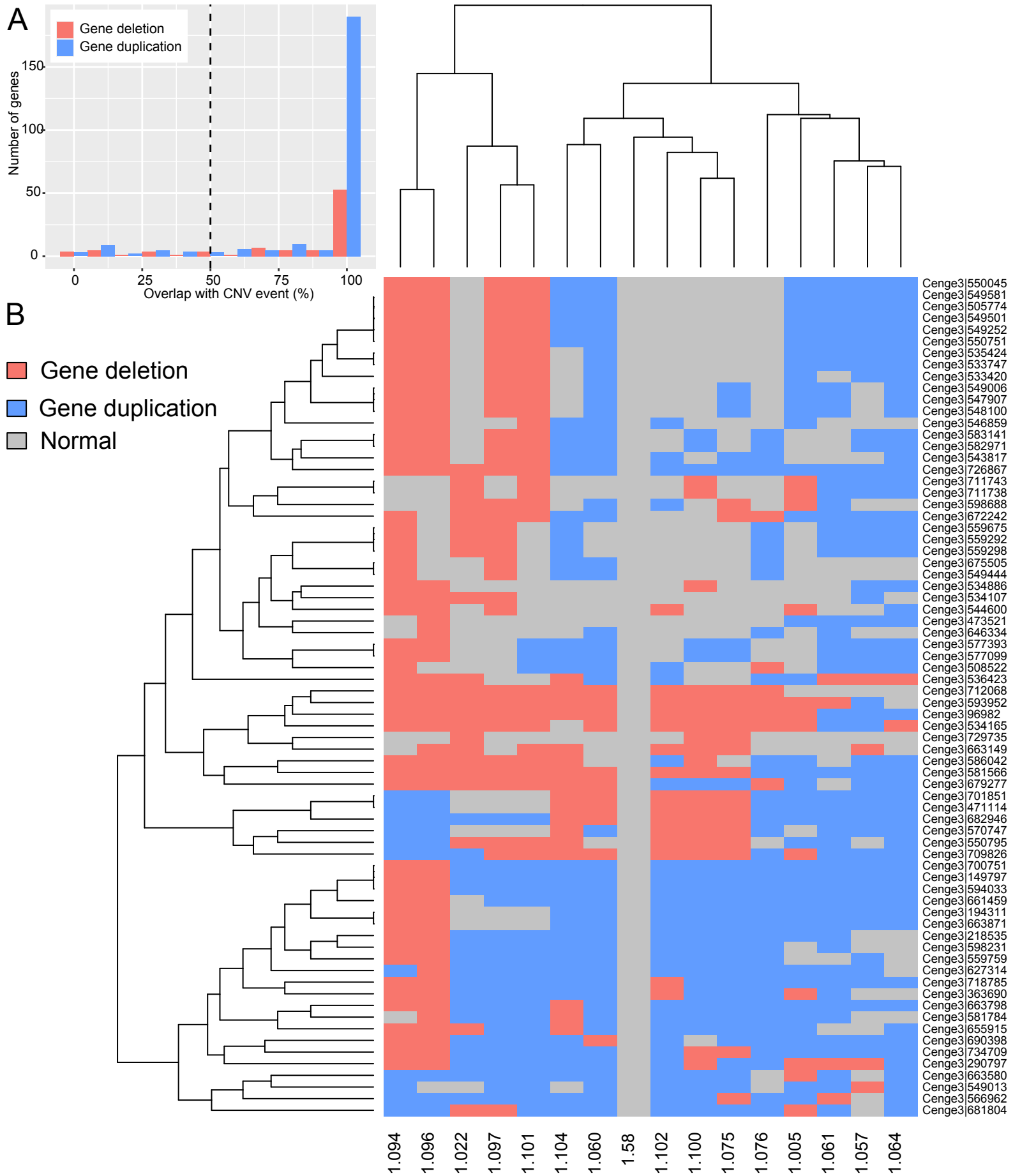
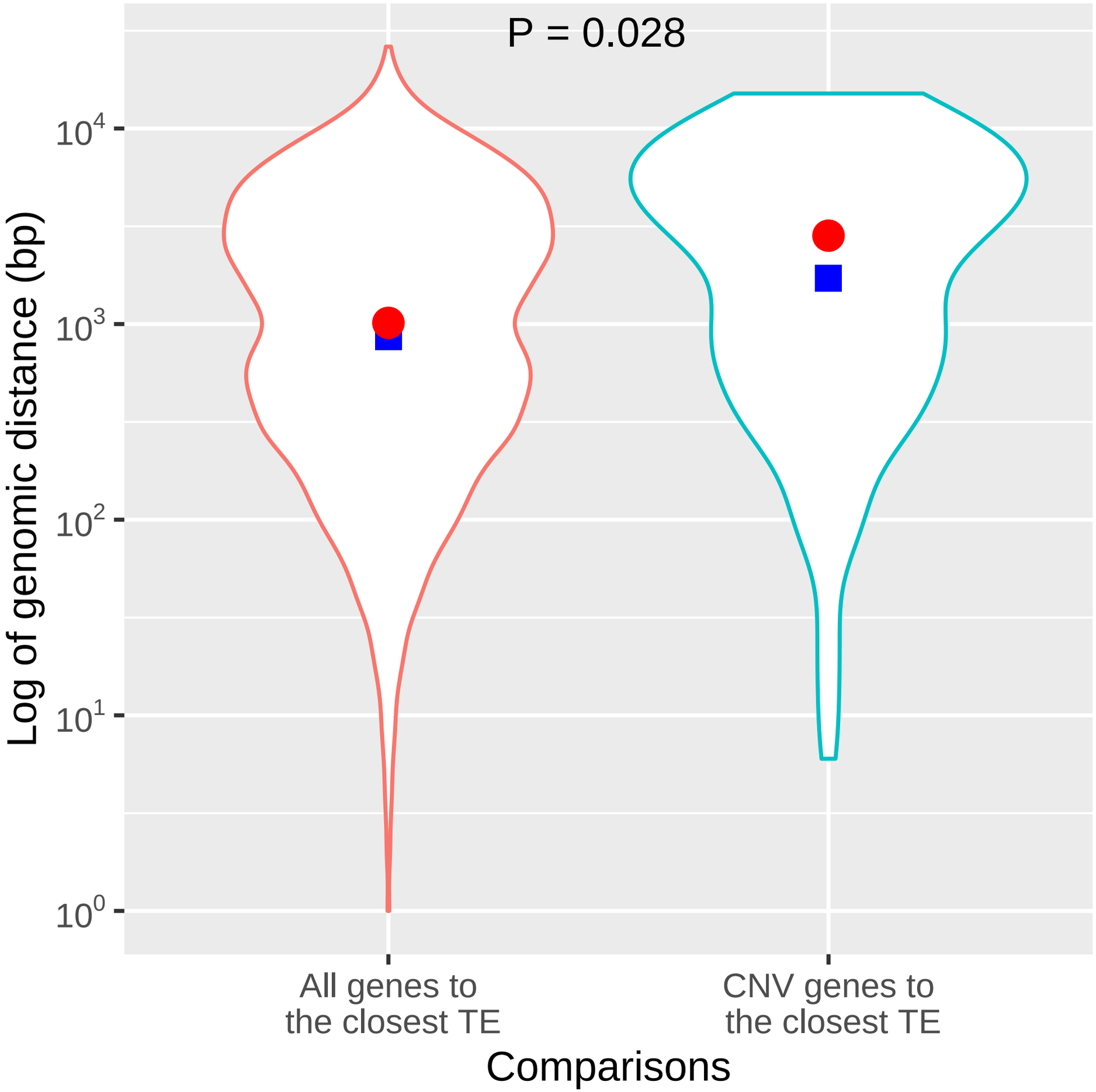


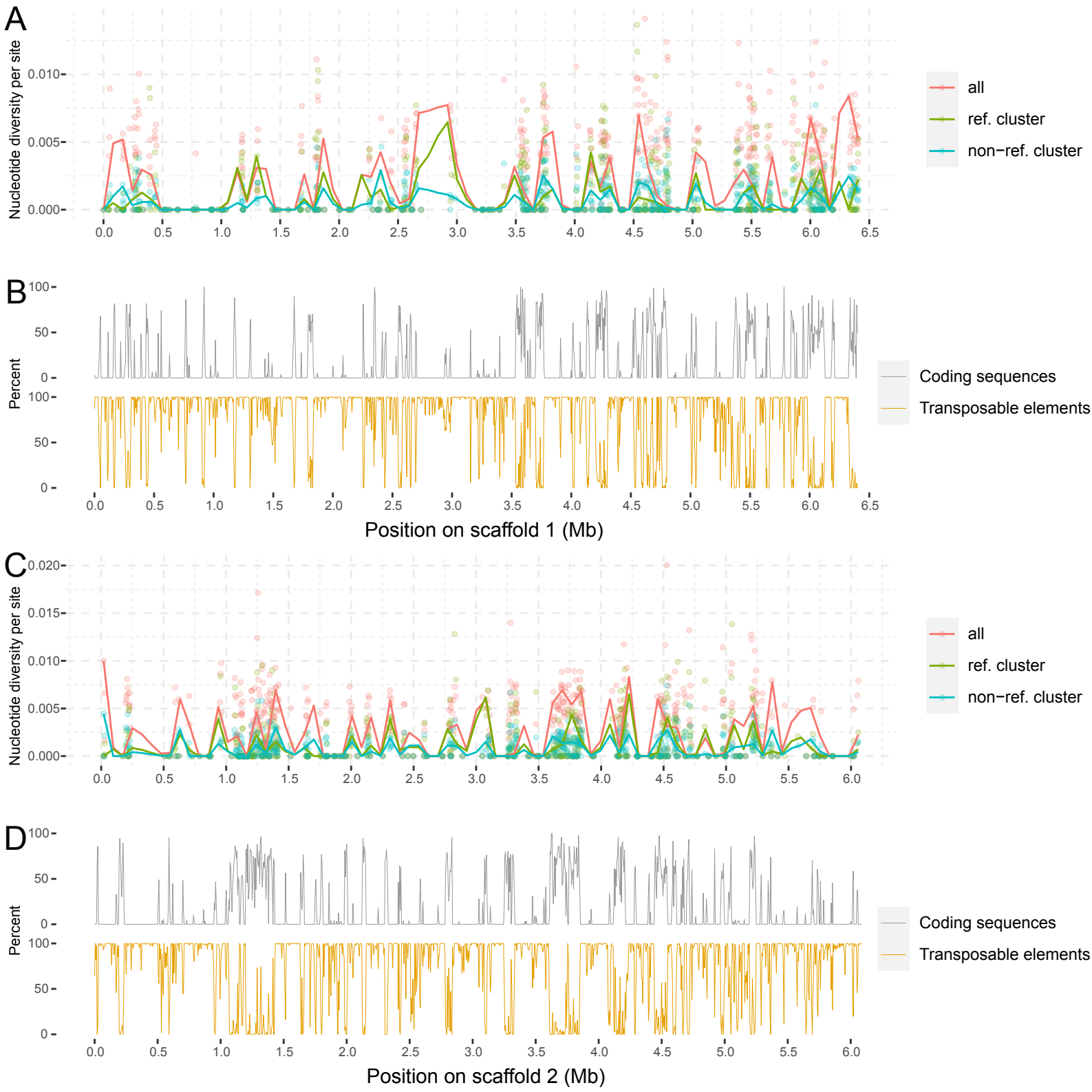
Figure 7: Chromosomal synteny analyses of *MAT1-1* and *MAT1-2* loci. Intraspecific analyses were performed with the reference genome strain (1.58) and interspecific comparisons were carried out between *G. stellatum*, *L. palustris* and *C. geophilum* strains. Colour gradients (white to red) correspond to sequence similarities ranging from 50–100%, with a minimum bit-score of 100.

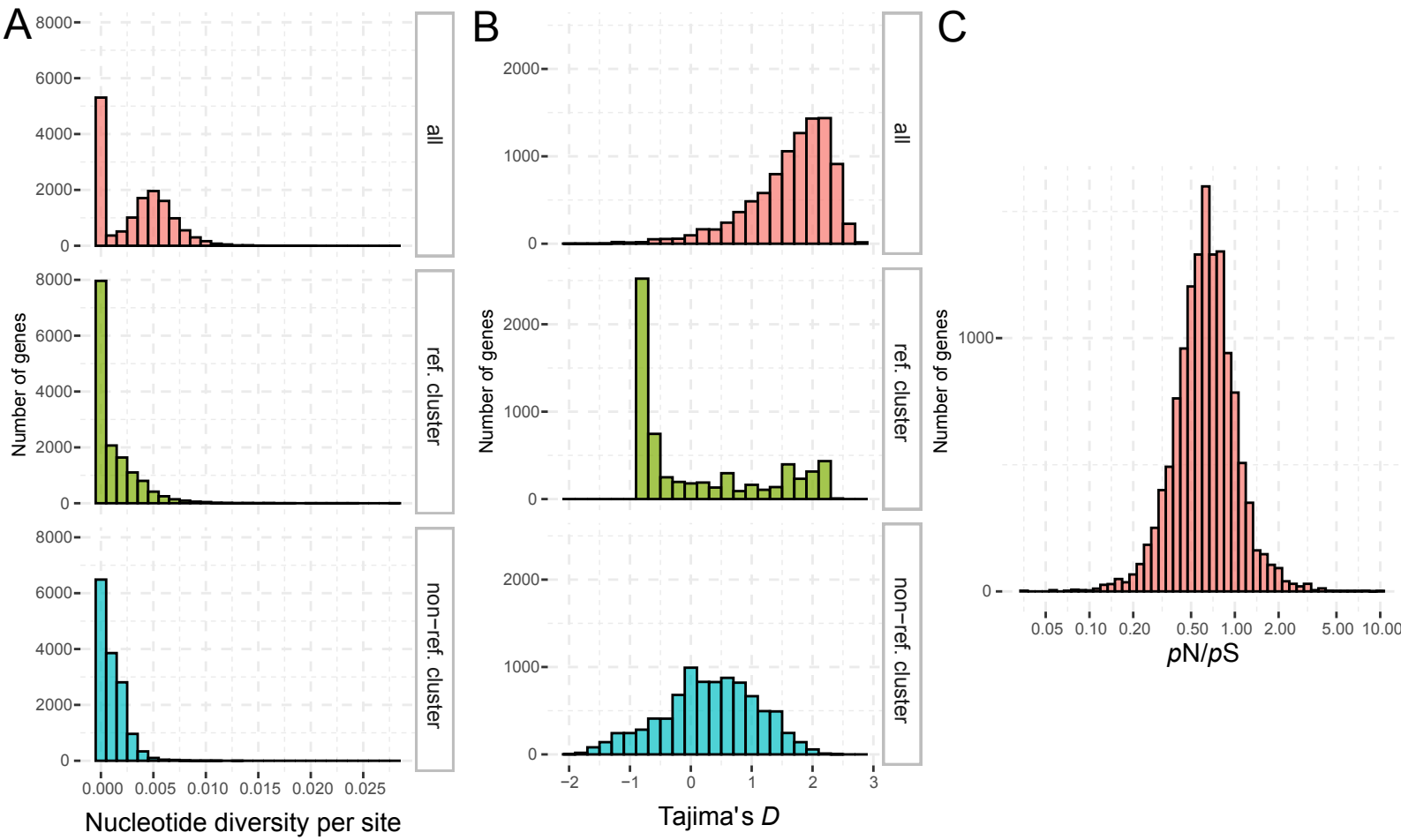

A
Latitude [°]
Longitude [°]
1.104 MAT1−2
1.101 MAT1−2
1.102 MAT1−2
1.094 MAT1−1
1.096 MAT1−2
1.097 MAT1−1
1.075 MAT1−1
1.076 MAT1−1
1.022 MAT1−2
1.005 MAT1−2
1.100 MAT1−1
1.057 MAT1−1
1.058 MAT1−1
1.060 MAT1−1
1.061 MAT1−1
1.064 MAT1−1
B
C. geophilum 1.057 CHE
C. geophilum 1.58 CHE
C. geophilum 1.064 CHE
C. geophilum 1.061 CHE
C. geophilum 1.005 CHE
C. geophilum 1.076 FRA
C. geophilum 1.104 FIN
C. geophilum A11−3 JPN
C. geophilum 1.102 CHE
C. geophilum CLW051 USA
C. geophilum 1.075 FRA
C. geophilum 1.100 FRA
C. geophilum 1.060 CHE
C. geophilum 1.022 CHE
C. geophilum 1.101 CHE
C. geophilum CGWG−51−09 CHE
C. geophilum 1.097 POL
C. geophilum 619M ESP
C. geophilum MC149 USA
C. geophilum COS039r USA
C. geophilum 2−2−1 USA
C. geophilum 2−11−1 USA
C. geophilum CLW048 USA
C. geophilum CLW001 USA
C. geophilum CM058 USA
C. geophilum CAA032 USA
C. geophilum 1.094 POL
C. geophilum 1.096 POL
C. geophilum 54CG NLD
C. geophilum CAA027 USA
C. geophilum CAA007 USA
C. geophilum COS049r USA
C. geophilum Am1−1 FRA
C. geophilum Coc2005 USA
C. geophilum SL3019 USA
C. geophilum CC041r USA
C. geophilum CLW020 USA
P. floridanum CAC094 USA
P. floridanum BA4b007 USA
G. stellatum 2.01 USA
0.02
C
Synonymous SNPs in coding regions
Genome−wide SNPs at 1kb distance
1.58
1.057
1.064
1.061
1.076
1.005
1.060
1.075
1.100
1.102
1.104
1.101
1.022
1.097
1.094
1.096
Glonium

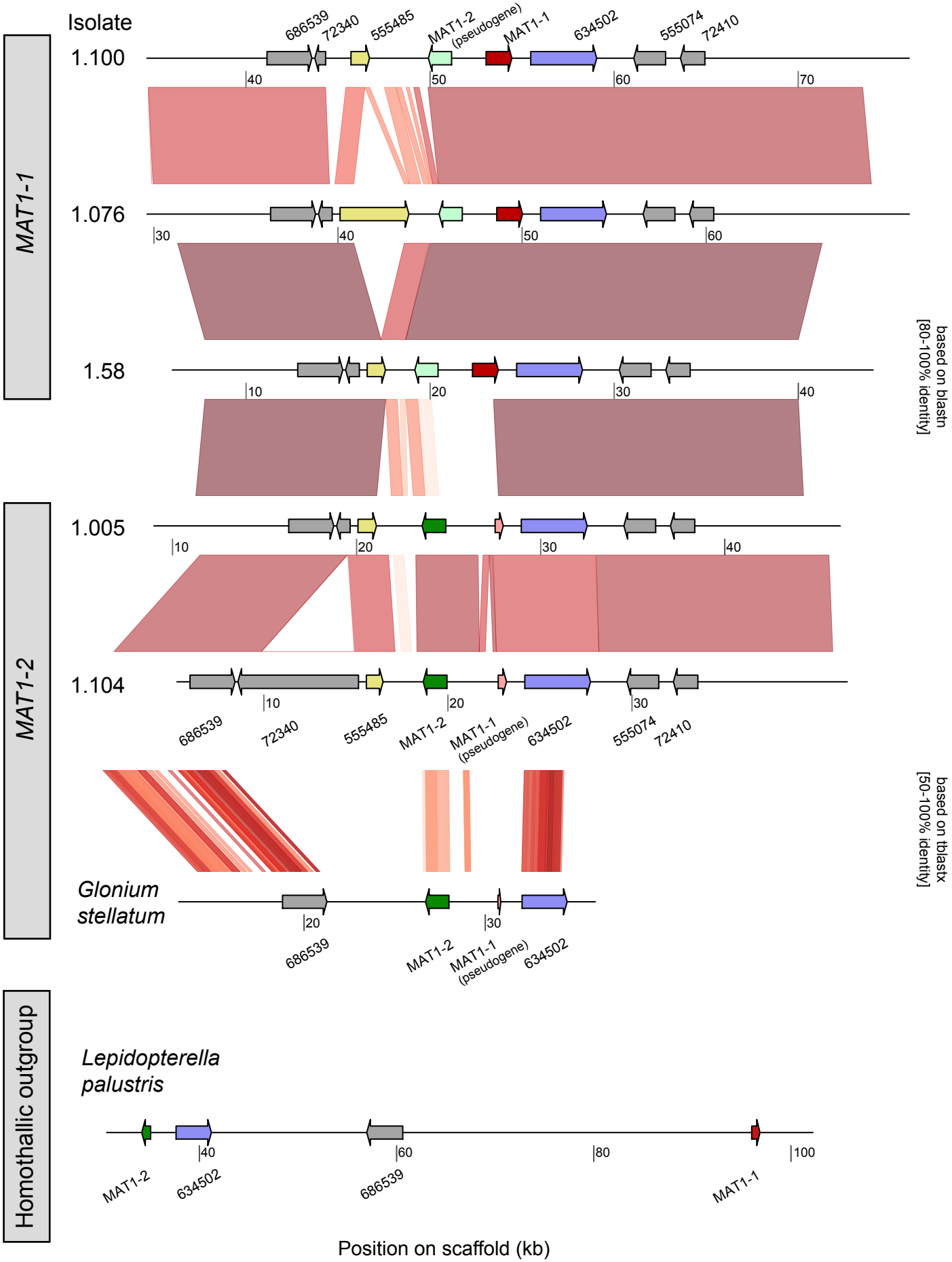

A
non-ref. cluster
ref. cluster
1.102
1.104
1.060
1.100
1.075
1.097
1.101
1.022
1.096
1.094
1.005
1.061
1.057
1.064
1.58
1.076
B
Locations
Finland
France (La Crête)
France (Ventoux)
Poland
Switzerland (Nessental)
Switzerland (Parabock)
Switzerland (Pfynwald)
Switzerland (Strada)
PC 2 (9.6%)
PC 1 (61.9%)
0.3
0.0
−0.3
−0.6
−0.2
0.0
0.2
0.4
C
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Q scores
1.00
0.75
0.50
0.25
0.00
K = 2
K = 3
K = 4
K = 5
1.057
1.061
1.064
1.58
1.005
1.060
1.102
1.101
1.022
1.100
1.075
1.076
1.094
1.096
1.097
1.104
Switerland
France
Poland
Finland


A
Gene deletion
Gene duplication
Number of genes
150
100
50
0
0
25
50
75
100
Overlap with CNV event (%)
B
Gene deletion
Gene duplication
Normal
Cenge3|550045
Cenge3|549581
Cenge3|505774
Cenge3|549501
Cenge3|549252
Cenge3|550751
Cenge3|535424
Cenge3|533747
Cenge3|533420
Cenge3|549006
Cenge3|547907
Cenge3|548100
Cenge3|546859
Cenge3|583141
Cenge3|582971
Cenge3|543817
Cenge3|726867
Cenge3|711743
Cenge3|711738
Cenge3|598688
Cenge3|672242
Cenge3|559675
Cenge3|559292
Cenge3|559298
Cenge3|675505
Cenge3|549444
Cenge3|534886
Cenge3|534107
Cenge3|544600
Cenge3|473521
Cenge3|646334
Cenge3|577393
Cenge3|577099
Cenge3|508522
Cenge3|536423
Cenge3|712068
Cenge3|593952
Cenge3|96982
Cenge3|534165
Cenge3|729735
Cenge3|663149
Cenge3|586042
Cenge3|581566
Cenge3|679277
Cenge3|701851
Cenge3|471114
Cenge3|682946
Cenge3|570747
Cenge3|550795
Cenge3|709826
Cenge3|700751
Cenge3|149797
Cenge3|594033
Cenge3|661459
Cenge3|194311
Cenge3|663871
Cenge3|218535
Cenge3|598231
Cenge3|559759
Cenge3|627314
Cenge3|718785
Cenge3|363690
Cenge3|663798
Cenge3|581784
Cenge3|655915
Cenge3|690398
Cenge3|734709
Cenge3|290797
Cenge3|663580
Cenge3|549013
Cenge3|566962
Cenge3|681804
1.094
1.096
1.022
1.097
1.101
1.104
1.060
1.58
1.102
1.100
1.075
1.076
1.005
1.061
1.057
1.064


P = 0.028
Log of genomic distance (bp)
10^4
10^3
10^2
10^1
10^0
All genes to
the closest TE
CNV genes to
the closest TE
Comparisons


A
Nucleotide diversity per site
0.010
0.005
0.000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
all
ref. cluster
non−ref. cluster
B
Percent
100
50
0
Coding sequences
Transposable elements
Position on scaffold 1 (Mb)
C
Nucleotide diversity per site
0.020
0.015
0.010
0.005
0.000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
all
ref. cluster
non−ref. cluster
D
Percent
100
50
0
Coding sequences
Transposable elements
Position on scaffold 2 (Mb)


A
B
C
all
ref. cluster
non−ref. cluster
Number of genes
Nucleotide diversity per site
Tajima's D
pN/pS
0.000 0.005 0.010 0.015 0.020 0.025
−2 −1 0 1 2 3
0.05 0.10 0.20 0.50 1.00 2.00 5.00 10.00


Isolate
MAT1-1
1.100
686539
72340
555485
MAT1-2 (pseudogene)
MAT1-1
634502
555074
72410
40
50
60
70
1.076
30
40
50
60
1.58
10
20
30
40
based on blastn [80-100% identity]
MAT1-2
1.005
10
20
30
40
1.104
10
20
30
686539
72340
555485
MAT1-2
MAT1-1 (pseudogene)
634502
555074
72410
based on tblastx [50-100% identity]
Glonium stellatum
20
30
686539
MAT1-2
MAT1-1 (pseudogene)
634502
Homothallic outgroup
Lepidopterella palustris
40
60
80
100
MAT1-2
634502
686539
MAT1-1
Position on scaffold (kb)