



HAL
open science

Specific populations of the yeast *Geotrichum candidum* revealed by molecular typing

Noémie Jacques, Sandrine Mallet, Fatima Laaghouti, Colin R. Tinsley, Serge
Casaregola

► **To cite this version:**

Noémie Jacques, Sandrine Mallet, Fatima Laaghouti, Colin R. Tinsley, Serge Casaregola. Specific populations of the yeast *Geotrichum candidum* revealed by molecular typing. *Yeast*, 2017, 34 (4), pp.165-178. 10.1002/yea.3223 . hal-03665115

HAL Id: hal-03665115

<https://agroparistech.hal.science/hal-03665115>

Submitted on 11 May 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

1 **Specific populations of the yeast *Geotrichum candidum* revealed by**
2 **molecular typing**

3
4 **Noémie Jacques[#], Sandrine Mallet[#], Fatima Laaghouti, Colin R. Tinsley**
5 **and Serge Casaregola^{*}**

6
7 Micalis Institute, INRA, AgroParisTech, CIRM-Levures, Université Paris-Saclay, 78350
8 Jouy-en-Josas, France.

9
10 ^{*}Corresponding author: serge.casaregola@grignon.inra.fr

11
12 **Running title:** Specific populations of *Geotrichum candidum*

13 **Abbreviations:** MLST, Multi Locus Sequence Typing, Sequence Type, ST; RAPD, Random
14 Amplification of Polymorphic DNA; LTR, Long Terminal Repeat; CLP, Chromosome
15 Length Polymorphism

16
17
18 The GenBank/EMBL/DDBJ accession numbers for the sequences determined in this study
19 are: HF558450 for *URA1* gene, HF558451 for *URA3* gene, HF558452 for *NUP116* gene,
20 HF558453 *SAPT4 (YSP3)* gene, HF586414 for *PLB3* gene and HF558454 for the retro-
21 transposon LTR. The *G. candidum* MLST database can be accessed online at
22 <http://pubmlst.org/geotrichum/>

23
24 [#] These authors participated equally to the work
25

26 **Abstract**

27 *Geotrichum candidum* is an ubiquitous yeast, and an essential component in the production of
28 many soft cheeses. We developed a Multi-Locus Sequence Typing (MLST) scheme with five
29 retained loci (*NUP116*, *URA1*, *URA3*, *SAPT4*, *PLB3*) which were sufficiently divergent to
30 distinguish 40 Sequence Types (STs) among the 67 *G. candidum* strains tested. Phylogenetic
31 analyses defined five main clades; one clade was restricted to environmental isolates, three
32 other clades included distinct environmental isolates and dairy strains, while the fifth clade
33 comprised 34 strains (13 STs), among which all but two were isolated from milk, cheese or
34 dairy environment. These findings suggest an adaptation to the dairy ecosystems by a group
35 of specialized European *G. candidum* strains. In addition, we developed a PCR inter-LTR
36 scheme, a fast and reproducible RAPD-like method for *G. candidum*, to type the closely
37 related dairy strains, which could not be distinguished by MLST. Overall, our findings
38 distinguished two types of dairy strains, one forming a homogeneous group with little genetic
39 diversity, and the other more closely related to environmental isolates. Neither regional nor
40 cheese specificity was observed in the dairy *G. candidum* strains analyzed. This present study
41 sheds light on the genetic diversity of both dairy and environmental strains of *G. candidum*
42 and thus extends previous characterizations that have focused on the cheese isolates of this
43 species.

44

45 **Key words:** MLST, inter-LTR, *Geotrichum candidum*, dairy, population

46

47

48 **Introduction**

49 The dimorphic yeast *Geotrichum candidum* (teleomorph *Galactomyces candidus*) is
50 commonly found in foodstuffs, either as part of their normal constitution or as a contaminant.
51 It is present in raw milk (Desmaures *et al.*, 1997) and is marketed as a starter for cheese
52 making, because of its proteolytic activities and its organoleptic properties. For these reasons,
53 *G. candidum* is desirable on the surface of semi-hard cheeses and mould-ripened or smeared
54 soft cheeses (Boutrou and Gueguen, 2005), and is of considerable importance in the cheese-
55 making industry (Pottier *et al.*, 2008).

56 *G. candidum* displays substantial morphological diversity and a wide ecological distribution.
57 The genetic basis of this variability has been investigated: work based on RAM-PCR and
58 RAPD-PCR (Gente *et al.*, 2002; Gente *et al.*, 2006; Marcellino *et al.*, 2001) demonstrated a
59 high degree of genetic diversity of *G. candidum*, to some extent correlated to its various
60 ecological niches and morphological polymorphism. PFGE profiling of 13 strains revealed
61 highly diverse chromosome numbers, and hence genome sizes. It was suggested that genome
62 size correlated with strain morphology (Gente *et al.*, 2006). It was recently shown that *G.*
63 *candidum* displays an unusual intra-specific polymorphism in the ribosomal DNA sequences
64 used for species identification (Alper *et al.*, 2011).

65 Among various typing methods, Multi-Locus Sequence Typing (MLST) is based on
66 comparisons of partial housekeeping gene sequences. It is accepted to be reliable and
67 reproducible, and has the advantage of portability, via the creation of a common database
68 which can be continuously enlarged and updated by independent laboratories (Bougnoux *et*
69 *al.*, 2004; Taylor and Fisher, 2003). MLST is the most popular typing method for yeast and
70 has been applied to more than 10 species, including *Candida albicans* (Bougnoux *et al.*,
71 2002), *Candida glabrata* (Dodgson *et al.*, 2003), *Candida tropicalis* (Tavanti *et al.*, 2005),
72 *Candida dubliniensis* (McManus *et al.*, 2008), *Candida krusei* (Jacobsen *et al.*, 2007),

73 *Saccharomyces cerevisiae* (Ayoub *et al.*, 2006; Fay and Benavides, 2005; Munoz *et al.*, 2009)
74 and *Cryptococcus gattii* (Feng *et al.*, 2008). The first MLST scheme for *G. candidum* was
75 developed using six gene fragments covering 3009 nucleotides, defining 17 ST among 40
76 strains (Alper *et al.*, 2013). This work confirmed the intra-specific variability observed in *G.*
77 *candidum* (Gente *et al.*, 2002; Gente *et al.*, 2006; Marcellino *et al.*, 2001).

78 Long Terminal Repeats (LTRs) are short transposon-associated sequences of around 300-400
79 bp widely repeated in some genomes, like those of the genus *Saccharomyces* and other
80 Saccharomycotina yeasts (Bleykasten-Grosshans and Neuveglise, 2011; Neuveglise *et al.*,
81 2002), either as flanking parts of the cognate retrotransposon or as isolated elements resulting
82 from deletion of the retrotransposon through recombination between the two flanking LTRs.
83 The latter solo elements are numerous in some genomes, and this feature has been used to
84 differentiate isolates within a single species by amplifying the regions between LTRs, using
85 primers in the conserved regions of LTRs (inter-LTR PCR). This approach has been applied
86 to various species including *S. cerevisiae* (Legras and Karst, 2003; Ness *et al.*, 1993),
87 *Debaryomyces hansenii* and *Kluyveromyces marxianus* (Sohier *et al.*, 2009).

88 In this study, we exploit the completed genome sequence of the *G. candidum* strain CLIB 918
89 (=ATCC 204307) (Morel *et al.*, 2015) to develop protocols for typing this species and to gain
90 insights on the phylogenetic relationship between isolates.

91

92 **Materials and methods**

93

94 **Yeast strains and growth conditions.**

95 Isolates used in this study are listed in Table 1. “FM” strains were provided during the “Food
96 Microbiomes” project ([http://www.agence-nationale-recherche.fr/?Project=ANR-08-ALIA-](http://www.agence-nationale-recherche.fr/?Project=ANR-08-ALIA-0007)
97 0007) by producers of starter cultures and cheeses, and were given FM numbers to keep their
98 origin confidential; UCMA strains were provided by the University of Caen (France) and
99 CLIB strains by CIRM-Levures (Thiverval-Grignon, France;
100 <http://www6.inra.fr/cirm/Levures>). Most of the *G. candidum* isolates were from dairy
101 products and are involved in the ripening of various cheeses, some are from other habitats
102 such as soil; these strains were isolated in various places in France and the rest of the world.
103 Other *G. candidum* strains were from Centraalbureau voor Schimmelcultures (Utrecht, the
104 Netherlands), Museum National d'Histoire Naturelle, Laboratoire de Cryptogamie (Paris,
105 France), VTT Technical Research Center (Finland), Kasetsart University (Bangkok,
106 Thailand), DSMZ (Braunschweig, Germany), BCCM/MUCL (Louvain la Neuve, Belgium),
107 NCYC (Norwich, United Kingdom) and NBRC (Chiba, Japan). Routinely, yeast strains were
108 grown in YPD (Yeast Peptone Dextrose: yeast extract 10 g l⁻¹, bacto peptone 10 g l⁻¹, glucose
109 10 g l⁻¹) at 28°C with shaking. For solid medium, 14 g l⁻¹ of agar was added to YPD.

110

111 **DNA extraction.**

112 Cultures grown in 3 ml YPD medium overnight at 28°C were centrifuged at 2,500 g for 3
113 min, and the cell pellets were washed in 750 µl 50 mM EDTA. Each cell pellet was
114 resuspended in 200 µl lysis buffer (1 % SDS, 2 % triton, 100 mM NaCl, 50 mM EDTA,
115 50mM Tris, pH=8), 200 µl chloroform/phenol (pH=8) and 300 mg glass beads and
116 mechanically shaken by vortexing for 4 min. Then, 200 µl TE buffer was added and the

117 samples centrifuged for 5 min at 12,000 g. The aqueous phase was transferred to a new tube,
118 and two chloroform extractions were carried out. DNA was precipitated with an equal volume
119 of 100% ethanol and centrifuged for 4 min at 12,000g. The pellet was rinsed with 400 µl of 70
120 % ethanol, dried at room temperature for 15 min, resuspended in 50 µl TE buffer, and
121 incubated with 2 µl RNase (10 mg ml⁻¹) for 30 min at 37°C. The DNA concentration was
122 quantified on a 0.8% agarose gel (wt/vol) with 1X TAE electrophoresis buffer containing 0.2
123 mg ml⁻¹ ethidium bromide, run at 100 V in a SUB-CELL GT electrophoresis system (Bio-
124 Rad) for 60 min.

125

126 **PCR amplification.**

127 The primers used in this study, based on the complete genome sequence of *G. candidum*
128 CLIB 918 (Morel *et al.*, 2015) were designed using Primer3
129 (<http://fokker.wi.mit.edu/primer3>). For inter-LTR PCR fingerprinting, sequences with
130 similarities to a *G. candidum* retrotransposon LTR identified by Morel *et al.*, (2015) were
131 aligned using MAFFT7 (Kato and Standley, 2013). Primers were designed in the conserved
132 regions of the aligned sequences. The primers used were GC_LTR3 (5'-
133 TCTTTCCTTTAATTAATCAA) and GC_LTR5 (5'-TGAAGTGAACCAACATAA). For
134 MLST, the primers used were NUP116_for (5'-ACCGCTACAACCTGGATTTGG),
135 NUP116_rev (5'-GAGACCTGTTTGAGGGCTTG), PLB3_for (5'-
136 AAGAATATCTGGGATCTTTC), PLB3_rev (5'-TGAAGAAGAAGTACCAAGAA),
137 SAPT4_for (5'-ATCATTAACACCCCGGCATA), SAPT4_rev (5'-
138 GTGTCACCAAGCAGAGCAAA), URA1_for (5'-CAAGCCAATTGTGCTGAGAA),
139 URA1_rev (5'-GGTGTCTAGGGCAGTTGAT), URA3_for (5'-
140 GCCAAAAGACCAACCTGTG), URA3_rev (5'-CCTCATCCATACGGTTCTGC).

141 For MLST, between 25 and 50 ng of genomic DNA was amplified in a 50 μ l PCR reaction
142 mix containing 0.8 mM dNTP mixture, 0.1 μ M forward and reverse primers in the
143 manufacturer's recommended buffer and 1 U TaKaRa Ex *Taq*. Reactions were run on a 2720
144 thermal cycler (Applied Biosystems) as follows: 5 min at 94°C followed by 30 cycles of 30
145 sec at 94°C, 40 s at temperatures between 35 and 50°C and 30 sec at 72°C, with a final
146 extension step of 7 min at 72 °C.

147 For inter-LTR PCR fingerprinting, 2 μ l of DNA preparation (containing 20 ng) was added to
148 a 23 μ l PCR reaction mix containing 0.8 mM dNTP mixture, 0.8 μ M GC1_LTR and
149 GC2_LTR primers in the manufacturer's recommended buffer and 0.625 U TaKaRa Ex *Taq*.
150 PCR conditions were as follows: 94°C for 5 min, 30 cycles of 94°C for 30 s, 37°C for 40 s
151 and 72°C for 2 min, followed by a final extension at 72°C for 7 min. PCR products were
152 visualized on a 1.2% agarose gel (wt/vol) 15 cm in length with 1X TAE electrophoresis buffer
153 containing 0.2 mg ml⁻¹ ethidium bromide, run at 100 V in a SUB-CELL GT electrophoresis
154 system (Bio-Rad) for 90 min. The molecular marker was Gene ruler DNA Ladder Mix
155 (Thermo Scientific). Acquisition of data by the Bionumerics software 6.5 (Applied Maths,
156 Belgium) was as described in (Sohier *et al.*, 2009). Briefly, electrophoresis patterns were
157 normalized to compensate for migration bias, non-linear background was subtracted from the
158 patterns and comparison was based on the rolling disk principle. Inter-LTR variability not
159 being appropriate for phylogenetic analysis, and the object being only to discriminate strains
160 within an MLST sequence type, interpretation of the Bionumerics output was by visual
161 observation ; no clustering algorithm was performed.

162

163 **Synonymous and non-synonymous base ratio**

164 Codon by codon analysis was performed using MEGA7 (Kumar *et al.*, 2016). Pairwise
165 peptide comparisons were conducted with MEGA7 (codon based Z-test for selection) and

166 with the package *Dnasp* (Librado and Rozas, 2009). The likelihood analysis was performed
167 using the REL algorithm of the HyPhy package implemented in *DataMonkey* (Pond and
168 Frost, 2005).

169

170 **DNA sequence determination and phylogenetic analysis**

171 PCR fragments were sequenced on both strands by Eurofins MWG Operon (Ebersberg,
172 Germany). Sequences were assembled with the phred/phrap/consed package (Ewing and
173 Green, 1998; Ewing *et al.*, 1998; Gordon *et al.*, 1998) and analyzed with BLAST
174 implemented at NCBI (<http://www.ncbi.nlm.nih.gov>). Sequence alignments were generated
175 using MAFFT7 (Kato and Standley, 2013). Phylogenetic trees were constructed with the
176 Neighbor-Joining program implemented in MEGA6 (Tamura *et al.*, 2013). Phylogenetic trees
177 were visualized with NJplot (Perriere and Gouy, 1996). Bionumerics software 6.5 (Applied
178 Maths, Belgium) was used for the acquisition of inter-LTR profiles and the integration of
179 MLST, inter-LTR and phenotypic data.

180

181 **Phylogenetic analysis with *Structure***

182 The MLST data were used to predict the population structure on the basis of the clustering
183 program *Structure* (Pritchard *et al.*, 2000). Analyses were performed with burn-in of 80,000
184 cycles and 80,000 rounds of calculation. In accordance with the authors' recommendations,
185 the best estimation of the number of populations, K, was taken as the value of K for which the
186 difference in log(probability of data) between K and K+1 diminished abruptly (*ie.*, the log
187 probability ceased to increase rapidly). The population structure was thus calculated with K =
188 5 populations. For all other parameters, the recommended default values were used.

189

190 **Phylogenetic analysis using *Splits tree***

191 The MLST data were used to predict genetic exchanges with the program *Splits tree* (Huson
192 and Bryant, 2006). Recommended default values were used.

193

194 **Results**

195 **Design of a MLST scheme for *G. candidum*.**

196 A literature survey was undertaken to identify genes which have been used for MLST
197 schemes in previous studies on yeasts (Bougnoux *et al.*, 2002; Dodgson *et al.*, 2003; Jacobsen
198 *et al.*, 2007; McManus *et al.*, 2008; Munoz *et al.*, 2009; Tavanti *et al.*, 2005). The sequences
199 of 28 *G. candidum* orthologs of such genes were extracted from the genome of CLIB 918.
200 PCR primers were designed to amplify fragments of around 500 bp and were tested on 67
201 strains, which were considered to be haploid on the basis of their mating type characteristics
202 (Morel *et al.*, 2015). The study population included 34 *G. candidum* strains isolated from
203 cheese, 4 from milk, 1 from the dairy environment, 1 from cow's udder, 3 dairy product
204 contaminants and 17 isolates from diverse origins (fruit, human faeces, industrial plant,
205 industrial waste...). The latter group contained three *G. silvicola* strains, now reassigned to
206 the *G. candidum* species. A preliminary analysis was performed on closely related dairy
207 strains from France to select the most divergent markers. Five markers were selected as
208 showing the greatest sequence divergence between strains: *URA1*, *URA3*, *SAPT4* (*YSP3* in *S.*
209 *cerevisiae*), *NUP116* and *PLB3*. Data for the five MLST markers are presented in Table 2.
210 The PCR fragments analyzed were between 393 bp for *PLB3* and 501 bp for *SAPT4*, yielding
211 a concatenated sequence length of 2254 bp which contained 90 polymorphic sites.
212 Table 2 summarizes information concerning the diversity at each locus. The number of
213 polymorphic sites, only consisting in substitutions, varied between 11 for *URA1* and 28 for
214 *PLB3*. The number of allelic profiles between loci was five for *SAPT4* and nine for *PLB3*,
215 *URA1*, *URA3* and *NUP116*. The average divergence for the alleles ranges from 0.2 % for
216 *SAPT4* to 4.7 % for *PLB3* (Supplementary Table 1).
217 Analysis using MEGA7 showed evidence of positive selection in 4 of the 14 variable codons
218 of the *NUP116* sequence, in 2 of the 28 codons of *PLB3*, in 2 of 21 of *SAPT4*, in 1 of 11 of

219 *URA1* and in 1 of 16 of *URA3*, though without good statistical support ($P \approx 0.7$). Pairwise
220 peptide comparisons showed a positive overall dN - dS for the sequences of *NUP116* and of
221 *PLB3* in strain MUCL8652 when compared with about half of the other strains ($P=0.16$).
222 Analysis with the package DnaSP gave values of dN/dS always smaller than unity, excepting
223 cases where dS was zero. In the case of *NUP116* over 90% of the values of dN/dS were
224 smaller than 0.25; for the other four genes, over 90% were less than 0.15. A likelihood
225 approach showed 2 positively selected sites (Bayes factors > 100) in the sequence of *SAPT4*
226 and none in the other proteins. Considering the feeble evidence from the codon by codon
227 analysis and the Z-test, the absence of evidence from the analysis in Dnasp and the prediction
228 of 2/21 codons subject to selection in *SAPT4* by the likelihood analysis, we may conclude that
229 the large majority of sites used in the analysis are not subject to positive selection.

230

231 **Geographical and industrial influence on strain grouping.**

232 Forty sequence types (ST) were defined amongst the 67 strains, based on the possession of
233 identical concatenated sequences. Overall, 31 strains were the sole member of their ST, while
234 the other 36 strains were distributed among the remaining 9 STs: ST 21 and ST 23 each
235 contained two strains, ST 14 and ST 19 contained three strains, ST 7, ST 18, ST 29 and ST 30
236 contained five strains and ST 22 contained six strains (Table 3). The pairwise divergence
237 between STs (averaged over the concatenated sequences) was from 0.04 % to 2.57 % and
238 average divergence for all STs was 0.96 %.

239 Phylogenetic trees were constructed by neighbor joining following alignment of the
240 sequences using MAFFT7 (Figures 1 and 2). The tree constructed with ST type sequences,
241 shown in Figure 1, differentiated five main clades. Two well-separated clades 1 and 2
242 included mainly environmental isolates and dairy strains. Clade 1 contained three STs
243 represented by a single isolate and one (ST 7) which contained five strains. Clade 2 contained

244 five single-isolate STs. A third clade also heterogeneous with regard to the origin of the
245 strains contained 14 single-strain STs and one (ST 14) composed of three strains. Clade 4
246 contained three single-strain STs and consisted of two strains isolated from soil and one from
247 fruit. Finally, in contrast to the previous clades, clade 5 contained only strains involved in
248 cheese making (either from cheese made with raw milk or from starters used in the cheese
249 industry), with the exception of two isolates (FM 268 and FM 269, in ST 18 and ST 27
250 respectively) which were isolated from human stools. It may be reasonable to propose that
251 these latter isolates have a cheese origin, considering their position in the tree and their
252 geographical origin. Isolates assigned to clade 5 were less divergent; seven of the thirteen STs
253 contained from two to six strains.

254 The slovakian strain CBS 11176 was classified among cheese isolates of clade 5. This strain
255 was first described as *Geotrichum bryndzea* (Sulo *et al.*, 2009), but it was recently shown to
256 belong to *Gal. candidus*, the teleomorph of *G. candidum* (Groenewald *et al.*, 2012); the
257 position of this strain in the phylogenetic tree shown in Figure 2 confirmed the recent
258 taxonomic change.

259

260 **Population analysis**

261 Since *G. candidum* is a sexually reproducing species, we performed analyses using the
262 programs *Structure* and *Splits tree*. *Structure* clusters the isolates based on their respective
263 genetic profiles which are considered in terms of the relative contributions from a predefined
264 number of hypothetical ancestral genomes. *Splits tree* calculates a phylogenetic tree which
265 permits the visualization of genetic exchanges in addition to mutations.

266 Preliminary analysis of the data with *Structure* suggested that the data were best explained
267 under the hypothesis of five ancestral genotypes. The results of clustering on this basis
268 (Figure 3a) distinguish the previously defined clades 1 and 5, and suggest that the clade 2

269 contains hybrid organisms with genetic contributions originating in clades 1 and 3. Individual
270 gene trees for each of the markers also indicated genetic exchanges between isolates
271 (Supplementary Figure 1). As suggested by the above Neighbor-Joining analysis (Figures 1
272 and 2), the clade 3 is composed of two subgroups, while clade 4 branches from this group.
273 Clades 1 and 5 contained almost exclusively dairy strains, or strains isolated from human
274 faeces, presumably as a consequence of having eaten dairy products, while clades 2, 3 and 4
275 contained a mixture of dairy and environmental strains. *Splits tree* gave similar results (Figure
276 3b), separating the strains into groups which, with few exceptions, corresponded to the clades
277 previously defined. The analysis separated clades 1 and 5 and illustrated the derivation of
278 clades 2 and 4 from the clade 3. Again, the distinction of clades 1 and 5, of dairy origin, from
279 the other groups, of which environmental strains constituted about half the isolates, suggests
280 that dairy strains have developed separately from non-dairy strains.

281

282 **Inter-LTR PCR distinguishes most of the *G. candidum* dairy strains.**

283 Although MLST was very informative about the divergence of environmental strains and
284 dairy strains, it was not clear whether the large ST groups were due to clonality or limits to
285 the discrimination of the MLST markers. We therefore developed an additional typing
286 method more sensitive to genetic changes occurring at short time-scales.

287 Sequence analysis of the complete genome of *G. candidum* CLIB 918 revealed the presence
288 of a Long Terminal Repeat (LTR)-retrotransposon likely related to the yeast Ty families
289 (Morel *et al.*, 2015). The corresponding 385 bp long LTR sequences found in the genome of
290 CLIB 918 were aligned with MAFFT7 and oligonucleotide primers were designed on the
291 basis of conserved regions (Supplementary Figure 2). These were used to amplify genomic
292 DNA from the 36 strains which were in multi-strain STs (listed in Table 3). Various primer
293 pairs and various PCR conditions were tested; those leading to the most discriminating results

294 (GC_LTR3 and GC_LTR5) were retained. In the subsequent analysis all the inter-LTR PCRs
295 were repeated at least twice as previously described (Sohier *et al.*, 2009).

296 Results of electrophoresis are shown in Figure 3 and Supplementary figure 3. The inter-LTR
297 profiles were compared to the MLST tree obtained previously to highlight the variability of
298 strains within the STs, now observable with inter-LTR analysis. The amplified bands ranged
299 from 200 to 4000 bp and the profiles of the various strains differed in number, size and
300 relative intensity of fragments. In the majority of cases, strains within the multi-strain ST were
301 distinguished by inter-LTR analysis. Thus, 28 strains were uniquely discriminated by a
302 combination of MLST and inter-LTR, while two strains of ST 29 (CLIB 1239 and CLIB
303 1241) shared the same inter-LTR profiles, as did FM 76 and FM 77 in ST 18, and CLIB 1244,
304 CLIB 1245, CLIB 1246, CLIB1247 and CLIB 1253 in ST 30. The two strains of ST 29
305 originated from the Franche-Comté region and the five strains in ST 30 were isolated from the
306 Haute-Savoie; their classification into the same clusters is therefore coherent. In conclusion,
307 the combination of MLST and inter-LTR allowed precise discrimination between cheese
308 strains.

309 As described previously (Marcellino *et al.*, 2001), we observed a limited, but reproducible,
310 phenotypic diversity in terms of carbon source assimilation (galactose, lactose and ribose;
311 Figure 3). The capacity to utilize ribose is rarely found in *G. candidum*. We found that six of
312 the 37 isolates could utilize ribose. They were mainly associated with ST 18 (4 of 5 isolates
313 tested). Only two, unrelated strains were lactose-negative. Galactose assimilation was absent
314 from ST 30 and from two thirds of the isolates belonging to the sister clade ST 22, these being
315 all cheese strains from the region Haute-Savoie region.

316

317

318 **Discussion**

319 We report the development of tools for investigating the evolution of *G. candidum* and for
320 studying industrial strains used in cheese production. The MLST analysis we describe here
321 involving five single copy genes confirms the high genetic diversity reported for *G.*
322 *candidum*. Previous typing studies based on RAM-PCR, RAPD and Chromosome Length
323 Polymorphism (CLP) techniques suggest substantial diversity within the species (Marcellino
324 *et al.*, 2001; Gente *et al.*, 2002; Gente *et al.*, 2006), which was again evidenced by the
325 variability of the ribosomal DNA sequences observed by (Alper *et al.*, 2011). Our results
326 demonstrate a large degree of genetic variation in the five genes considered (between 2.4 and
327 7.1 % of polymorphic sites, this being higher than the 0.9 to 3.2 % found in the study of
328 (Alper *et al.*, 2013).

329 In this work, *G. candidum* strains could be readily differentiated on the basis of the partial
330 sequences of five genes: 40 different STs were obtained for 67 isolates based on 90
331 polymorphic sites within the 2254 bp. This compares favorably with the 17 STs distinguished
332 in the MSLT scheme of Alper *et al.* (2013) using 58 polymorphic sites in 6 loci (3009 bp), our
333 choice of candidate genes being facilitated by the availability of the whole genome sequence
334 of *G. candidum*. In addition, the analysis was performed on a larger number of isolates from a
335 wide range of geographical or ecological origins, including the environment, industrial plant
336 or food contaminant, which permitted a more general view of the phylogenetic diversity of
337 the species.

338 Twenty-eight of the strains studied here have previously been studied by RAPD (Marcellino
339 *et al.*, 2001). The previous RAPD results and our findings are generally coherent, although
340 there were some differences. The three strains, CLIB 1267 (GC129 in Marcellino *et al.*,
341 2001), CLIB 1258 (=GC101) and CLIB 1237 (=GC37), isolated from milk used respectively
342 for manufacturing Camembert (Normandy), St Nectaire (Auvergne) and Chaource cheeses

343 (Champagne-Ardenne), were previously found by (Marcellino *et al.*, 2001) to be closely
344 related to other cheese strains. However, we found that they are well separated from other
345 cheese strains and closely related to environmental strains (Figure 2). Another difference
346 concerns the yeasts that we assign to ST30 (composed of isolates from ripening cheese from
347 Annecy). The isolate CLIB 1253 (=GC90) was suggested to be different from the other strains
348 that we classified as ST 30 (Marcellino *et al.*, 2001). We clearly show here that the five
349 strains belonging to ST 30 are indistinguishable on the basis of MLST, and inter-LTR profile.
350 In addition they have identical phenotypic characteristics (this work), and we have previously
351 shown that they carry the same mating type (Morel *et al.*, 2015). These strains were isolated
352 from two different cheeses, Reblochon and Tomme de Savoie, made in different dairy
353 facilities. Our results therefore indicate that these are clones of the same strain present at more
354 than one particular site and in more than one cheese, suggesting that *G. candidum* isolates are
355 not region or cheese specific.

356 The 67 isolates studied here were classified in 40 STs. A total of 31 isolates were typed by
357 MLST alone as they all display a unique ST. Though all of the environmental isolates could
358 be distinguished, there were a number of dairy isolates (eight isolates in clades 1, 2 or 3 and
359 28 isolates belonging to the dairy clade 5) that were not typed by this approach. We were able
360 to improve strain differentiation by developing inter-LTR PCR for *G. candidum*, which
361 successfully differentiated all but nine of the strains that were not separated by MLST. Inter-
362 LTR PCR is simple, reliable and fast, as demonstrated previously for other species (Ness *et*
363 *al.*, 1993; Legras and Karst, 2003; Sohier *et al.*, 2009) and may therefore be useful for strain
364 differentiation, for example, in such applications as analysis of yeast population dynamics
365 during cheese ripening.

366 It has been reported that *G. candidum* is highly variable as assessed by chromosome
367 separation using PFGE (Gente *et al.*, 2006). However, CLP very often results from ectopic

368 recombination between similar sequences, such as transposons (Zolan, 1995; Casaregola et
369 al., 1998; Rachidi et al., 1999). Accordingly, we observed a very large number of various
370 kinds of transposons in the genome of *G. candidum* (Morel *et al.*, 2015). The discrepancy
371 between MLST and RAPD may be inherent to the techniques used because methods based on
372 the amplification of repeats do not measure genetic diversity directly, but rather, like PFGE,
373 indicate chromosomal differences. It has been demonstrated, for example, that the species
374 *Schizosaccharomyces pombe* displays substantial CLP which is not associated with either
375 sequence or phenotypic variability (Brown *et al.*, 2011). In this study, strains which could not
376 be separated on the basis of sequence divergence, *i.e.* MLST, displayed little phenotypic
377 diversity; they are therefore presumably closely related. The use of inter-LTR PCR permitted
378 a fine-scale discrimination of these related isolates within a single ST.

379 Our MLST analysis differentiated a clade containing dairy isolates from those containing
380 environmental isolates mixed with dairy isolates, in agreement with the suggestion that *G.*
381 *candidum* species could be separated into two groups, on the basis of rRNA variability and
382 genomic sequence diversity (Alper *et al.*, 2013). The observation of these authors concerns
383 dairy yeasts, and may correspond to the two types of dairy yeasts that we evidenced (clade 5
384 vs. clades 2, 3 and 4). In summary, our work has brought to light a population of dairy yeasts
385 which is clearly distinct from the clades containing a mixture of dairy and environmental
386 strains. This is reminiscent of population genomics findings for *S. cerevisiae* (Liti et al., 2009;
387 Schacherer et al., 2009), in which specialized isolates formed separated clades. *G. candidum*
388 clade 5 is reminiscent of the *S. cerevisiae* European wine group in which strains are closely
389 related and specialized in wine making.

390 In addition to shedding light on the population structure and probable evolution of the species
391 *G. candidum*, our work provides information and methods with potential to facilitate genetic
392 improvement and studies of genetic diversity within this species. For example, it provides a

393 framework which will make it easier to correlate genetic divergence with industrially
394 important characteristics of *G. candidum*, such as aroma production, morphology and growth
395 characteristics.

396

397 **Acknowledgements**

398

399 This work received funding from the *Agence National pour la Recherche* grant “Food
400 Microbiomes” (ANR-08-ALIA-007-02). This publication made use of the *Geotrichum* MLST
401 website (<http://pubmlst.org/geotrichum/>) developed by Keith Jolley and sited at the University
402 of Oxford (Jolley and Maiden, 2010). The development of this site has been funded by the
403 Wellcome Trust. We thank Christelle Louis-Mondésir for expert technical assistance. We are
404 grateful to Prof. Savitree Limtong (Kasetsart University, Thailand) for providing us with the
405 isolate NT12. We are grateful to Guillaume Morel (supported by a CIFRE fellowship with
406 CNIEL) for the extraction from the genome sequence of CLIB 918 of the genes of interest.
407 We also thank the *Centre National Interprofessionnel de l'Economie Laitière* (CNIEL) and
408 the *Syndicat Professionnel des Producteurs d'Auxiliaires pour l'Industrie Laitière* (SPPAIL)
409 for providing us with dairy strains.

410

411 **References**

- 412 Alper I, Frenette M, Labrie S. 2011. Ribosomal DNA polymorphisms in the yeast *Geotrichum*
413 *candidum*. *Fungal Biol* **115**: 1259-69.
- 414 Alper I, Frenette M, Labrie S. 2013. Genetic diversity of dairy *Geotrichum candidum* strains
415 revealed by multilocus sequence typing. *Appl Microbiol Biotechnol* **97**: 5907-20.
- 416 Ayoub MJ, Legras JL, Saliba R, *et al.* 2006. Application of Multi Locus Sequence Typing to
417 the analysis of the biodiversity of indigenous *Saccharomyces cerevisiae* wine yeasts
418 from Lebanon. *J Appl Microbiol* **100**: 699-711.
- 419 Bleykasten-Grosshans C, Neuveglise C. 2011. Transposable elements in yeasts. *C R Biol* **334**:
420 679-86.
- 421 Bougnoux ME, Aanensen DM, Morand S, *et al.* 2004. Multilocus sequence typing of *Candida*
422 *albicans*: strategies, data exchange and applications. *Infect Genet Evol* **4**: 243-52.
- 423 Bougnoux ME, Morand S, d'Enfert C. 2002. Usefulness of multilocus sequence typing for
424 characterization of clinical isolates of *Candida albicans*. *J Clin Microbiol* **40**: 1290-7.
- 425 Boutrou R, Gueguen M. 2005. Interests in *Geotrichum candidum* for cheese technology. *Int J*
426 *Food Microbiol* **102**: 1-20.
- 427 Brown WR, Liti G, Rosa C, *et al.* 2011. A Geographically Diverse Collection of
428 *Schizosaccharomyces pombe* Isolates Shows Limited Phenotypic Variation but
429 Extensive Karyotypic Diversity. *G3 (Bethesda)* **1**: 615-26.
- 430 Desmasures N, Bazin F, Gueguen M. 1997. Microbiological composition of raw milk from
431 selected farms in the Camembert region of Normandy. *J Appl Microbiol* **83**: 53-8.
- 432 Dodgson AR, Pujol C, Denning DW, *et al.* 2003. Multilocus sequence typing of *Candida*
433 *glabrata* reveals geographically enriched clades. *J Clin Microbiol* **41**: 5709-17.
- 434 Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error
435 probabilities. *Genome Res* **8**: 186-94.
- 436 Ewing B, Hillier L, Wendl MC, *et al.* 1998. Base-calling of automated sequencer traces using
437 phred. I. Accuracy assessment. *Genome Res* **8**: 175-85.
- 438 Fay JC, Benavides JA. 2005. Evidence for domesticated and wild populations of
439 *Saccharomyces cerevisiae*. *PLoS Genet* **1**: 66-71.
- 440 Feng X, Yao Z, Ren D, *et al.* 2008. Genotype and mating type analysis of *Cryptococcus*
441 *neoformans* and *Cryptococcus gattii* isolates from China that mainly originated from
442 non-HIV-infected patients. *FEMS Yeast Res* **8**: 930-8.
- 443 Gente S, Desmasures N, Panoff JM, *et al.* 2002. Genetic diversity among *Geotrichum*
444 *candidum* strains from various substrates studied using RAM and RAPD-PCR. *J Appl*
445 *Microbiol* **92**: 491-501.
- 446 Gente S, Sohier D, Coton E, *et al.* 2006. Identification of *Geotrichum candidum* at the species
447 and strain level: proposal for a standardized protocol. *J Ind Microbiol Biotechnol* **33**:
448 1019-31.
- 449 Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing.
450 *Genome Res* **8**: 195-202.
- 451 Groenewald M, Coutinho T, Smith MT, *et al.* 2012. Species reassignment of *Geotrichum*
452 *bryndzae*, *Geotrichum phurueaensis*, *Geotrichum silvicola* and *Geotrichum vulgare*
453 based on phylogenetic analyses and mating compatibility. *Int J Syst Evol Microbiol*
454 **62**: 3072-80.
- 455 Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies.
456 *Mol Biol Evol* **23**: 254-67.
- 457 Jacobsen MD, Gow NA, Maiden MC, *et al.* 2007. Strain typing and determination of
458 population structure of *Candida krusei* by multilocus sequence typing. *J Clin*
459 *Microbiol* **45**: 317-23.

460 Jolley KA, Maiden MC. 2010. BIGSdb: Scalable analysis of bacterial genome variation at the
461 population level. *BMC Bioinformatics* **11**: 595.

462 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
463 improvements in performance and usability. *Mol Biol Evol* **30**: 772-80.

464 Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis
465 Version 7.0 for Bigger Datasets. *Mol Biol Evol* **33**: 1870-4.

466 Larkin MA, Blackshields G, Brown NP, *et al.* 2007. Clustal W and Clustal X version 2.0.
467 *Bioinformatics* **23**: 2947-2948.

468 Legras JL, Karst F. 2003. Optimisation of interdelta analysis for *Saccharomyces cerevisiae*
469 strain characterisation. *FEMS Microbiol Lett* **221**: 249-55.

470 Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA
471 polymorphism data. *Bioinformatics* **25**: 1451-2.

472 Marcellino N, Beuvier E, Grappin R, *et al.* 2001. Diversity of *Geotrichum candidum* strains
473 isolated from traditional cheesemaking fabrications in France. *Appl Environ Microbiol*
474 **67**: 4752-9.

475 McManus BA, Coleman DC, Moran G, *et al.* 2008. Multilocus sequence typing reveals that
476 the population structure of *Candida dubliniensis* is significantly less divergent than
477 that of *Candida albicans*. *J Clin Microbiol* **46**: 652-64.

478 Morel G, Sterck L, Swennen D, *et al.* 2015. Differential gene retention as an evolutionary
479 mechanism to generate biodiversity and adaptation in yeasts. *Sci Rep* **5**: 11571.

480 Munoz R, Gomez A, Robles V, *et al.* 2009. Multilocus sequence typing of oenological
481 *Saccharomyces cerevisiae* strains. *Food Microbiol* **26**: 841-6.

482 Ness F, Lavallée F, Dubourdiou D, *et al.* 1993. Identification of yeast strains using the
483 polymerase chain reaction. *Journal Sci Food Agri.* **62**: 69-94.

484 Neuveglise C, Feldmann H, Bon E, *et al.* 2002. Genomic evolution of the long terminal repeat
485 retrotransposons in hemiascomycetous yeasts. *Genome Res* **12**: 930-43.

486 Perriere G, Gouy M. 1996. WWW-query: an on-line retrieval system for biological sequence
487 banks. *Biochimie* **78**: 364-9.

488 Pond SL, Frost SD. 2005. Datamonkey: rapid detection of selective pressure on individual
489 sites of codon alignments. *Bioinformatics* **21**: 2531-3.

490 Pottier I, Gente S, Vernoux JP, *et al.* 2008. Safety assessment of dairy microorganisms:
491 *Geotrichum candidum*. *Int J Food Microbiol* **126**: 327-32.

492 Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using
493 multilocus genotype data. *Genetics* **155**: 945-59.

494 Sohier D, Dizes AS, Thuault D, *et al.* 2009. Important genetic diversity revealed by inter-LTR
495 PCR fingerprinting of *Kluyveromyces marxianus* and *Debaryomyces hansenii* strains
496 from French traditional cheeses. *Dairy Science & Technology* **89**: 569-581.

497 Tamura K, Stecher G, Peterson D, *et al.* 2013. MEGA6: Molecular Evolutionary Genetics
498 Analysis version 6.0. *Mol Biol Evol* **30**: 2725-9.

499 Tavanti A, Davidson AD, Johnson EM, *et al.* 2005. Multilocus sequence typing for
500 differentiation of strains of *Candida tropicalis*. *J Clin Microbiol* **43**: 5593-600.

501 Taylor JW, Fisher MC. 2003. Fungal multilocus sequence typing--it's not just for bacteria.
502 *Curr Opin Microbiol* **6**: 351-6.

503

504

Table 1 List of strains studied

Species	Isolate	Other names	Origin	Source
<i>Geotrichum bryndzae</i> ^a	CLIB 1366 ^T	CBS 11176 ^T	Slovakia	Bryndza cheese
<i>Geotrichum candidum</i>	CLIB 1368 ^T	CBS 615.84 ^T	Ile de France	Brie cheese
//	CLIB 918	ATCC 204307	Normandy	Pont l'Evêque cheese
//	CLIB 1235	GC12	Normandy	Camembert cheese
//	CLIB 1236	GC 21	Normandy	Goat's cheese
//	CLIB 1237	GC37	Normandy	Camembert cheese
//	CLIB 1239	GC43	Franche-Comté	Mont d'Or cheese
//	CLIB 1240	GC44	Haute-Savoie	Reblochon cheese
//	CLIB 1241	GC45	Franche-Comté	Mont d'Or cheese
//	CLIB 1242	GC42	Haute-Savoie	Reblochon cheese
//	CLIB 1243	GC51	Haute-Savoie	Reblochon cheese
//	CLIB 1244	GC59	Haute-Savoie	Tomme de Savoie cheese
//	CLIB 1245	GC60	Haute-Savoie	Reblochon cheese
//	CLIB 1246	GC63	Haute-Savoie	Reblochon cheese
//	CLIB 1247	GC64	Haute-Savoie	Tomme de Savoie cheese
//	CLIB 1248	GC74	Haute-Savoie	Reblochon cheese
//	CLIB 1251	GC79	Burgundy	Epoisses cheese
//	CLIB 1252	GC84	Burgundy	Epoisses cheese
//	CLIB 1253	GC90	Haute-Savoie	Reblochon cheese
//	CLIB 1255	GC96	Haute-Savoie	Reblochon cheese
//	CLIB 1256	GC97	Haute-Savoie	Reblochon cheese
//	CLIB 1257	GC100	Auvergne	Saint Nectaire cheese
//	CLIB 1258	GC101	Auvergne	Saint Nectaire cheese
//	CLIB 1260	GC105	Auvergne	Saint Nectaire cheese
//	CLIB 1262	GC110	Auvergne	Saint Nectaire cheese
//	CLIB 1263	GC120	Auvergne	Saint Nectaire cheese
//	CLIB 1267	GC129	Champagne-Ardenne	Chaource cheese
//	CLIB 1270	GC146	Auvergne	Saint Nectaire cheese
//	CLIB 1274	GC164	Haute-Savoie	Reblochon cheese
//	CLIB 1283	UCMA 103	Normandy	Pont l'Evêque cheese
//	CLIB 1285	UCMA 3936	Normandy	Dairy
//	CLIB 1361	CBS 182.33	Italy	Yoghurt
//	CLIB 1415	LCP 51.590	Spain	Sand
//	FM 03		unknown	Cheese contaminant
//	FM 115		unknown	unknown
//	FM 122		unknown	unknown
//	FM 125		unknown	unknown
//	FM 127		unknown	unknown
//	FM 128		unknown	unknown
//	FM 136		unknown	unknown
//	FM 212		France	Corn silage

//	FM 214		Normandy	Cow's milk
//	FM 260		Normandy	Cow's udder
//	FM 268		Normandy	Stools
//	FM 269		Normandy	Stools
//	FM 270		Normandy	Stools
//	FM 29		Auvergne	Goat's cheese
//	FM 30		Auvergne	Goat's cheese
//	FM 31		Auvergne	Goat's cheese
//	FM 34		Auvergne	Goat's cheese
//	FM 76		Normandy	Cow's milk
//	FM 77		unknown	unknown
//	CBS 117138		Italy	Municipal compost
//	CBS 11628		South Africa	Soil
//	CBS 476.83		Senegal	Soil
//	CBS 557.83		Egypt	Fruit
//	DSM 10452		Germany	Sauerkraut
//	DSM 13629		United Kingdom	Polyurethane
//	MUCL 11539		United Kingdom	Polluted water
//	MUCL 14462		United States	Squash, Curcubita
//	MUCL 881		Belgium	Milk
//	MUCL 8652		Belgium	Soaked hay
//	NBRC 5368		unknown	Butter
//	NCYC 49		unknown	Milk
<i>Geotrichum silvicola</i> ^a	CLIB 1378 ^T	CBS 9194 ^T	Brazil	Insect
//		NT 12	Thailand	Rain forest
//		VTTC 4559	Sweden	Malting system

^a previous species denomination. GC, strains described in Marcelino et al. (2001).

Table 2 Characteristics of the five loci studied

Locus	Size of examined amplicon (bp)	Sequence divergence (%)	Number of polymorphic sites	Number of allelic profile per locus
<i>NUP116</i>	425	3.29	14	9
<i>URA1</i>	465	2.36	11	9
<i>URA3</i>	470	3.40	16	9
<i>SAPT4</i>	501	4.19	21	5
<i>PLB3</i>	393	7.12	28	9

1 **Table 3** MLST profiles for each strains of this study and distribution of isolates in STs

2

Strains	Markers					STs
	<i>NUPI16</i>	<i>PLB3</i>	<i>SAPT4</i>	<i>URA1</i>	<i>URA3</i>	
NT 12	6	2	2	5	4	1
CLIB 1267	2	2	4	1	1	2
CLIB 1237	2	2	2	2	1	3
LCP 51.590	2	2	5	2	3	4
CLIB 1235	1	2	1	2	1	5
CBS 182.33	2	2	2	1	1	6
CLIB 1236	2	2	1	2	1	7
CLIB 1251	2	2	1	2	1	7
CLIB 1252	2	2	1	2	1	7
CLIB 1263	2	2	1	2	1	7
FM 125	2	2	1	2	1	7
CBS 9194	4	3	3	3	1	8
FM 122	3	7	2	2	1	9
CLIB 1283	2	1	2	1	1	10
FM 270	2	1	2	4	3	11
FM 03	5	4	2	2	1	12
FM 212	2	4	2	2	1	13
FM 136	1	4	2	2	1	14
CLIB 1258	1	4	2	2	1	14
CBS 117138	1	4	2	2	1	14
VTTC 4559	1	4	2	2	5	15
CLIB 1274	2	5	1	2	1	16
FM 260	3	4	1	2	3	17
FM 214	2	1	1	2	3	18
FM 268	2	1	1	2	3	18
FM 34	2	1	1	2	3	18
FM 76	2	1	1	2	3	18
FM 77	2	1	1	2	3	18
CBS 11176	2	1	1	2	1	19
CLIB 1248	2	1	1	2	1	19
CLIB 1260	2	1	1	2	1	19
CBS 615.84	3	1	3	2	1	20
CLIB 1257	1	1	1	2	1	21
FM 128	1	1	1	2	1	21
CLIB 1240	1	1	1	1	1	22
CLIB 1242	1	1	1	1	1	22
CLIB 1243	1	1	1	1	1	22
CLIB 1255	1	1	1	1	1	22
FM 127	1	1	1	1	1	22
MUCL 881	1	1	1	1	1	22
FM 29	1	1	1	1	3	23
FM 31	1	1	1	1	3	23

FM 30	1	6	1	1	3	24
CLIB 1285	2	1	1	4	1	25
FM 115	3	6	5	1	1	26
FM 269	2	1	1	1	3	27
CLIB 918	2	1	1	1	2	28
CLIB 1239	2	1	1	1	1	29
CLIB 1241	2	1	1	1	1	29
CLIB 1256	2	1	1	1	1	29
CLIB 1262	2	1	1	1	1	29
CLIB 1270	2	1	1	1	1	29
CLIB 1244	5	1	1	1	1	30
CLIB 1245	5	1	1	1	1	30
CLIB 1246	5	1	1	1	1	30
CLIB 1247	5	1	1	1	1	30
CLIB 1253	5	1	1	1	1	30
DSM 10452	2	2	1	2	6	31
MUCL 11539	2	4	2	2	6	32
CBS 11628	7	7	2	6	7	33
CBS 476.83	8	7	2	6	8	34
CBS 557.83	7	8	2	9	8	35
NBRC 5368	9	4	2	7	1	36
MUCL 8652	3	9	2	1	1	37
DSM 13629	2	6	2	7	1	38
MUCL 14462	2	3	5	8	9	39
NCYC 49	2	2	1	2	2	40

1

2

1 **Figure legends**

2 **Figure 1:** Neighbor-Joining tree of the 40 STs generated from the analysis of the
3 concatenation of the five selected sequences in 67 *G. candidum* isolates. Bootstrap values
4 (1000 repetitions) are indicated at the main nodes. The evolutionary distances were computed
5 using the Kimura 2-parameter method and are expressed as number of base substitutions per
6 site. The rate variation among sites was modeled with a gamma distribution (shape parameter
7 = 5). The analysis involved 40 nucleotide sequences with a total of 2254 positions in the final
8 dataset. Evolutionary analyses were conducted using MEGA6 (Tamura *et al.*, 2011). Numbers
9 in brackets indicate the number of isolates with the same ST. Bar, 0.001 substitutions per site.

10 **Figure 2:** Neighbor-Joining tree of 67 *G. candidum* isolates based on the concatenated
11 sequences. Details of the calculations are as for figure 1.. The analysis involved 67 nucleotide
12 sequences with a total of 2254 positions in the final dataset. Evolutionary analyses were
13 conducted in MEGA6 (Tamura *et al.*, 2011). Strain numbers in bold are from dairy origin.
14 Strains with boxed numbers were isolated from stools. Strains with numbers in italic are of
15 unknown origin. Bar, 0.001 substitutions per site.

16 **Figure 3:** Population structure in connection with the phylogenetic analysis of 67 strains of
17 the 67 *G. candidum* strains. a) Clustering of the strain genotypes by *Structure*. For each of the
18 strains, presented on the abscissa, the estimated contribution (ordinate) to its genotype from
19 each of five hypothetical ancestral genotypes is represented as rectangles of different colours.
20 The origins of each strain are shown above the histogram: “D” milk or dairy, “E”
21 environment, “F” non-dairy food processing, “H” human stools, “n” not known. b)
22 Phylogenetic analysis of the yeasts using *Splits tree*. Strain names are colour-coded as for the
23 histogram labels in a) (D, E, F, H and n): blue, milk or dairy; orange, environmental strains;
24 green, non-dairy food processing; pink, human stools; grey, origin unknown. Groups 1 to 5

1 are highlighted for comparison with the results of the Neighbor-Joining phylogenetic tree of
2 Figure 2, and of the analysis with Structure.

3

4 **Figure 4:** Composite-analysis of the 36 isolates which were in non-unique STs. The MLST
5 dendrogram was generated as described in the Materials and Methods and is shown on the
6 left. The inter-LTR profile of each strain was super-imposed on the MLST tree. Variability of
7 the inter-MLST profile must therefore be observed within each represented ST. The ability to
8 assimilate galactose, lactose and ribose is indicated by a black rectangle. The geographical
9 origin and the substrate of isolation of the strains and the ST to which the strains belong are
10 shown at the right.

11

Figure 1

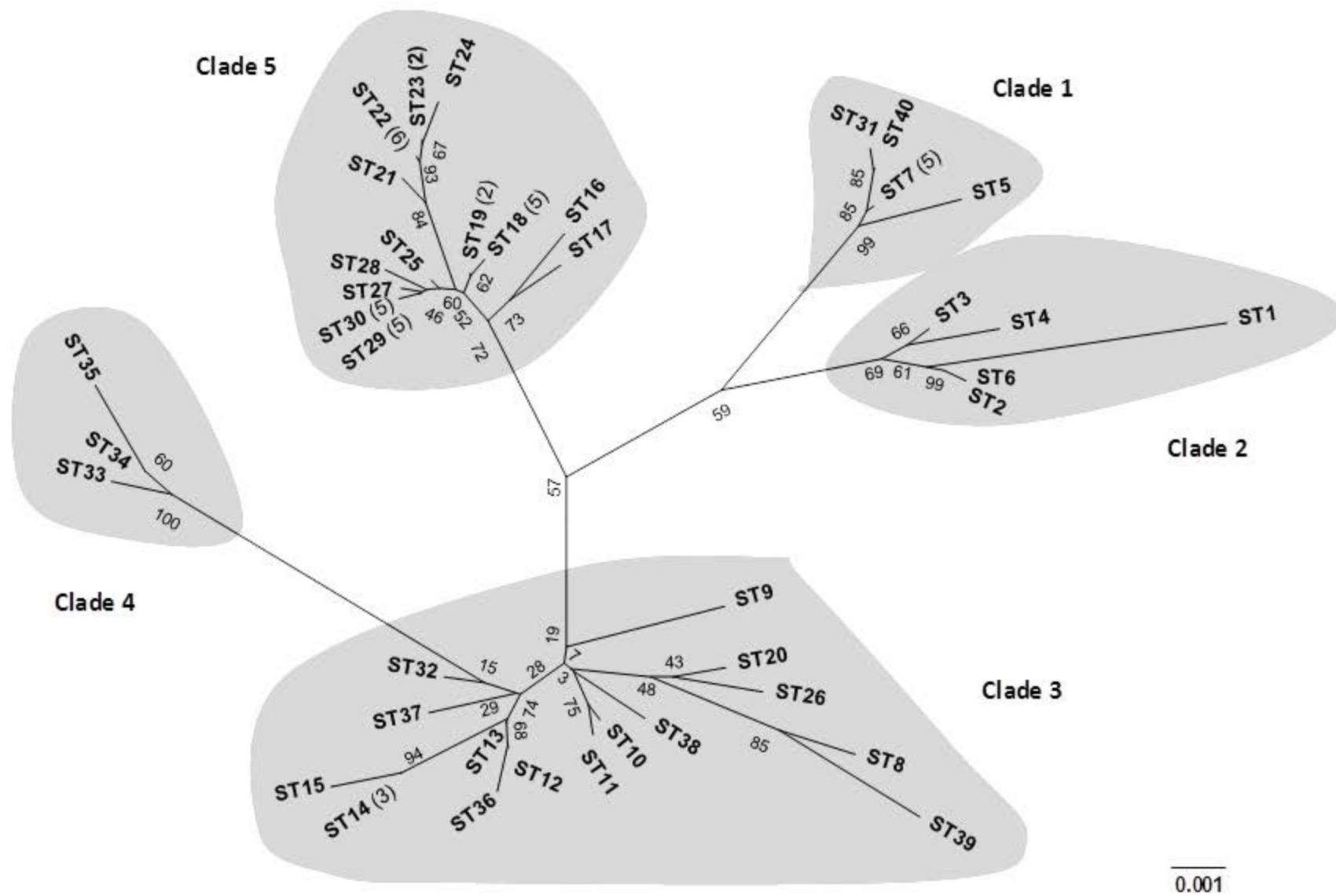


Figure 2

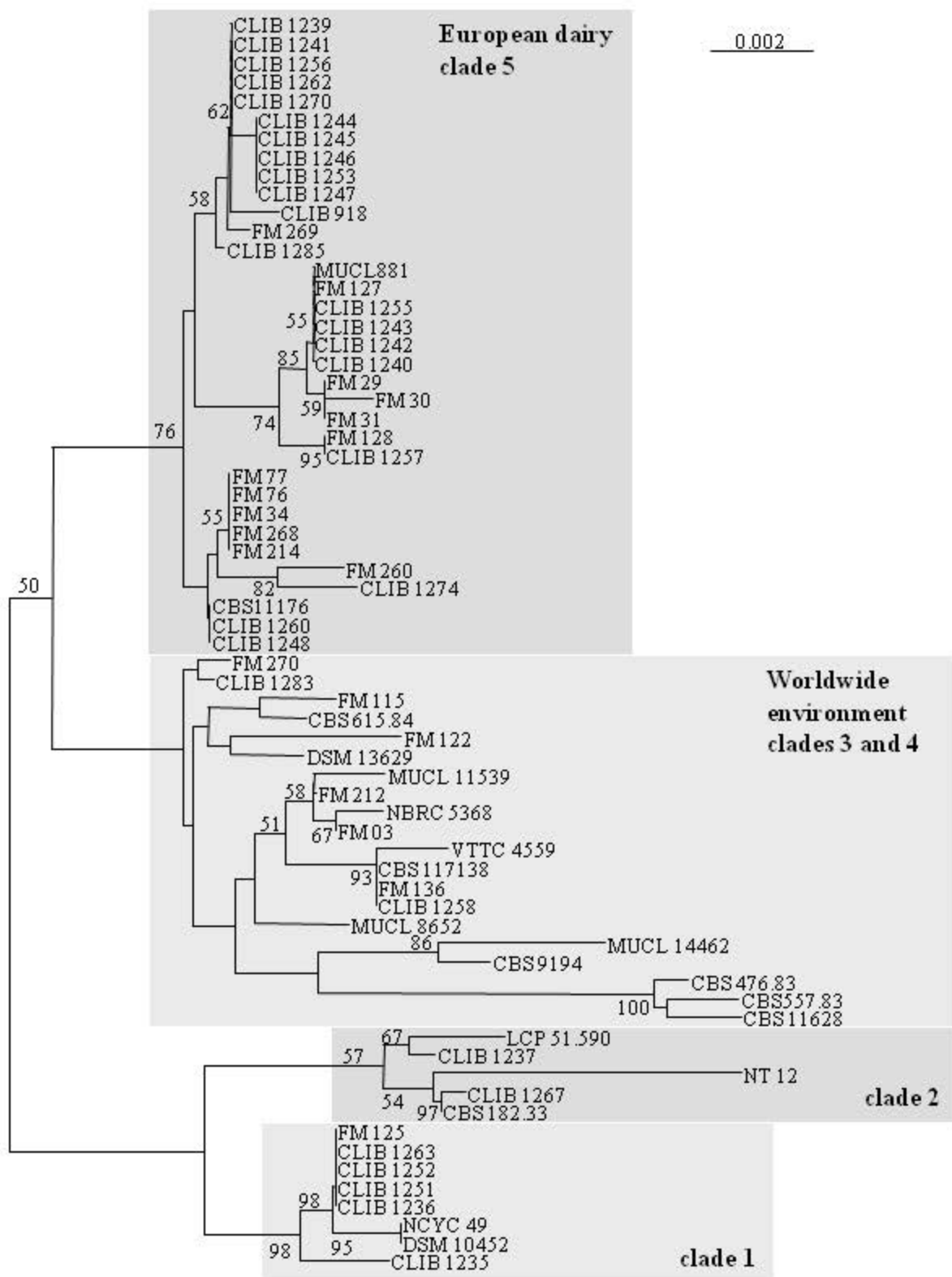
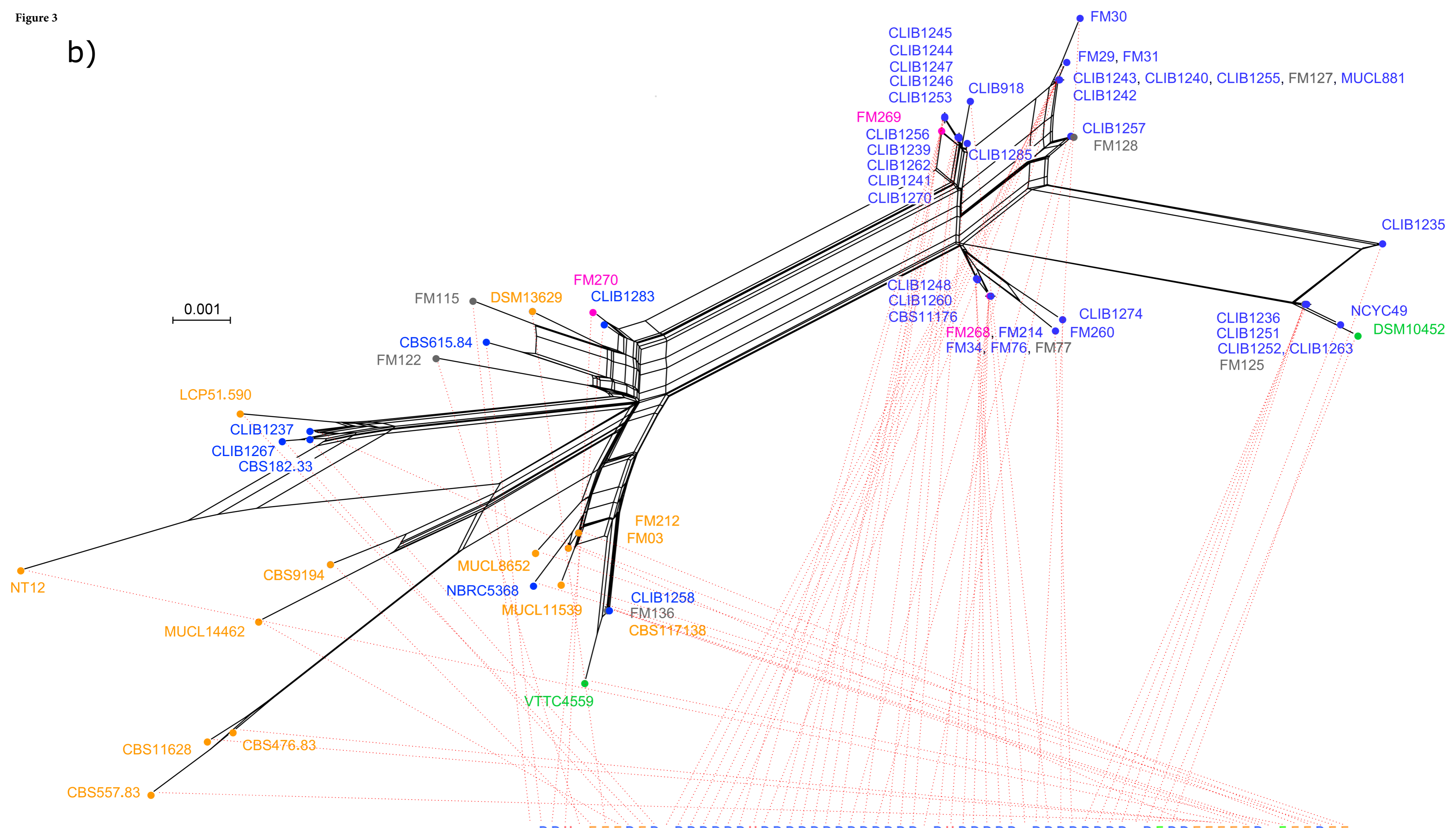


Figure 3

b)



a)

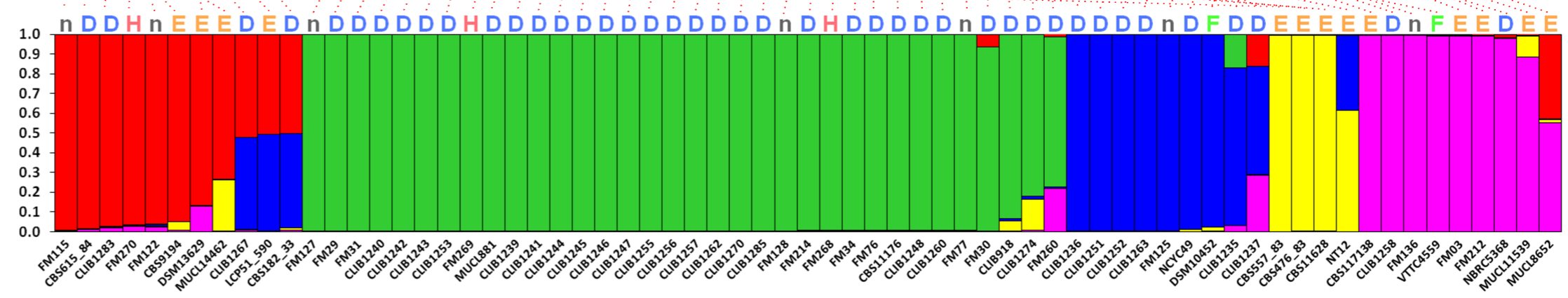


Figure 4

