



HAL
open science

Evolutionary genomics of sex-related chromosomes at the base of the green lineage 2

L Felipe Benites, François Bucchini, Sophie Sanchez-Brosseau, Nigel Grimsley, Klaas Vandepoele, Gwenael Piganeau

► **To cite this version:**

L Felipe Benites, François Bucchini, Sophie Sanchez-Brosseau, Nigel Grimsley, Klaas Vandepoele, et al.. Evolutionary genomics of sex-related chromosomes at the base of the green lineage 2. *Genome Biology and Evolution*, 2021, 10.1093/gbe/evab216/6380139 . hal-03369980

HAL Id: hal-03369980

<https://hal.sorbonne-universite.fr/hal-03369980>

Submitted on 7 Oct 2021

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.

1
2
3 **1 Evolutionary genomics of sex-related chromosomes at the base of the green lineage**
4
5 2

6 3 L. Felipe Benites^{1,a}, François Bucchini^{2,3,a}, Sophie Sanchez-Brosseau¹, Nigel Grimsley¹, Klaas
7 4 Vandepoele^{2,3,4}, Gwenaël Piganeau^{1*}
8
9 5

10 6
11 7 **Author affiliations:**
12 8

13 9 ¹Integrative Biology of Marine Organisms (BIOM), Sorbonne University, CNRS,
14 10 Oceanological Observatory of Banyuls, Banyuls-sur-Mer, France

15 11 ²Ghent University, Department of Plant Biotechnology and Bioinformatics, Technologiepark
16 12 71, 9052 Ghent, Belgium

17 13 ³VIB Center for Plant Systems Biology, Technologiepark 71, 9052 Ghent, Belgium

18 14 ⁴Bioinformatics Institute Ghent, Ghent University, Technologiepark 71, 9052 Ghent, Belgium

19 15 ^a equal contribution

20 16 * **Corresponding author:** gwenael.piganeau@obs-banyuls.fr
21 17

22 18 **Keywords : recombination suppression, mating-type loci, chlorophyta, GC content**
23 19 **phylogenetic profiling**
24 20
25 21
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57

Abstract

While sex is now accepted as a ubiquitous and ancestral feature of eukaryotes, direct observation of sex is still lacking in most unicellular eukaryotic lineages. Evidence of sex is frequently indirect and inferred from the identification of genes involved in meiosis from whole genome data and/or the detection of recombination signatures from genetic diversity in natural populations. In haploid unicellular eukaryotes, sex-related chromosomes are named mating-type (*MTs*) chromosomes and generally carry large genomic regions where recombination is suppressed. These regions have been characterized in Fungi and Chlorophyta and determine gamete compatibility and fusion. Two candidate *MT+* and *MT-* alleles, spanning 450-650 kb, have recently been described in *Ostreococcus tauri*, a marine phytoplanktonic alga from the Mamiellophyceae class, an early diverging branch in the green lineage.

Here, we investigate the architecture and evolution of these candidate *MT+* and *MT-* alleles. We analysed the phylogenetic profile and GC content of *MT* gene families in eight different species, whose divergence has been previously estimated at up to 640 million years, and found evidence that the divergence of the two *MTs* alleles predates speciation in the *Ostreococcus* genus. Phylogenetic profiles of *MT* trans-specific polymorphisms in gametologs disclosed candidate *MTs* in two additional species, and possibly a third. These Mamiellales *MT* candidates are likely to be the oldest mating-type loci described to date, which makes them fascinating models to investigate the evolutionary mechanisms of haploid sex determination in eukaryotes.

keywords: sex determining chromosome, recombination suppression, mating types, Chlorophyta, Mamiellophyceae

Significance statement:

Direct evidence of sexual reproduction is difficult to observe in many unicellular eukaryotes, while indirect evidence relies on gene content or recombination signatures. Here we report the gene content of two candidate mating type loci in a unicellular phytoplanktonic eukaryote. Identification and phylogenetic analyses of the gametologs shared between the two mating types suggest signatures of trans-specific evolution, i.e. an ancient divergence, prior to the speciation events within the *Ostreococcus* lineage. The divergence between gametologs can be leveraged to assign strains from distantly related species to each of the two mating types. Thus, they are likely to be the oldest mating-type loci described to date, which makes them

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

55 fascinating models to investigate the evolutionary mechanisms of haploid sex determination
56 in eukaryotes.

Downloaded from <https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evab216/6380139> by BIUS Jussieu user on 07 October 2021

57 Introduction

58 Meiotic sex and its associated intra-chromosomal and inter-chromosomal
59 recombination events are considered ubiquitous, ancestral features of eukaryotes (Speijer et al.
60 2015). Across the eukaryotic tree of life, meiotic sex has been reported in many algal lineages
61 (reviewed in Umen and Coelho 2019), such as chlorophytes (Sager and Granick 1954; Suda et
62 al. 1989; Fučíková et al. 2015), bacillariophytes (Chepurnov et al. 2004), chlorarachniophytes
63 (Beutlich and Schnetter 1993), cryptophytes (Hill and Wetherbee 1986; Kugrens and Lee 1988),
64 cyanidiophytes (Malik et al. 2007), dinoflagellates (Pfiester 1989) and euglenoids (Ebenezer et
65 al. 2019).

66 There have been intense efforts to study sex determining mechanisms and underlying
67 genetic make-up in multicellular animals and plants (Bachtrog et al. 2014 for a review).
68 However, less is known about sex-determining mechanisms in microbial eukaryotes. Ancestral
69 sex-determining mechanisms have evolved in unicellular eukaryotes, so that “*it is clear that the*
70 *evolution of different sexes in its most basic form is represented by the evolution of mating-*
71 *types*” (Hoekstra 1987). Obviously, it is less straightforward to identify morphological
72 differences between sexes in microorganisms than in macro-organisms. The term “mating type”
73 describes different “sexual types” in unicellular eukaryotes, and was first coined by Tracy
74 Sonneborn. He used this term to indicate that only certain lines (or “stocks”) of the ciliate
75 *Paramecium aurelia* mated with each other, but never with themselves (Sonneborn 1937). He
76 noted that the *Paramecium* mating system was “*strikingly similar to the sexual differences*
77 *between gametes in some of the unicellular green alga*”. He referred to earlier work by Strehlow
78 (1929) on *plus* and *minus* “sexes” reported in unicellular soil and freshwater green algae from
79 the order Chlamydomonadales. In the Fungal kingdom, there has been a rapidly growing
80 experimental evidence of mating types for many species (reviewed in Billiard et al. 2012; Wolfe
81 and Butler 2017), initially in the yeasts *Saccharomyces cerevisiae* (Astell et al. 1981) and
82 *Neurospora crassa* (Staben and Yanofsky 1990). Mating types were identified later in the green
83 algal lineage, as in *Chlamydomonas reinhardtii* (Ferris et al. 2002), and across the eukaryotic
84 tree of life (reviewed in Umen and Coelho 2019). Interestingly, the evolutionary link between
85 mating types and male and female sexes has been unambiguously demonstrated in the volvocine
86 green lineage (Nozaki et al. 2006)(Ferris et al. 2010)(Hamaji et al. 2018). However, the origin
87 of mating-types remains unresolved. Three main hypotheses have been formulated for the
88 origin and maintenance of this genetic setup, which requires outcrossing. First, it may mediate
89 the prevention of genetic conflicts (Hurst and Hamilton 1992); second, the prevention of

haploid selfing, that is mating among clonal cells e.g. (Billiard et al. 2011)(Billiard et al. 2012). A third proximate hypothesis is that this genetic system has evolved from a cell signalling system for partner recognition and pairing by producing recognition/attraction molecules and their receptors, as initially suggested by Hoekstra (Hoekstra 1987) and expanded by Hadjivasiliou and Pomiankowski (2016). Common themes of mating-type loci were quickly noticed: they often come in two types (with notable exceptions in Fungi e.g. Billiard et al. (2011) for a review) with hardly any sequence conservation. While orthologous genes may be identified between the two mating-type regions, gametologs, mating type regions share little synteny as a consequence of rearrangements and insertion of repetitive DNA (Ferris and Goodenough 1994; Lengeler et al. 2002; Ferris et al. 2010; Badouin et al. 2015; Fontanillas et al. 2015; Hamaji et al. 2016; Geng et al. 2018). Moreover, mating-type loci may also experience recombination suppression both in diploid sexual system, as well as in haploid sexual systems and the UV sex chromosomes (Bachtrog et al. 2011)(Coelho et al. 2018). Recombination suppression may be stepwise and thus generate ‘evolutionary strata’ of differentiation between the two mating types (Hartmann et al. 2021 for a review in Fungi). The consequence of recombination suppression are manifold (Charlesworth and Charlesworth 2000) (Charlesworth 2016) and may include a higher probability of fixation of deleterious mutations, massive rearrangements, which may be associated to lower gene density (Yamamoto et al. 2021), GC composition changes, as well as differential gene expression. GC composition results from the balance between mutation biases, selection and GC biased gene conversion (Galtier et al. 2001), a molecular process linked to recombination. Therefore, regions with suppressed recombination are expected to display a significant lower GC content as compared to recombining regions, and a 4 to 10% lower GC content over the mating type locus has been reported in the mating type region of four species of volvocine algae (Hamaji et al. 2018).

The genomic features associated to mating type regions may thus guide the identification of candidate mating type loci in lineages in which genomic data is available, while the experimental conditions eliciting syngamy and meiosis have not yet been found, precluding experimental validation. While there is no direct evidence of sexual reproduction in the cosmopolitan marine picoeukaryote *Ostreococcus tauri* (Mamiellophyceae, Chlorophyta) there are three lines of indirect evidence for sexual reproduction (Grimsley et al. 2010). The first line of evidence comes from screening the whole genome sequence for genes encoding proteins involved in meiosis. These proteins have been described in all Mamiellophyceae species for which full genomes sequences are available, including *O. tauri* (Derelle et al. 2006), *O. lucimarinus* (Palenik et al. 2007), *Micromonas pusilla*, *M. commoda* (Worden et al. 2009), and

1
2
3 124 in *Bathycoccus* spp. metagenomes from the Arctic (Joli et al. 2017). The second line of evidence
4
5 125 comes from population polymorphism data that indicate inter-chromosomal and intra-
6
7 126 chromosomal recombination (Grimsley et al. 2010). Indeed, when sequencing can be performed
8
9 127 in several strains from the same population, analyses of the polymorphism spectrum allow the
10
11 128 estimation of the frequency of sex in natural populations (Tsai et al. 2008; Grimsley et al. 2010;
12
13 129 Drott et al. 2020; Hasan and Ness 2020; Koufopanou et al. 2020). Finally, the third line of
14
15 130 evidence comes from a population genomic analysis that demonstrated the existence of a
16
17 131 candidate mating type loci (450 and 650 kb) in *O. tauri* (Grimsley et al. 2010). *Ostreococcus*
18
19 132 *tauri* RCC4221 was suggested to represent the candidate *minus* mating type (hereafter *MT*-)
20
21 133 together with *O. lucimarinus* CCE9901, because of the presence of a gene encoding for a plant-
22
23 134 specific transcription factor from the RWP-RK gene family (Worden et al. 2009). This gene
24
25 135 family includes the “sex determining gene” (minus dominance *MID*) of minus mating type loci
26
27 136 in Volvocales algae (Ferris and Goodenough 1997; Umen 2011). The candidate opposite mating
28
29 137 type (hereafter *MT*+) was identified from the genome analysis of 12 *O. tauri* strains lacking
30
31 138 sequence homology with *O. tauri* RCC4221 over the 650 kb region. These strains also lacked
32
33 139 a gene containing an RWP-RK domain (Blanc-Mathieu et al. 2017). Phylogenetic analysis of
34
35 140 five gametologs revealed that *O. tauri* *MT*- and *MT*+ genes clustered with different
36
37 141 *Ostreococcus* species of the same mating type, respectively. This suggests that mating type
38
39 142 differentiation predates speciation within *Ostreococcus*, suggesting that *Ostreococcus* *MT*+ and
40
41 143 *MT*- are remarkably ancient. However, the total number of gametologs, their synteny and
42
43 144 sequence conservation among Mamiellales and Mamiellophyceae remains unknown.

44
45 145 Here, we investigated the architecture and phylogenetic profiles of the *MT*+ and *MT*-
46
47 146 alleles to unfold their evolutionary history. We analyzed the gene set of the two candidate
48
49 147 mating type loci, and identified the complete set of gametologs between them. This allowed us
50
51 148 to define the set of orthologous genes located inside each of the available candidate *MT* loci in
52
53 149 Mamiellales. This dataset was then leveraged (i) to investigate the presence of evolutionary
54
55 150 strata, (ii) construct gene genealogies to search for trans-specific evolution signatures (iii)
56
57 151 identify the opposite mating types from additional Mamiellophyceae sequence data. This
58
59 152 allowed to trace back the age of the divergence of the *MT*+ and *MT*- alleles in this early
60
61 153 diverging branch of the green lineage.

62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154

155 **Results**

156 **Sorting out gene families in *O. tauri* MT according to their prevalence across species**

157 The GC content can be used as a predictor of recombination rates in genomes
158 undergoing GC-biased gene conversion (e.g. Meunier and Duret 2004; Charlesworth et al.
159 2020), and it was suggested that there is an inverse relationship between chromosome length
160 and GC content, which is consistent with GC biased GC conversion in *Ostreococcus* (Jancek et
161 al. 2008). The genome-wide spontaneous mutation rate is GC->AT biased, which is consistent
162 with a mechanism like GC-biased gene conversion that could explain the difference between
163 the observed 0.60 GC frequency in the genome and the expected equilibrium 0.36 GC frequency
164 under mutation bias (Krasovec et al. 2017). The detection of the sharp (~9 to 17%) decrease in
165 GC content on the big outlier chromosome was used to define *MT* boundaries in *O. tauri*
166 RCC4221 (*MT*-), *O. tauri* RCC1115 (*MT*+), and six Mamiellales genomes (Figure 1,
167 supplementary table S1, Supplementary Material online). Using OrthoFinder, we assigned
168 genes from the *Ostreococcus* spp., *Bathycoccus prasinus*, *Micromonas commoda*, and
169 *Micromonas pusilla* to gene families (GFs). Mating type gene families were defined as GFs
170 with members located within the *MT* region of either *O. tauri* RCC4221 (*MT*-) or *O. tauri*
171 RCC1115 (*MT*+). The presence/absence of the genes of these GFs in the lineage provides
172 important information about *MT*+ and *MT*- specific GFs, as well as four additional distinct non-
173 overlapping GF categories (table 1).

174 **Table 1:** Classification, description, and quantities of genes and gene families (GFs) in *O. tauri* RCC4221 (*MT*-)
 175 and RCC1115 (*MT*+) strains.

Gene Family class	Features of included genes	RCC4221 (<i>MT</i> -)	RCC1115 (<i>MT</i> +)
<i>MT</i> specific GFs	Present in either all <i>Ostreococcus MT</i> - or all <i>Ostreococcus MT</i> +	6 genes in 6 GFs	2 genes in 2 GFs
Core <i>MT</i> GFs	Present in all Mamiellales genomes and located only in <i>MT</i> region	23 genes in 23 GFs	23 genes in 23 GFs
Shared <i>MT</i> GFs (non-core)	Present in both <i>Ostreococcus MT</i> loci, but not in all Mamiellales <i>MT</i> regions	75 genes in 69 GFs	79 genes in 69 GFs
GFs extending outside <i>MT</i>	Present in one <i>Ostreococcus MT</i> locus but with homologous genes in other regions in the opposite strain	28 genes in 27 GFs	8 genes in 4 GFs
GFs not retained for analysis	Present in only one <i>Ostreococcus MT</i> locus and Mamiellales genomes but absent from the genomes of the opposite strains/ <i>MT</i> ; divergent GFs or singletons	112 genes	128 genes
Total number of genes		244	240

176

177 The “*MT* specific” GF class contains genes that are shared only by *Ostreococcus*
 178 genomes from the same *MT*. The *MT*-specific GFs contain the smallest number of genes: 6 and
 179 2 genes for *MT*- and *MT*+, respectively. These GFs are expected to contain genes involved in
 180 sex determination and functional control associated with each *MT*, as well as dispensable genes
 181 trapped into this locus (Wilson et al. 2019 for a review in Ascomycetes). Functional annotation
 182 revealed that most of these genes encode for hypothetical proteins or do not have any predicted
 183 function. The *MT*- specific GFs contain a gene with an RWP-RK domain (ostta02g01710), as
 184 previously reported (Worden et al. 2009), and a gene (ostta02g00990) that encodes for an SRP-
 185 dependent co-translational protein involved in targeting proteins to the membrane. Within the
 186 *MT*+-specific GFs, there are only 2 genes, which encode for hypothetical proteins annotated
 187 with Gene Ontology terms linked to mismatch repair, protein binding, and transport
 188 (supplementary table S2, Supplementary Material online).

189 The “core *MT*” GF class contains GFs exclusively composed of gametologs that are
 190 located inside the boundaries of all candidate *MT* regions in all eight Mamiellales genomes
 191 (supplementary table S1, Supplementary Material online). There are 23 “core *MT*” GF, which
 192 make up less than 10% of genes of the *MT* (Table 1) and these likely belonged to the ancestral
 193 locus which evolved into a *MT* in the lineage. Functional annotation indicates that these genes
 194 have housekeeping functions, such as ATP and DNA binding, transcription, glycolipid
 195 biosynthesis, protein transport, and RNA methylation, but no obvious link to mating
 196 (supplementary table S3, Supplementary Material online).

197 The largest GF class (69 GFs) regroups gametologs that are shared by both
 198 *Ostreococcus MT* loci, and that can be absent from the *MT* regions in some Mamiellales species

1
2
3 199 (Shared *MT* GFs, non-core). A fourth class of GFs contains genes located within the *O. tauri*
4 200 *MT* locus or on standard chromosomes (GF extending outside *MT*), and provides evidence of
5 201 translocations between standard chromosomes and the *MT* loci. The remaining GFs are present
6 202 in only one *O. tauri* *MT* locus and other Mamiellales genomes, or contain genes that are too
7 203 divergent to generate phylogenies, as the alignments are too short. Therefore, they were
8 204 excluded from further analyses, together with singleton genes (except the *MT*-specific GFs).

9
10 205 While the core and specific GFs categories should contain the most ancient genes on
11 206 the *MT*, the other GF categories likely reflect gain, loss, and translocation of genes in and out
12 207 of the *MT*. This prompted us to undertake synteny and phylogenetic profiling of each GF to
13 208 understand its evolutionary dynamics.

21 209 **Genomic architecture of *O. tauri* mating type regions**

22
23 210 Syntenic regions outside the *MT* loci have been reported between species of the same
24 211 genus : *O. tauri* and *O. lucimarinus* (Palenik et al. 2007), *M. pusilla* and *M. commoda* (Worden
25 212 et al. 2009). Within *O. tauri*, regions outside the *MT* locus have been shown to be perfectly
26 213 syntenic and share >99% nucleotide identity, in sharp contrast with the *MT* region (*O. tauri*
27 214 Chromosome 2, fig. 1), which cannot be aligned at the nucleotide level between *MT*-
28 215 (RCC4221) and *MT*+ (RCC1115) (Blanc-Mathieu et al. 2017). We further investigated the
29 216 relative position of orthologous genes in the *MT*+ and *MT*- regions, but found no evidence for
30 217 synteny in genes from shared and core GFs between both regions (fig. 2A): *MT* specific genes
31 218 do not cluster but are interspersed throughout the *MT*+ and *MT*- loci.

32
33 219 Ancient inversion events are a well-known trigger for suppression of recombination in
34 220 genome evolution, but the relative position of orthologous genes in *MT*- and *MT*+ regions
35 221 provide no evidence of a past inversion event. Instead, visual examination of the global pattern
36 222 suggested a large translocation of the [b,c] segment in 5' followed by the [a,b] segment in 3'
37 223 (fig. 2A). To investigate this hypothesis, we defined a simple statistic, *Sdist*, based on the
38 224 relative distance between orthologous genes on the *MT*+ and *MT*-: *Sdist* is equal to 0 for perfect
39 225 co-linearity (see methods). Random permutations of the gene orders enabled the estimation of
40 226 the null distribution. The observed *Sdist* was not significantly different from the average *Sdist*
41 227 for orthologous genes placed randomly on the two *MT*s (10,000 permutations, $p > 0.10$).
42 228 However, the translocation of the 5' extremity of *MT*- (segment [b,c]) to the start of *MT*- (arrow
43 229 on fig. 2A) was associated with a significantly smaller *Sdist* than the random *Sdist* (100,000
44 230 permutations $p=0.0054$). This demonstrates that this translocation significantly improves the

231 overall co-linearity between MT^+ and MT^- , supporting the idea of a past large-scale
232 translocation in one of the MT loci.

233 To track gene translocation events between the MT s and the autosomal regions, we
234 located the positions of 46 (MT^-) and 30 (MT^+) genes from GFs sharing genes inside and
235 outside the MT regions. Genes of the same GFs as MT^- genes were located on diverse autosomes
236 (fig. 2B). We also observed a similar patchy distribution for GFs of gene members extending
237 outside the MT^+ (fig. 2C). This provides evidence for past gene translocations between many
238 autosomes and the MT regions.

239 To search for evidence of evolutionary strata, defined as discrete regions containing
240 orthologous genes with similar substitution rates (Lahn and Page 1999), we computed the rate
241 of synonymous substitutions (Ks) (Tzeng et al. 2004) of the genes belonging to the 69 shared
242 MT GFs on MT^- and MT^+ in *O. tauri* (shared GFs). We were able to compute the number of
243 non-synonymous substitutions (Ks) for only 22 gene pairs, given that for other gene pairs Ks
244 values were close to saturation. From these 22, 19 had a Ks < 1, and only 2 were adjacent on
245 both the MT^+ and the MT^- (supplementary table S4, Supplementary Material online). This is
246 consistent with a scenario of independent gene conversion events between the two MT s, except
247 for one event spanning two genes. Interestingly, within these recently diverged genes, only 2
248 pairs were adjacent in only one of the mating types (MT^+). This suggests that the source or the
249 destination of the conversion events between MT s tends to span several kb. These observations
250 indicate an absence of evidence for strata throughout the large MT regions of *O. tauri*. However,
251 this absence of evidence may be reconsidered in the future if additional genome data in novel
252 species can be informative to infer the ancestral gene order on the mating type (Branco et al.
253 2017).

254 **Phylogenetic insights into evolutionary dynamics of mating types**

255 The topology of each GF phylogenetic tree is informative about the relative chronology
256 of the speciation and the divergence events between the MT^+ and MT^- alleles. We assessed
257 whether the topology supported either of the two scenarios: (i) in the “*mating type allele*
258 *diverged post speciation*” scenario: mating type alleles diverged after speciation events within
259 *Ostreococcus* (no mating type alleles = Post); or (ii) in the “*mating type allele diverged ante*
260 *speciation*” scenario: mating type alleles diverged prior to the speciation event (mating type
261 allele separation = Ante). This later scenario has previously been coined as trans-specific
262 evolution resulting from long term balancing selection (Richman 2000). Consequently, the
263 variation within the genes following the “Ante” scenario may be named trans-specific

1
2
3 264 polymorphisms (Devier et al. 2009). The number of GFs for each topology is displayed in fig.
4
5 265 3. Interestingly, this dual phylogenetic signal (mating type allele divergence ante versus post-
6
7 266 speciation) is mirrored by a GC3 content signature of the genes. Indeed, genes belonging to
8
9 267 GFs that support ancient mating type origin have a significantly lower GC3 content than genes
10
11 268 whose evolutionary history is concordant with the speciation history of the genus. For the 23
12
13 269 core *MT* GFs (listed in supplementary table S3, Supplementary Material online), the majority
14
15 270 of phylogenies (21 trees, supplementary fig. S1, Supplementary Material online) support the
16
17 271 “ancient mating type” evolutionary scenario that mating type region diverged before the
18
19 272 speciation events within *Ostreococcus*, whereas only two phylogenies support the scenario of
20
21 273 a mating type differentiation after the speciation events.

22
23 274 Thus, most core and shared *MT* GFs support an ancient mating type origin (fig. 4A with
24
25 275 mating type separation and 3B without mating type separation). In contrast, the phylogenies of
26
27 276 most GFs containing paralogous genes outside the *MT* region are consistent with the speciation
28
29 277 tree, suggesting their translocation inside the *MT* locus occurred recently.

28 278 **Expanding the number of Mamiellales species with two mating type alleles**

30
31 279 Since the core *MT* GFs allow *MT*⁺ and *MT*⁻ delineation in the Mamiellales, we used the
32
33 280 sequence data to screen 33 transcriptomes (MMETSP and 1KP datasets) from several
34
35 281 Mamiellophyceae species for homologous sequences (listed in supplementary table S5,
36
37 282 Supplementary Material online). The taxonomic affiliation of each transcriptome was inferred
38
39 283 from 18S rDNA sequences (supplementary table S6 and supplementary fig. S2, Supplementary
40
41 284 Material online). The phylogenetic range of the transcriptomes spanned from the early
42
43 285 divergent freshwater species, such as *Monomastix opisthostigma* (Monomastigales),
44
45 286 *Crustomastix*, and *Dolichomastix* (Dolichomastigales), to early Mamiellales, such as
46
47 287 *Mantoniella*. It also included several *Micromonas* strains from novel species, such as *M. bravo*
48
49 288 and *M. polaris*. In total, at least one homologous gene was recovered for each GF (with an
50
51 289 average of 11 GFs per transcriptome) in 28 of 33 transcriptomes (fig. 5).

290

291 The most striking pattern came from *O. mediterraneus* MMETSP0929 (strain
292 RCC2572) and *O. lucimarinus* MMETSP0939 (strain BCC118000) transcriptomes. While both
293 datasets displayed hits for almost all core genes (17 out of 23), the taxonomic affiliation inferred
294 for these genes by best blast hit (BBH) was not consistent with the 18S taxonomic affiliation.
295 Instead, it suggested affiliation to a different species of the opposite mating type (supplementary
296 fig. S3, Supplementary Material online). In *O. mediterraneus* MMETSP0929, 14 of 17 genes
297 were affiliated to species from the opposite *MT* groups (*MT*-), such as *O. tauri* and *O.*
298 *lucimarinus*, not to the reference genome *O. mediterraneus* RCC2590 *MT*+. Likewise, 15 of
299 17 best blast hits of *O. lucimarinus* MMETSP0939 came from *MT*+ genomes, and not from the
300 *MT*- *O. lucimarinus* reference genome. To confirm the taxonomic affiliation of these genes, we
301 built maximum likelihood phylogenies, including homologs extracted from the transcriptomes
302 (supplementary fig. S3, Supplementary Material online). From the 17 gene families with a best
303 blast hit, 12 passed the alignment length and identity thresholds (see methods). Of these, 10
304 phylogenies included both *O. mediterraneus* MMETSP0929 and *O. lucimarinus*
305 MMETSP0939, and two phylogenies included only *O. lucimarinus* MMETSP0939. From
306 these, 11 phylogenies were consistent with ancient *MT*+ and *MT*- divergence (example in fig.
307 6A), while one phylogeny regrouped genes according to species (fig. 6B).

308

309 These phylogenetic analyses confirmed the taxonomic affiliation inferred from amino-
310 acid sequence conservation and support an ancient divergence of genes from two *MT* regions.
311 This led us to conclude that *O. lucimarinus* strain RCC2572 and *O. mediterraneus* strain
312 BCC118000 (MMETSP0929 and MMETSP0939, respectively) are of the opposite mating type
313 to the strains for which the reference genome is available. This extends the evidence of the
314 existence of two mating types in *O. tauri* to two additional *Ostreococcus* species.

315

1
2
3 316 **Identification of candidate mating types based on gene genealogies in *Micromonas***
4
5 317 ***commoda***
6

7
8 318 *Micromonas* is the most represented Mamiellophyceae genus in the available
9
10 319 transcriptomic datasets, with 14 transcriptomes. Therefore, we further examined the individual
11
12 320 GF phylogenetic topologies and sequence similarities by using the core *MT* GF set (23 GFs) to
13
14 321 search for clustering that might suggest an ancient divergence of *MTs* in *Micromonas*. To this
15
16 322 end, we selected *Micromonas* transcriptomes with more than one positive hit with the GFs, and
17
18 323 the highest number of hits in the majority of transcriptomes (9 transcriptomes), together with
19
20 324 one outgroup from the genus (*Mantoniella* sp. MMETSP1468). Finally, we built individual GF
21
22 325 phylogenies from these sequences and the core genes GF dataset (supplementary fig. S4,
23
24 326 Supplementary Material online).

25 327 A consistent sub-clustering of strains within the *Micromonas commoda* group was observed.
26
27 328 MMETSP 1084, 1387, 1403, and 1400 clustered together in 11 of 13 phylogenies, while
28
29 329 MMETSP1404 and 1393 clustered with genes from the reference genome of *M. commoda*
30
31 330 RCC299 (fig. 7A and supplementary fig. S4, Supplementary Material online). In only two
32
33 331 phylogenies, there was no apparent sub-clustering (fig. 7C). Additionally, the branch lengths of
34
35 332 the 11 phylogenies displaying sub-clustering were longer and similar to the branch lengths
36
37 333 separating *M. polaris* from *M. bravo*, or *M. commoda* from *M. pusilla*. Consistent with this, the
38
39 334 average pairwise amino-acid identities between *M. commoda* genes from the two different sub-
40
41 335 clusters ranged from 65% to 89% (supplementary table S7, Supplementary Material online).
42
43 336 For comparison, we built phylogenies of the actin and β -tubulin genes (fig. 7B and 7D), which
44
45 337 are highly conserved, and their phylogenetic topology showed a species topology signature,
46
47 338 where these strains did not support two sub-clusters. Pairwise amino-acid identity for the latter
48
49 339 GFs between strains ranged from 98% to 99.4% (for actin and β -tubulin, respectively), as
50
51 340 expected for strains from the same species. This phylogenetic signal was similar to the
52
53 341 *Ostreococcus* core GF phylogenies, consistent with an ancient mating type separation (fig. 4A).
54
55 342 Despite the low number of genes (13 genes from 23 GFs), this sub-clustering suggests that there
56
57 343 are two *MTs* in *Micromonas commoda*: strains MMETSP1404, 1393, and *M. commoda*
58
59 344 RCC299 (the reference genome); and strains MMETSP 1084, 1387, 1400, and 1403,
60
345 representing the opposite *MT*. As Worden et al. (2009) suggested, *M. commoda* RCC299 would
346 represent the *MT*⁻, given the presence of an RWP-RK motif gene in its candidate *MT* region.
347 Thus, the strains MMETSP 1084, 1387, 1403, and 1400 would represent the *MT*⁺ type. Taken

1
2
3 348 together, phylogenetic analyses of GFs are consistent with an ancient gene divergence of *MT*
4 349 gametologs in the *M. commoda* lineage, as expected under recombination suppression.

7 350 **Clues about earlier origin of Mating Type loci in Mamiellophyceae**

9
10 351 As the phylogenetic signal may be lost over time as a consequence of the decay of
11 352 similarity between orthologs (Jain et al. 2019), we investigated indirect signatures of *MTs*. *MTs*
12 353 evolve without recombination, and this has been shown to decrease GC content. We therefore
13 354 investigated whether a GC signature could be detected in homologous genes to the core GFs
14 355 outside the Mamiellales (comprising *Ostreococcus*, *Bathycoccus* and *Micromonas*). Thus, we
15 356 analysed the GC content of the synonymous third codon position (GC3) of core GF hits in
16 357 several Mamiellophyceae species, and compared this to the GC3 content of genes from the
17 358 background genome or transcriptome. Core *MT* GFs have significantly lower GC3 (around
18 359 20%) than genes of the background genome (or transcriptome) in *Bathycoccus*, *Ostreococcus*
19 360 and *Micromonas* (fig. 8 and supplementary table S8, Supplementary Material online).
20 361 Interestingly, we found evidence of a similar difference in GC3 content between gene hits
21 362 against the core *MT* GFs and the background transcriptome in *Mantoniella squamata* CCAP
22 363 1965/1 and the uncultured Mamiellophyceae (uncultured eukaryote RCC2288), with ~10% and
23 364 20% differences between genes from the GFs and genes from the background transcriptome,
24 365 respectively. This suggested that genes that are homologous to the core GFs are also located in
25 366 a low GC chromosome region in these Mamiellophyceae species (fig. 8 and supplementary
26 367 table S8, Supplementary Material online). However, there is no evidence for a GC3 content
27 368 difference between homologous genes to the core GFs and the genes from the background
28 369 transcriptome in *Crustomastix* or *Monomastix* (fig. 8).

29
30
31
32
33
34
35
36
37
38
39
40
41
42 370
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

371 Discussion

372 Direct evidence of meiosis is not available for most marine planktonic microbial
373 eukaryotes. This is either due to the difficulty in culturing certain species, or because
374 experimental studies are hampered by a lack of knowledge about sex determination and the
375 conditions required to induce a sexual cycle. In the case of haploid green picoalgae (cell
376 diameter $< 2 \mu\text{m}$) of the Mamiellales lineage, population genomics data in one species allowed
377 the identification of two candidate mating type alleles with suppressed recombination (Blanc-
378 Mathieu et al. 2017). Here, comparative genomics of seven related species within the
379 Mamiellales lineage unravelled different facets in the mode and tempo of evolution in this
380 enigmatic locus.

381 First, while no *MT+* and *MT-* specific genes could be identified for all seven species,
382 *MT+* and *MT-* specific genes could be identified within the *Ostreococcus* genus. *MT-* specific
383 genes may be implicated in mating type differentiation, such as the previously identified gene
384 encoding an RWP-RK domain (Worden et al. 2009). The two *MT+* specific genes that have
385 been identified in *Ostreococcus* encode for unknown proteins. One of these proteins
386 (gm1.767_g) harbours WD40 repeats and is predicted to bind to other proteins. The second
387 protein has a DNA binding domain, which is also found in DNA mismatch repair proteins
388 (gm1.689_g, PF00488). A WD40 protein has been shown to regulate mating in the fungus
389 *Ustilago maydis* (Wang et al. 2011). Nevertheless, the functional range of WD40 proteins is
390 too wide to confidently infer a role of the *Ostreococcus* protein to act as a *MT+* signal protein.

391 Second, comparative phylogenetics of core gametologs allowed the identification the
392 opposite mating types in two additional species for which transcriptomes were available: the
393 *MT-* in *O. mediterraneus* and the *MT+* in *O. lucimarinus*. This mating type profiling is made
394 possible by the high divergence between the *MT+* and *MT-* regions, as gametologs cluster by
395 *MT* and not by species. By screening available environmental data from the TARA Oceans
396 project for the presence of these gametologs, we previously found that, in fact, both mating
397 types of *O. lucimarinus* were present at the stations where this species had been detected
398 (Leconte et al. 2020). Mating type profiling was also suggested between strains from *M.*
399 *commoda*: phylogenies of the gametologs suggest two clusters of strains, in contrast with
400 phylogenies of highly conserved housekeeping genes (actin, β -tubulin, and 18s rDNA) (fig. 7;
401 supplementary table S7 and supplementary fig. S2, Supplementary Material online).

402 Third, analysing additional transcriptome data from early diverging branches of the
403 Mamiellophyceae class, we could detect orthologous genes to the Mamiellales gametologs in

1
2
3 404 eight additional transcriptomes. However, we could not detect any significant difference in GC3
4
5 405 signatures in the earliest Mamiellophyceae, as would be expected under suppressed
6
7 406 recombination; on the contrary, GC3 values appear to be higher in homologous genes in
8
9 407 Dolichomastigales. This suggests the Mamiellales gametologs are not part of a lower GC region
10
11 408 in earlier branching Mamiellophyceae. The conservation of 23 gametologs within the
12
13 409 Mamiellales lineage prompted us to investigate the dynamic of these genes. The additional
14
15 410 gametologs within the *Ostreococcus* lineage support an ancient large translocation event.
16
17 411 Inversions have been previously suggested to trigger recombination suppression and have been
18
19 412 recently reported in the origin of a young sex-determining chromosome (Natri et al. 2019).
20
21 413 However, translocations are also expected to disrupt recombination (McKim et al. 1988).

22
23 414 One intriguing feature of sex determining chromosomes is their organization as multiple
24
25 415 discrete regions, where genes can be clustered by genetic divergence (measured by the rate of
26
27 416 non-synonyms substitutions), defined as “evolutionary strata”. In humans, strata were first
28
29 417 described by Lahn and Page (1999), who suggested that suppression of recombination was
30
31 418 initiated in one region (stratum) and later expanded in discrete steps, by strata. This could
32
33 419 happen through additional chromosomal inversions, which are known to suppress
34
35 420 recombination in mammalian chromosomes. Only a few X-Y sequence similarities persist, and
36
37 421 these alleles are orderly stratified by age in the X chromosome and scrambled in the Y.
38
39 422 Although strata have been observed in several vertebrates, plants, and fungi (Bachtrog et al.
40
41 423 2014; Badouin et al. 2015; Coelho et al. 2018), they do not appear to be a common feature of
42
43 424 algal mating types and sex chromosomes. Indeed, we found no evidence of evolutionary strata
44
45 425 in *Ostreococcus* MTs, as neither ancient nor recent genes cluster in any of the MTs This may be
46
47 426 due to their ancient divergence, associated with a limited more recent expansion dynamic, as
48
49 427 suggested in the UV chromosomes of the brown algae *Ectocarpus* (Ahmed et al. 2014).
50
51 428 Alternatively, it could also be due to the lack of information about the ancestral gene order on
52
53 429 the mating type (Branco et al. 2017).

54
55 430 To counteract the effects of reduced recombination inside MTs, gene conversion
56
57 431 between mating types has been suggested to act as a homogenizing force in *Chlamydomonas*
58
59 432 (De Hoff et al. 2013). In fungal mating types, the suppression of recombination maintains
60
433 linkage of mating-type genes within each locus, which is required for correct mating-type
434
435 determination (Kües 2000; Branco et al. 2017). However, gene flow between mating type loci
436
and gene conversion events have recently been reported in several species (Sun et al. 2012;
Hartmann et al. 2020). This suggests an important difference in the evolutionary processes of

haploid sex determining systems versus diploid sex determining systems, where gene flow between sex determining regions is rare (Hartmann et al. 2020).

The diversification within Mamiellales is estimated to have occurred between 330 and 640 million years ago (Lang et al. 2010)(Blank 2013)(Parfrey et al. 2011), much earlier than the diversification within Volvocales where deep homology of mating type loci has been reported (Ferris et al. 2010), and with a higher upper limit to the estimated 370 million years divergence of the STE3-like pheromone receptors from basidiomycete fungi (Devier et al. 2009). Therefore, our data suggest the Mamiellales mating type sex-determining region to be among the oldest mating type reported.

In conclusion, we analysed the phylogenetic profiles of the gene families within the *Ostreococcus* mating types, and gained insights into the evolutionary history of this sex-determining region in one of the earliest diverging orders of Chlorophytes. The identification of strains from the two opposite mating types in three species will guide future experimental approaches for mating and strain crossing, since a highly efficient transformation protocol is now available in *Ostreococcus* (Sanchez et al. 2019). Complete genome sequences in additional Mamiellophyceae are now essential to investigate the early dynamics of the sex-determining regions in the green lineage.

Materials and Methods

Mating type gene family definition

The full set of predicted genes from eight Mamielliales genomes (supplementary table S1, Supplementary Material online) was loaded into a custom version of the pico-PLAZA framework (Proost et al. 2009; Vandepoele et al. 2013) to define and analyse gene families (GFs). Following an ‘all-against-all’ protein sequence similarity search, performed with BLASTP (version 2.6.0+, maximum E-value threshold 1e-4, keeping up to 2,500 hits), we delineated GFs using OrthoFinder version 2.1.2 (Emms and Kelly 2015).

The boundaries of the mating type (*MT*) region of *Ostreococcus tauri* RCC4221 and RCC1115 served as a starting point for defining candidate *MT* GFs (supplementary table S1, Supplementary Material online). All genes located within either *MT* region were extracted, based on the coordinates of their coding sequence (CDS). For each gene included in these two gene sets, the GF they were assigned to was subsequently retrieved, consisting of a validated homologous group of ortholog and paralog genes in eight available genomes. Based on the location of the GF members (chromosome or scaffold and coordinates), a ‘*MT* signal’ value

470 was then computed for every genome in which the GF was represented. This value corresponds
 471 to the fraction of members located within the *MT* region (for the given genome-GF
 472 combination), and was used to filter and classify the list of candidate GFs. The complete list of
 473 *MT* GFs is reported in supplementary table S9, Supplementary Material online.

474 For every retained GF, protein sequences were aligned using MAFFT version 7.187
 475 (Kato and Standley 2013) with the L-INS-i alignment method and a maximum of 1,000
 476 iterative refinements. We edited the multiple sequence alignments (MSAs) using several filters
 477 on both sequences and positions, implemented in the PLAZA framework and described by
 478 Proost (Proost et al. 2009). Briefly, highly divergent and partial sequences were filtered out,
 479 and positions containing gaps in minimum 10% of the sequences or containing potentially
 480 misaligned amino acids removed. We also applied a minimum length cut-off to the edited MSA:
 481 the edited MSA had to be 50-amino-acid-long at least, otherwise we ignored it. In case the
 482 original unedited MSA was shorter, we used this length as a cut-off value instead. Finally, we
 483 retained only MSAs that showed at least 50% alignment of amino-acid identity in half of the
 484 sequences of the MSA. The circular plots depicting the location of homologous genes from GFs
 485 having copies outside of the *O. tauri MT* loci (fig. 2B and 2C) were generated with the circlize
 486 package in R (Gu et al. 2014; <https://r-project.org/>).

487 To test different gene order rearrangement scenarios between the *MT+* and *MT-* regions, we
 488 defined S_{dist} , which is the absolute value of the difference of the position of orthologous genes
 489 on the *MT+* and *MT-* regions. If there are n orthologous genes between the two loci with p_i-
 490 the position (in rank) of gene i on *MT-* and p_i+ the position of its ortholog on *MT+*, $S_{dist} =$
 491 $\sum_{i=1}^n |p_i- - p_i+|$. $S_{dist}=0$ if all orthologs are perfectly collinear. The expected S_{dist} under
 492 random position of orthologous genes in the two mating types was assessed by simulations. If
 493 there has been an inversion of gene order between the two regions, S_{dist} is maximal, $S_{dist}=z(2n-$
 494 $2z)$, with $z=n/2$ if n is even, and $z=(n-1)/2$ if z is odd.

495 Gene family clustering and phylogeny

496 For each GF MSA that passed our filtering criteria, we built a Maximum Likelihood
 497 (ML) phylogenetic tree using IQ-TREE version 1.6.5 (Nguyen et al. 2015). Trees were built
 498 under the best-fitting substitution model selected by ModelFinder (Kalyaanamoorthy et al.
 499 2017), chosen among commonly used models (JTT, LG, WAG, Blosum62, VT, and Dayhoff).
 500 Empirical amino-acid frequencies were calculated from the data, the FreeRate model (Yang
 501 1995; Soubrier et al. 2012) was used to account for rate heterogeneity across sites, and branch

1
2
3 502 supports were assessed using ultrafast bootstrap approximation (UFBoot) (Soubrier et al. 2012)
4 503 with 1,000 bootstrap replicates.

5 504 We used similar alignment, MSA editing, and phylogenetic tree building procedures
6 505 when considering sequences from external sources (e.g. transcripts from MMETSP samples).
7 506 The divergent gene removal criterion was based on the results of the all-against-all protein
8 507 sequence similarity search performed using data from the eight reference genomes only
9 508 (supplementary table S1, Supplementary Material online). Therefore, it was not used to filter
10 509 out these sequences from the MSAs. Phylogenetic trees were built for full alignments in case
11 510 the editing was deemed too stringent, for instance discarding transcripts flagged as partial
12 511 sequences. Finally, when investigating the molecular phylogeny of the 18S rDNA genes, we
13 512 used IQ-TREE's ModelFinder Plus parameter to select the best DNA substitution model.

23 513 **Gene family phylogenetic tree classification**

24 514 We visualised and inspected the *MT* GF trees using FigTree version 1.4.4
25 515 (<http://tree.bio.ed.ac.uk/software/figtree/>). We examined ultrafast bootstrap support values and
26 516 topology type, and counted the number of times genes clustered by mating type or according to
27 517 their taxonomic classification (by species).

33 518 **Searching for homologs in publicly available transcriptomes**

34 519 We used sequences of core *MT* GF members as queries to search for homologs in
35 520 Mamiellophyceae transcriptomes (33 transcriptomes in total, listed in supplementary table S5,
36 521 Supplementary Material online). Transcriptomes were retrieved from the MMETSP (Keeling
37 522 et al. 2014; Johnson et al. 2019) and 1KP datasets (Matasci et al. 2014). Re-assembled
38 523 MMETSP transcriptomes were downloaded from <https://doi.org/10.5281/zenodo.251828>
39 524 (version 1; January 2017) and 1KP transcriptomes via 1KP's R interface
40 525 (<https://github.com/ropensci/onekp>). CDS from each Mamiellophyceae MMETSP
41 526 transcriptome were predicted using TransDecoder (Haas et al. 2013) with default parameters.
42 527 Sequence similarity searches were performed using tblastx (maximum E-value threshold 1e-4)
43 528 and results were filtered to retain hits with alignment length > 50 and amino-acid identity >
44 529 60%. In-depth phylogenetic analyses of individual hits from *O. mediterraneus* strain RCC2572
45 530 (MMETSP0929), *O. lucimarinus* strain BCC118000 (MMETSP0939), *Micromonas* MMETSP
46 531 transcriptomes (1084, 1327, 1387, 1393, 1400, 1401, 1402, 1403, 1404), and *Mantoniella*
47 532 MMETSP transcriptomes (1106, 1468) were performed as previously described for the
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 533 reference genomes. The presence/absence matrix of each informative orthologous group against
4 534 the transcriptomes was generated using the ggplot2 package in R) (Wickham 2011).

5 535 To validate and elucidate each MMETSP transcriptome's taxonomic affiliation, we
6 536 downloaded Mamiellophyceae 18S rDNA sequences from reference genomes in GenBank, the
7 537 SILVA database (Wickham 2011), and *Micromonas* spp. sequences provided in Simon et al.
8 538 (2017) (supplementary table S6, Supplementary Material online). Transcripts matching selected
9 539 18S sequences were extracted with blastn (maximum E-value 1e-5) and 18S rDNA sequences
10 540 were subsequently predicted using RNAMmer (Lagesen et al. 2007). A ML phylogenetic tree
11 541 was built using IQ-TREE and following each clustering of this Mamiellophyceae reference tree
12 542 (rooted in *Monomastix* spp.), transcriptomes were tentatively classified according to a species
13 543 clustering (supplementary fig. S2, Supplementary Material online). Phylogeny indicated that
14 544 MMETSP transcriptomes matched their species classification, and transcriptomes from novel
15 545 *Micromonas* species as *M. polaris* and *M. bravo* were designated using the data and new
16 546 classification of (Simon et al. 2017).

27 547 **Compositional analysis (GC3) of gene families in Mamiellophyceae**

28 548 To evaluate compositional differences between third codon positions (GC3) of GF
29 549 members and CDS from the overall genome or transcriptome (supplementary table S8,
30 550 Supplementary Material online) we used a custom python script to perform GC3 calculations.
31 551 We subsequently evaluated the results using Student's *t*-test as implemented in R.

32 552 **Synonymous and non-synonymous divergence of shared *MT* gene families**

33 553 We used homologous pairs of the 69 shared *MT* GFs to calculate sequence genetic
34 554 divergence with the seqinr package v3.4-5 (kaks function) using (Li 1993) method (LWL85) in
35 555 R.

36 556

37 557 **Acknowledgements**

38 558 This work was funded by the European Union's Horizon 2020 research and innovation
39 559 programme under the Marie Skłodowska-Curie ITN project SINGEK (H2020-MSCA-ITN-
40 560 2015-675752 to L.F.B. and F.B.). We thank the Moore foundation for sequencing most of the
41 561 Mamiellophycean transcriptomes analysed in this study and the Genotoul Bioinformatic
42 562 platform for providing computing and data storage resources.

43 563

44 564 **Data availability statement**

45 565 All genomic and transcriptomic sequence data is available on GenBank under accession number
46 566 CAID00000000.1 (*O. tauri*), PRJNA337288 (*O. lucimarinus*), PRJNA15676 (*M. commoda*),

47 567

1
2
3 567 PRJNA15678 (*M. pusilla*), PRJNA394752 (*B. prasinus*), PRJNA248394 (MMESTP). The
4 568 accession numbers of the 18S rDNA sequences are summarized in Supplemental Table S6.
5 569
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Downloaded from <https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evab216/6380139> by BIUS Jussieu user on 07 October 2021

1
2
3 **570 References**

- 4
5 571 Ahmed S, Cock JM, Pessia E, Luthringer R, Cormier A, Robuchon M, Sterck L, Peters AF,
6 572 Dittami SM, Corre E, et al. 2014. A haploid system of sex determination in the brown alga
7 573 *Ectocarpus* sp. *Curr. Biol.* CB 24:1945–1957.
- 8
9 574 Astell CR, Ahlstrom-Jonasson L, Smith M, Tatchell K, Nasmyth KA, Hall BD. 1981. The
10 575 sequence of the DNAs coding for the mating-type loci of *Saccharomyces cerevisiae*. *Cell*
11 576 27:15–23.
- 12 577 Bachtrog D, Kirkpatrick M, Mank JE, McDaniel SF, Pires JC, Rice W, Valenzuela N. 2011.
13 578 Are all sex chromosomes created equal? *Trends Genet.* 27:350–357.
- 14 579 Bachtrog D, Mank JE, Peichel CL, Kirkpatrick M, Otto SP, Ashman T-L, Hahn MW, Kitano J,
15 580 Mayrose I, Ming R, et al. 2014. Sex determination: why so many ways of doing it? *PLoS*
16 581 *Biol.* 12:e1001899.
- 17 582 Badouin H, Hood ME, Gouzy J, Aguilera G, Siguenza S, Perlin MH, Cuomo CA, Fairhead C,
18 583 Branca A, Giraud T. 2015. Chaos of Rearrangements in the Mating-Type Chromosomes of
19 584 the Anther-Smut Fungus *Microbotryum lychnidis-dioicae*. *Genetics* 200:1275.
- 20 585 Beutlich A, Schnetter R. 1993. The Life Cycle of *Cryptochlora perforans* (Chlorarachniophyta).
21 586 *Bot. Acta* 106:441–447.
- 22 587 Billiard S, López-Villavicencio M, Devier B, Hood ME, Fairhead C, Giraud T. 2011. Having
23 588 sex, yes, but with whom? Inferences from fungi on the evolution of anisogamy and mating
24 589 types. *Biol. Rev. Camb. Philos. Soc.* 86:421–442.
- 25 590 Billiard S, López-Villavicencio M, Hood ME, Giraud T. 2012. Sex, outcrossing and mating
26 591 types: unsolved questions in fungi and beyond. *J. Evol. Biol.* 25:1020–1038.
- 27 592 Blanc-Mathieu R, Krasovec M, Hebrard M, Yau S, Desgranges E, Martin J, Schackwitz W,
28 593 Kuo A, Salin G, Donnadiou C, Desdevises Y, Sanchez-Ferandin S, Moreau Hervé, et al.
29 594 2017. Population genomics of picophytoplankton unveils novel chromosome
30 595 hypervariability. *Sci. Adv.* 3:e1700239.
- 31 596 Blanc-Mathieu R, Verhelst B, Derelle E, Rombauts S, Bouget F-Y, Carré I, Château A, Eyre-
32 597 Walker A, Grimsley N, Moreau H, et al. 2014. An improved genome of the model marine
33 598 alga *Ostreococcus tauri* unfolds by assessing Illumina de novo assemblies. *BMC Genomics*
34 599 15:1103.
- 35 600 Blank CE. 2013. Origin and early evolution of photosynthetic eukaryotes in freshwater
36 601 environments: reinterpreting proterozoic paleobiology and biogeochemical processes in
37 602 light of trait evolution. *J. Phycol.* 49:1040–1055.

- 1
2
3 603 Branco S, Badouin H, Rodríguez de la Vega RC, Gouzy J, Carpentier F, Aguileta G, Siguenza
4 604 S, Brandenburg J-T, Coelho MA, Hood ME, et al. 2017. Evolutionary strata on young
5 605 mating-type chromosomes despite the lack of sexual antagonism. *Proc. Natl. Acad. Sci.*
6 606 114:7067.
- 7
8
9
10 607 Charlesworth B, Charlesworth D. 2000. The degeneration of Y chromosomes. *Philos. Trans. R.*
11 608 *Soc. B Biol. Sci.* 355:1563–1572.
- 12
13 609 Charlesworth D. 2016. Plant Sex Chromosomes. *Annu. Rev. Plant Biol.* 67:397–420.
- 14
15 610 Charlesworth D, Zhang Y, Bergero R, Graham C, Gardner J, Yong L. 2020. Using GC Content
16 611 to Compare Recombination Patterns on the Sex Chromosomes and Autosomes of the Guppy,
17 612 *Poecilia reticulata*, and Its Close Outgroup Species. *Mol. Biol. Evol.* [Internet].
- 18
19
20 613 Chepurnov VA, Mann DG, Sabbe K, Vyverman W. 2004. Experimental studies on sexual
21 614 reproduction in diatoms. *Int. Rev. Cytol.* 237:91–154.
- 22
23
24 615 Coelho SM, Gueno J, Lipinska AP, Cock JM, Umen JG. 2018. UV Chromosomes and Haploid
25 616 Sexual Systems. *Trends Plant Sci.* 23:794–807.
- 26
27 617 De Hoff PL, Ferris P, Olson BJSC, Miyagi A, Geng S, Umen JG. 2013. Species and population
28 618 level molecular profiling reveals cryptic recombination and emergent asymmetry in the
29 619 dimorphic mating locus of *C. reinhardtii*. *PLoS Genet.* 9:e1003724.
- 30
31
32 620 Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, Robbens S, Partensky F, Degroeve S,
33 621 Echeynie S, Cooke R, et al. 2006. Genome analysis of the smallest free-living eukaryote
34 622 *Ostreococcus tauri* unveils many unique features. *Proc Natl Acad Sci U A* 103:11647–
35 623 11652.
- 36
37
38 624 Devier B, Aguileta G, Hood ME, Giraud T. 2009. Ancient Trans-specific Polymorphism at
39 625 Pheromone Receptor Genes in Basidiomycetes. *Genetics* 181:209–223.
- 40
41
42 626 Ebenezer TE, Zoltner M, Burrell A, Nenarokova A, Novák Vanclová AMG, Prasad B, Soukal
43 627 P, Santana-Molina C, O'Neill E, Nankissoor NN, et al. 2019. Transcriptome, proteome and
44 628 draft genome of *Euglena gracilis*. *BMC Biol.* 17:11.
- 45
46
47 629 Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome
48 630 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157.
- 49
50
51 631 Ferris P, Olson BJSC, De Hoff PL, Douglass S, Casero D, Prochnik S, Geng S, Rai R,
52 632 Grimwood J, Schmutz J, et al. 2010. Evolution of an expanded sex-determining locus in
53 633 *Volvox*. *Science* 328:351–354.
- 54
55
56 634 Ferris PJ, Armbrust EV, Goodenough UW. 2002. Genetic structure of the mating-type locus of
57 635 *Chlamydomonas reinhardtii*. *Genetics* 160:181–200.
- 58
59
60

- 1
2
3 636 Ferris PJ, Goodenough UW. 1997. Mating type in *Chlamydomonas* is specified by mid, the
4 minus-dominance gene. *Genetics* 146:859–869.
5 637
6 638 Fontanillas E, Hood ME, Badouin H, Petit E, Barbe V, Gouzy J, de Vienne DM, Aguilera G,
7 Poulain J, Wincker P, et al. 2015. Degeneration of the nonrecombining regions in the mating-
8 639 type chromosomes of the anther-smut fungi. *Mol. Biol. Evol.* 32:928–943.
9 640
10 641 Fučíková K, Pažoutová M, Rindi F. 2015. Meiotic genes and sexual reproduction in the green
11 algal class Trebouxiophyceae (Chlorophyta). *J. Phycol.* 51:419–430.
12 642
13 643 Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian
14 genomes: the biased gene conversion hypothesis. *Genetics* 159:907–911.
15 644
16 645 Geng S, Miyagi A, Umen JG. 2018. Evolutionary divergence of the sex-determining gene MID
17 uncoupled from the transition to anisogamy in volvocine algae. *Dev. Camb. Engl.* 145.
18 646
19 647 Grimsley N, Péquin B, Bachy C, Moreau H, Piganeau G. 2010. Cryptic sex in the smallest
20 eukaryotic marine green alga. *Mol. Biol. Evol.* 27:47–54.
21 648
22 649 Gu Z, Gu L, Eils R, Schlesner M, Brors B. 2014. circlize Implements and enhances circular
23 visualization in R. *Bioinforma. Oxf. Engl.* 30:2811–2812.
24 650
25 651 Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles
26 D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-Seq:
27 reference generation and analysis with Trinity. *Nat. Protoc.* [Internet] 8. Available from:
28 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3875132/>
29 652
30 653 Hadjivasiliou Z, Pomiankowski A. 2016. Gamete signalling underlies the evolution of mating
31 types and their number. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 371.
32 654
33 655 Hamaji T, Kawai-Toyooka H, Uchimura H, Suzuki M, Noguchi H, Minakuchi Y, Toyoda A,
34 Fujiyama A, Miyagishima S, Umen JG, et al. 2018. Anisogamy evolved with a reduced sex-
35 determining region in volvocine green algae. *Commun. Biol.* 1:17.
36 656
37 657 Hamaji T, Mogi Y, Ferris PJ, Mori T, Miyagishima S, Kabeya Y, Nishimura Y, Toyoda A,
38 Noguchi H, Fujiyama A, et al. 2016. Sequence of the *Gonium pectorale* Mating Locus
39 Reveals a Complex and Dynamic History of Changes in Volvocine Algal Mating
40 Haplotypes. *G3 GenesGenomesGenetics* 6:1179–1189.
41 658
42 659 Hartmann FE, Duhamel M, Carpentier F, Hood ME, Foulongne-Oriol M, Silar P, Malagnac F,
43 Grognet P, Giraud T. 2021. Recombination suppression and evolutionary strata around
44 mating-type loci in fungi: documenting patterns and understanding evolutionary and
45 mechanistic causes. *New Phytol.* 229:2470–2491.
46 660
47 661
48 662
49 663
50 664
51 665
52 666
53 667
54
55
56
57
58
59
60

- 1
2
3 668 Hartmann FE, Rodríguez de la Vega RC, Gladieux P, Ma W-J, Hood ME, Giraud T. 2020.
4 669 Higher Gene Flow in Sex-Related Chromosomes than in Autosomes during Fungal
5 670 Divergence. *Mol. Biol. Evol.* 37:668–682.
- 6
7
8 671 Hill DRA, Wetherbee R. 1986. *Proteomonas sulcata* gen. et sp. nov. (Cryptophyceae), a
9 672 cryptomonad with two morphologically distinct and alternating forms. *Phycologia* 25:521–
10 673 543.
- 11
12
13 674 Hoekstra RF. 1987. The evolution of sexes. *Experientia. Suppl.* 55:59–91.
- 14
15 675 Hurst LD, Hamilton WD. 1992. Cytoplasmic fusion and the nature of sexes. *Proc. R. Soc. Lond.*
16 676 *B Biol. Sci.* 247:189–194.
- 17
18 677 Jain A, Perisa D, Fliedner F, von Haeseler A, Ebersberger I. 2019. The Evolutionary
19 678 Traceability of a Protein. *Genome Biol. Evol.* 11:531–545.
- 20
21
22 679 Jancek S, Gourbière S, Moreau H, Piganeau G. 2008. Clues about the genetic basis of adaptation
23 680 emerge from comparing the proteomes of two *Ostreococcus* ecotypes (Chlorophyta,
24 681 Prasinophyceae). *Mol. Biol. Evol.* 25:2293–2300.
- 25
26
27 682 Johnson LK, Alexander H, Brown CT. 2019. Re-assembly, quality evaluation, and annotation
28 683 of 678 microbial eukaryotic reference transcriptomes. *GigaScience* [Internet] 8. Available
29 684 from: <https://academic.oup.com/gigascience/article/8/4/giy158/5241890>
- 30
31
32 685 Joli N, Monier A, Logares R, Lovejoy C. 2017. Seasonal patterns in Arctic prasinophytes and
33 686 inferred ecology of *Bathycoccus* unveiled in an Arctic winter metagenome. *ISME J.*
34 687 11:1372–1385.
- 35
36
37 688 Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermin LS. 2017. ModelFinder:
38 689 fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14:587–589.
- 39
40
41 690 Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7:
42 691 Improvements in Performance and Usability. *Mol. Biol. Evol.* 30:772–780.
- 43
44
45 692 Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV,
46 693 Archibald JM, Bharti AK, Bell CJ, et al. 2014. The Marine Microbial Eukaryote
47 694 Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of
48 695 eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.* 12:e1001889.
- 49
50
51 696 Krasovec M, Eyre-Walker A, Sanchez-Ferandin S, Piganeau G. 2017. Spontaneous Mutation
52 697 Rate in the Smallest Photosynthetic Eukaryotes. *Mol. Biol. Evol.* 34:1770–1779.
- 53
54
55 698 Kües U. 2000. Life history and developmental processes in the basidiomycete *Coprinus*
56 699 *cinereus*. *Microbiol. Mol. Biol. Rev.* MMBR 64:316–353.
- 57
58 700 Kugrens P, Lee RE. 1988. Ultrastructure of Fertilization in a Cryptomonad1. *J. Phycol.* 24:385–
59 701 393.

- 1
2
3 702 Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A Resource for Timelines,
4 703 Timetrees, and Divergence Times. *Mol. Biol. Evol.* 34:1812–1819.
- 5
6 704 Lahn BT, Page DC. 1999. Four evolutionary strata on the human X chromosome. *Science*
7 705 286:964–967.
- 8
9
10 706 Lang D, Weiche B, Timmerhaus G, Richardt S, Riaño-Pachón DM, Corrêa LGG, Reski R,
11 707 Mueller-Roeber B, Rensing SA. 2010. Genome-wide phylogenetic comparative analysis of
12 708 plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with
13 709 complexity. *Genome Biol. Evol.* 2:488–503.
- 14
15
16
17 710 Leconte J, Benites LF, Vannier T, Wincker P, Piganeau G, Jaillon O. 2020. Genome Resolved
18 711 Biogeography of Mamiellales. *Genes* 11:66.
- 19
20 712 Lengeler KB, Fox DS, Fraser JA, Allen A, Forrester K, Dietrich FS, Heitman J. 2002. Mating-
21 713 Type Locus of *Cryptococcus neoformans*: a Step in the Evolution of Sex Chromosomes.
22 714 *Eukaryot. Cell* 1:704–718.
- 23
24
25 715 Li W-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous
26 716 substitution. *J. Mol. Evol.* 36:96–99.
- 27
28
29 717 Ma W-J, Carpentier F, Giraud T, Hood ME. 2020. Differential Gene Expression between
30 718 Fungal Mating Types Is Associated with Sequence Degeneration. *Genome Biol. Evol.*
31 719 12:243–258.
- 32
33
34 720 Malik S-B, Ramesh MA, Hulstrand AM, Logsdon JM. 2007. Protist homologs of the meiotic
35 721 Spo11 gene and topoisomerase VI reveal an evolutionary history of gene duplication and
36 722 lineage-specific loss. *Mol. Biol. Evol.* 24:2827–2841.
- 37
38
39 723 Matasci N, Hung L-H, Yan Z, Carpenter EJ, Wickett NJ, Mirarab S, Nguyen N, Warnow T,
40 724 Ayyampalayam S, Barker M, et al. 2014. Data access for the 1,000 Plants (1KP) project.
41 725 *GigaScience* 3:17.
- 42
43
44 726 McKim KS, Howell AM, Rose AM. 1988. The effects of translocations on recombination
45 727 frequency in *Caenorhabditis elegans*. *Genetics* 120:987–1001.
- 46
47
48 728 Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human
49 729 genome. *Mol Biol Evol* 21:984–990.
- 50
51
52 730 Natri HM, Merilä J, Shikano T. 2019. The evolution of sex determination associated with a
53 731 chromosomal inversion. *Nat. Commun.* 10:145.
- 54
55 732 Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective
56 733 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*
57 734 32:268–274.
- 58
59
60

- 1
2
3 735 Nozaki H, Mori T, Misumi O, Matsunaga S, Kuroiwa T. 2006. Males evolved from the
4 dominant isogametic mating type. *Curr. Biol.* 16:R1018–R1020.
5 736
6 737 Palenik B, Grimwood J, Aerts A, Rouze P, Salamov A, Putnam N, Dupont C, Jorgensen R,
7 Derelle E, Rombauts S, et al. 2007. The tiny eukaryote *Ostreococcus* provides genomic
8 738 insights into the paradox of plankton speciation. *Proc Natl Acad Sci U S A* 104:7705–7710.
9 739
10 740 Parfrey LW, Lahr DJG, Knoll AH, Katz LA. 2011. Estimating the timing of early eukaryotic
11 741 diversification with multigene molecular clocks. *Proc. Natl. Acad. Sci.* 108:13624–13629.
12 742
13 743 Pfiester LA. 1989. Dinoflagellate Sexuality. In: Bourne GH, Jeon KW, Friedlander M, editors.
14 International Review of Cytology. Vol. 114. Academic Press. p. 249–272.
15 744
16 745 Proost S, Bel MV, Sterck L, Billiau K, Parys TV, Peer YV de, Vandepoele K. 2009. PLAZA:
17 746 A Comparative Genomics Resource to Study Gene and Genome Evolution in Plants. *Plant*
18 *Cell* 21:3718–3731.
19 747
20 748 Richman A. 2000. Evolution of balanced genetic polymorphism. *Mol. Ecol.* 9:1953–1963.
21 749
22 750 Sager R, Granick S. 1954. Nutritional control of sexuality in *Chlamydomonas reinhardi*. *J. Gen.*
23 *Physiol.* 37:729–742.
24 751
25 752 Sanchez F, Geffroy S, Norest M, Yau S, Moreau H, Grimsley N. 2019. Simplified
26 753 Transformation of *Ostreococcus tauri* Using Polyethylene Glycol. *Genes* 10.
27 754
28 755 Slapeta J, Lopez-Garcia P, Moreira D. 2006. Global dispersal and ancient cryptic species in the
29 756 smallest marine eukaryotes. *Mol Biol Evol* 23:23–29.
30 757
31 758 Sonneborn TM. 1937. Sex, Sex Inheritance and Sex Determination in *Paramecium Aurelia*.
32 *Proc. Natl. Acad. Sci. U. S. A.* 23:378–385.
33 759
34 760 Soubrier J, Steel M, Lee MSY, Der Sarkissian C, Guindon S, Ho SYW, Cooper A. 2012. The
35 761 Influence of Rate Heterogeneity among Sites on the Time Dependence of Molecular Rates.
36 762 *Mol. Biol. Evol.* 29:3345–3358.
37 763
38 764 Speijer D, Lukes J, Elias M. 2015. Sex is a ubiquitous, ancient, and inherent attribute of
39 765 eukaryotic life. *Proc. Natl. Acad. Sci. U. S. A.* 112:8827–8834.
40 766
41 767 Staben C, Yanofsky C. 1990. *Neurospora crassa* a mating-type region. *Proc. Natl. Acad. Sci.*
42 768 87:4917–4921.
43 769
44 770 Strehlow, K. 1929. Über die Sexualität einiger Volvocales. *Zeitschrift für Botanik* 21:625:692.
45 771
46 772 Suda S, Watanabe MM, Inouye I. 1989. Evidence for Sexual Reproduction in the Primitive
47 773 Green Alga *Nephroselmis Olivacea* (prasinophyceae)1. *J. Phycol.* 25:596–600.
48 774
49 775 Sun Y, Corcoran P, Menkis A, Whittle CA, Andersson SGE, Johannesson H. 2012. Large-Scale
50 776 Introgression Shapes the Evolution of the Mating-Type Chromosomes of the Filamentous
51 777 Ascomycete *Neurospora tetrasperma*. *PLOS Genet.* 8:e1002820.
52 778
53
54
55
56
57
58
59
60

- 1
2
3 769 Tzeng Y-H, Pan R, Li W-H. 2004. Comparison of three methods for estimating rates of
4 770 synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 21:2290–2298.
5
6 771 Umen J, Coelho S. 2019. Algal Sex Determination and the Evolution of Anisogamy. *Annu.*
7 772 *Rev. Microbiol.* 73:267–291.
8
9 773 Umen JG. 2011. Evolution of sex and mating loci: an expanded view from Volvocine algae.
10 774 *Curr. Opin. Microbiol.* 14:634–641.
11
12 775 Vandepoele K, Bel MV, Richard G, Landeghem SV, Verhelst B, Moreau H, Peer YV de,
13 776 Grimsley N, Piganeau G. 2013. pico-PLAZA, a genome database of microbial
14 777 photosynthetic eukaryotes. *Environ. Microbiol.* 15:2147–2153.
15
16 778 Wang L, Berndt P, Xia X, Kahnt J, Kahmann R. 2011. A seven-WD40 protein related to human
17 779 RACK1 regulates mating and virulence in *Ustilago maydis*. *Mol. Microbiol.* 81:1484–1498.
18
19 780 Wickham H. 2011. ggplot2. *WIREs Comput. Stat.* 3:180–185.
20
21 781 Wilson AM, Wilken PM, van der Nest MA, Wingfield MJ, Wingfield BD. 2019. It's All in the
22 782 Genes: The Regulatory Pathways of Sexual Reproduction in Filamentous Ascomycetes.
23 783 *Genes* 10.
24
25 784 Wolfe KH, Butler G. 2017. Evolution of Mating in the Saccharomycotina. *Annu. Rev.*
26 785 *Microbiol.* 71:197–214.
27
28 786 Worden AZ, Lee JH, Mock T, Rouze P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML,
29 787 Derelle E, Everett MV, et al. 2009. Green evolution and dynamic adaptations revealed by
30 788 genomes of the marine picoeukaryotes *Micromonas*. *Science* 324:268–272.
31
32 789 Yamamoto K, Hamaji T, Kawai-Toyooka H, Matsuzaki R, Takahashi F, Nishimura Y, Kawachi
33 790 M, Noguchi H, Minakuchi Y, Umen JG, et al. 2021. Three genomes in the algal genus *Volvox*
34 791 reveal the fate of a haploid sex-determining region after a transition to homothallism. *Proc.*
35 792 *Natl. Acad. Sci.* 118:e2100712118.
36
37 793 Yang Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics*
38 794 139:993–1005.
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 801 **List of Figure Legends**
4

5
6 802 **Figure 1:** Size and GC content in the candidate mating type chromosomes (candidate mating
7
8 803 type locus positions as in Sup. Table S1) in the 8 Mamiellales genomes. Sequences of CH02 of
9
10 804 *O. lucimarinus* and *M. pusilla* have been reversed complemented to take colinearity of flanking
11
12 805 regions as described in (Palenik et al. 2007) and (Worden et al. 2009) into account. Node
13
14 806 divergence estimations are from (Slapeta et al. 2006) for *Micromonas* and (Parfrey et al. 2011)
15
16 807 for the basal node.
17

18
19 809 **Figure 2:** Gene organization in the mating type region of *O. tauri* RCC4221 (*MT*⁻) and
20
21 810 RCC1115 (*MT*⁺). (A) Location of gene pairs from the 23 core gene families (GFs) and 57
22
23 811 shared GFs in the mating type region of *O. tauri* RCC4221 (*MT*⁻, blue rectangle) and RCC1115
24
25 812 (*MT*⁺, red rectangle). Genes from core and shared GFs are represented by bright and dark ticks,
26
27 813 respectively, and homologous gene pairs are connected by grey lines. Genes from *MT*⁺ or *MT*⁻
28
29 814 specific GFs are also shown, represented by black ticks. Shared gene families having multiple
30
31 815 copies in either *O. tauri* RCC4221 (*MT*⁻) and RCC1115 (*MT*⁺) are not depicted. (B, C)
32
33 816 Location of homologous *MT* genes from GFs with copies outside of either *O. tauri* *MT* region
34
35 817 in Mamiellales genomes, for *O. tauri* RCC4221 (*MT*⁻, 39 GFs, B) and RCC1115 (*MT*⁺, 16 GFs,
36
37 818 C). Each peripheral segment represents a chromosome or scaffold of one of eight Mamiellales
38
39 819 genomes. The *MT* genes from *O. tauri* RCC4221 (*MT*⁻, B) or RCC1115 (*MT*⁺, C) are
40
41 820 connected to their homologs by grey lines. If a homolog is located within a *MT* locus, the link
42
43 821 is coloured in orange. The abbreviations are as follows: chromosome (CH), contig (CG), and
44
45 822 unitig (UG).

46
47 823 **Figure 3:** Phylogenetic signal and GC3 content of gene family (GF) members in *O. tauri*
48
49 824 RCC4221 (*MT*⁻) and RCC1115 (*MT*⁺). ‘Post’ for GF genes with mating type separation after
50
51 825 speciation and ‘Ante’ for GF genes with mating type separation prior to speciation. Circle size
52
53 826 is proportional to the number of GF genes (numerical value within each circle), and circle colour
54
55 827 depicts the average GC3 content from low (yellow-golden) to high (green).
56

57
58 829 **Figure 4:** Unrooted maximum-likelihood phylogenetic trees of representative core *MT* gene
59
60 830 families 000581 (A) and 000945 (B). Genes from *MT*⁻ strains are coloured in blue, genes from
831
832 *MT*⁺ strains are coloured in red. Ultrafast bootstrap support values are denoted on branches.
Abbreviations: *O. tauri* RCC4221 (OT4221), *O. tauri* RCC1115 (OT1115), *O. lucimarinus*

833 (OL), *O. sp* RCC809 (O809), *O. mediterraneus* RCC2590 (OMED), *B. prasinus* RCC1105
 834 (B1105), *M. commoda* RCC299 (MC299), and *M. pusilla* (MPU).

835

836 **Figure 5:** Presence-absence matrix of best BLAST hits (BBH) of core *MT* gene families in each
 837 Mamiellophyceae transcriptome. Species' names of sequenced strains (left column) as inferred
 838 from 18S rDNA sequence analysis extracted from the transcriptome (supplementary fig. S2,
 839 Supplementary Material online). The colour of each rectangle indicates the taxonomic
 840 affiliation of the BBH (colour key at the bottom). Transcriptomes containing genes with a BBH
 841 affiliated to a different species are highlighted in grey.

842

843 **Figure 6:** Unrooted maximum-likelihood phylogenetic trees of representative core *MT* gene
 844 families 001374 (A) and 003390 (B) including homologous sequences from *O. lucimarinus*
 845 MMETSP0939 (strain BCC118000) and *O. mediterraneus* MMETSP0929 (strain RCC2572).
 846 Candidate mating type genes *MT*⁺ are in red, *MT*⁻ in blue. Topology (A) clusters genes
 847 according to mating type, whereas topology (B) corresponds to the species phylogeny. Ultrafast
 848 bootstrap support values are indicated on branches. Abbreviations: *O. tauri* RCC4221
 849 (OT4221), *O. tauri* RCC1115 (OT1115), *O. lucimarinus* (OL), *O. lucimarinus* BCC118000
 850 (OLMMETSP0939), *O. sp* RCC809 (O809), *O. mediterraneus* RCC2590 (OMED), *O.*
 851 *mediterraneus* RCC2572 (OMMMETSP0929), *B. prasinus* RCC1105 (B1105), *M. commoda*
 852 RCC299 (MC299), and *M. pusilla* (MPU).

853

854 **Figure 7:** Phylogenetic trees of representative core *MT* gene families 001102 (A) and 003908
 855 (C), and actin (B) and β -tubulin (D) genes, for *M. commoda* and *M. pusilla* reference genomes
 856 and homologous genes retrieved from diverse *Micromonas* spp. transcriptomes. In the
 857 phylogeny "A", two *M. commoda* sub-clusters are highlighted in dark green (MMETSP1403,
 858 1400, 1084, 1387) and light green (MMETSP1393, 1404, and the reference *M. commoda*
 859 RCC299). In phylogeny "C", there is no "A type" sub-clustering. In the actin and β -tubulin
 860 trees, sub-clusters are absent and branch lengths are shorter than "A", but parallel with
 861 phylogeny "C".

862

863 **Figure 8:** GC3 content comparison between genes from core *MT* gene families (dark green)
 864 and overall genome or transcriptome (light green) in Mamiellophyceae species. Phylogenetic
 865 relationships are inferred from 18S rDNA phylogeny. An asterisk (*) indicates a significant
 866 GC3 content difference (Student's *t*-test *p*-value < 0.05). Abbreviations: *O. tauri* RCC4221

1
2
3 867 (*Ostreococcus*), *B. prasinus* RCC1105 (*Bathycoccus*), *M. pusilla* CCMP1545 (*Micromonas*),
4
5 868 *M. squamata* strain CCMP1436 - MMETSP1468 (*Mantoniella squamata*), *M. squamata* strain
6
7 869 CCCAP 1965/1 - QXSZ (*M. squamata*), *M. antarctica* strain SL-175 - MMETSP1106
8
9 870 (*Mantoniella antarctica*), Uncultured eukaryote RCC2288 - MMETSP1326 (Uncult.
10
11 871 Mamiellophyceae), *C. stigmatica* CCMP3273 - MMETSP0803 (*Crustomastix*), *D. tenuilepis*
12
13 872 CCMP3274 - MMETSP0033 (*Dolichomastix*), *D. tenuilepis* strain M1680 - XOAL (*D.*
14
15 873 *tenuilepis*), and *M. opisthostigma* CCAC 0206 - BTFM (*Monomastix*).
16

17 875 **List of Tables**

18
19 876 **Table 1:** Classification, description, and quantities of genes and gene families (GFs) in *O. tauri*
20
21 877 RCC4221 (*MT*-) and RCC1115 (*MT*+) strains.
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

879 **Table 1:** Classification, description, and quantities of genes and gene families (GFs) in *O. tauri* RCC4221 (*MT*-)
 880 and RCC1115 (*MT*+) strains.

Gene Family class	Features of included genes	RCC4221 (<i>MT</i> -)	RCC1115 (<i>MT</i> +)
<i>MT</i> specific GFs	Present in either all <i>Ostreococcus MT</i> - or all <i>Ostreococcus MT</i> +	6 genes in 6 GFs	2 genes in 2 GFs
Core <i>MT</i> GFs	Present in all Mamiellales genomes and located only in <i>MT</i> region	23 genes in 23 GFs	23 genes in 23 GFs
Shared <i>MT</i> GFs (non-core)	Present in both <i>Ostreococcus MT</i> loci, but not in all Mamiellales <i>MT</i> regions	75 genes in 69 GFs	79 genes in 69 GFs
GFs extending outside <i>MT</i>	Present in one <i>Ostreococcus MT</i> locus but with homologous genes in other regions in the opposite strain	28 genes in 27 GFs	8 genes in 4 GFs
GFs not retained for analysis	Present in only one <i>Ostreococcus MT</i> locus and Mamiellales genomes but absent from the genomes of the opposite strains/ <i>MT</i> ; divergent GFs or singletons	112 genes	128 genes
Total number of genes		244	240

881

882

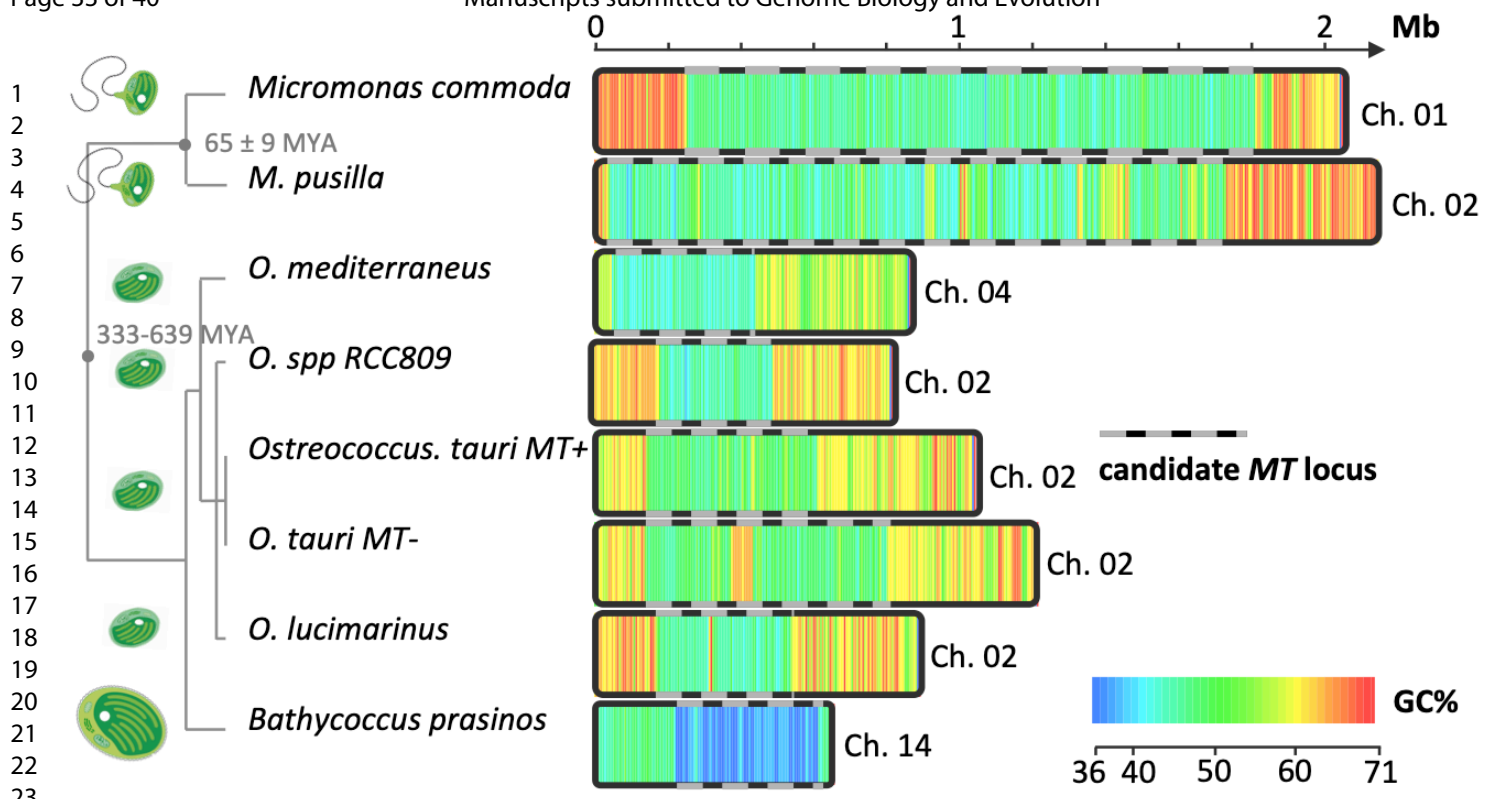
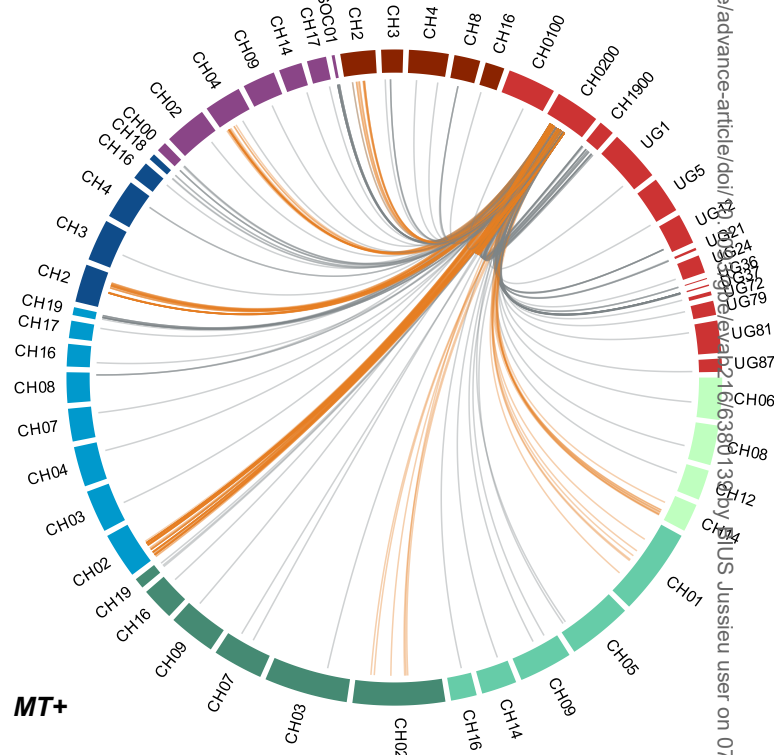
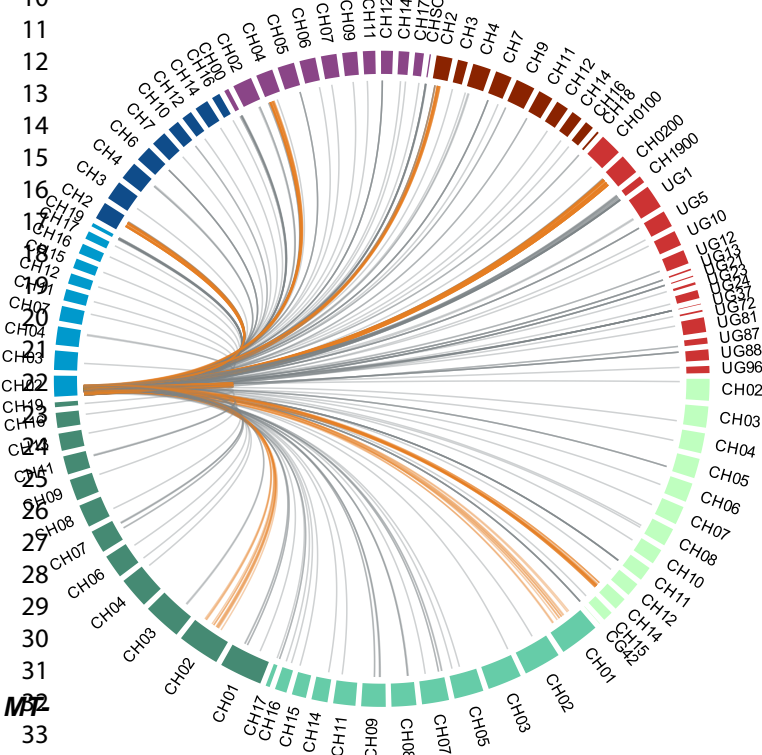
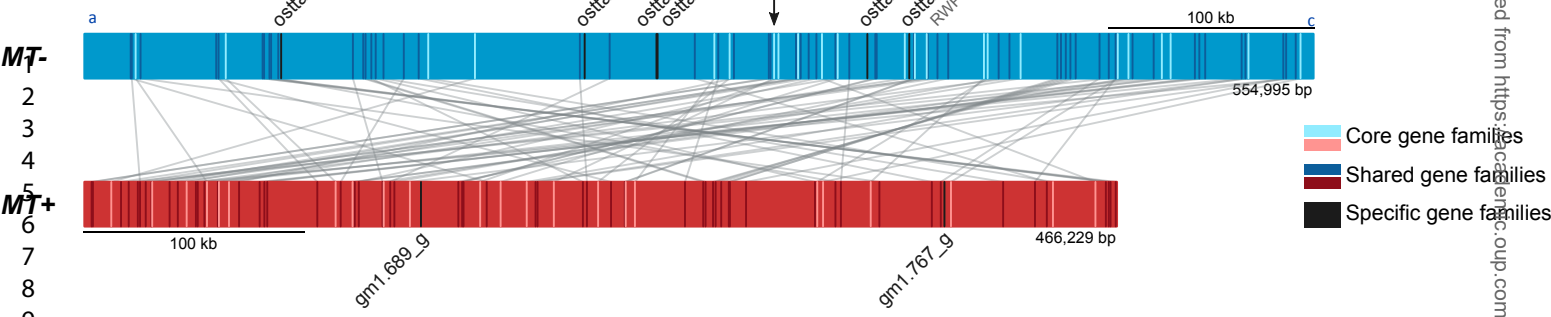


Figure 1. "Big Outlier Chromosome" size and GC content in eight completely sequences Mamiellales strains.

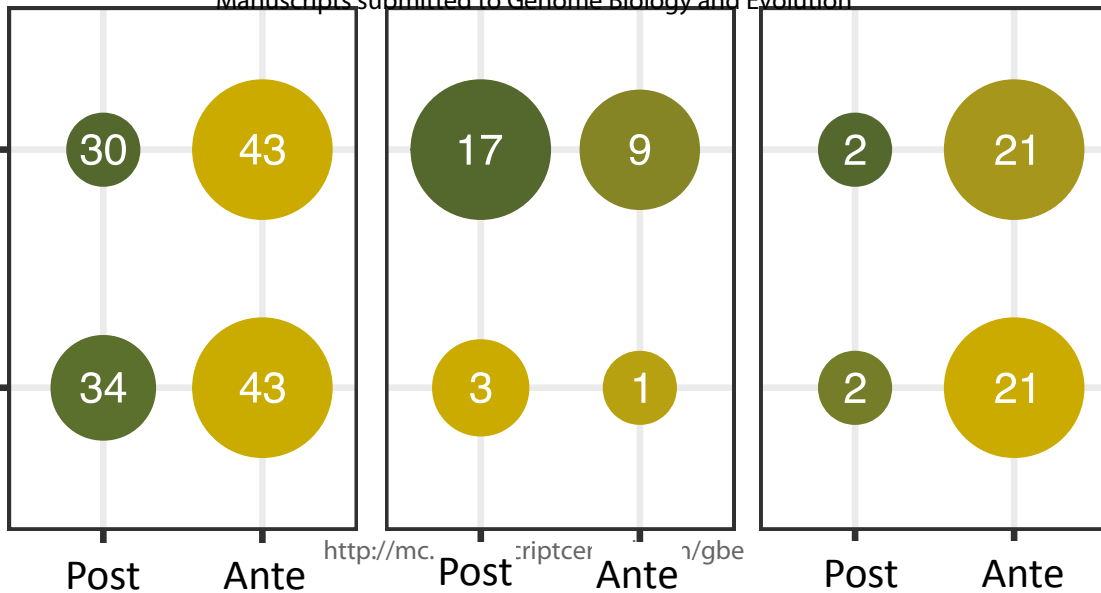


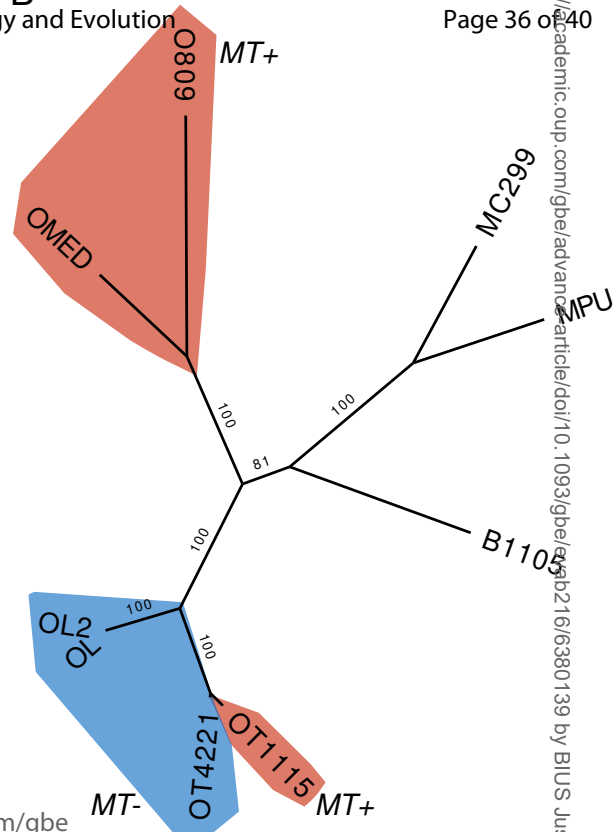
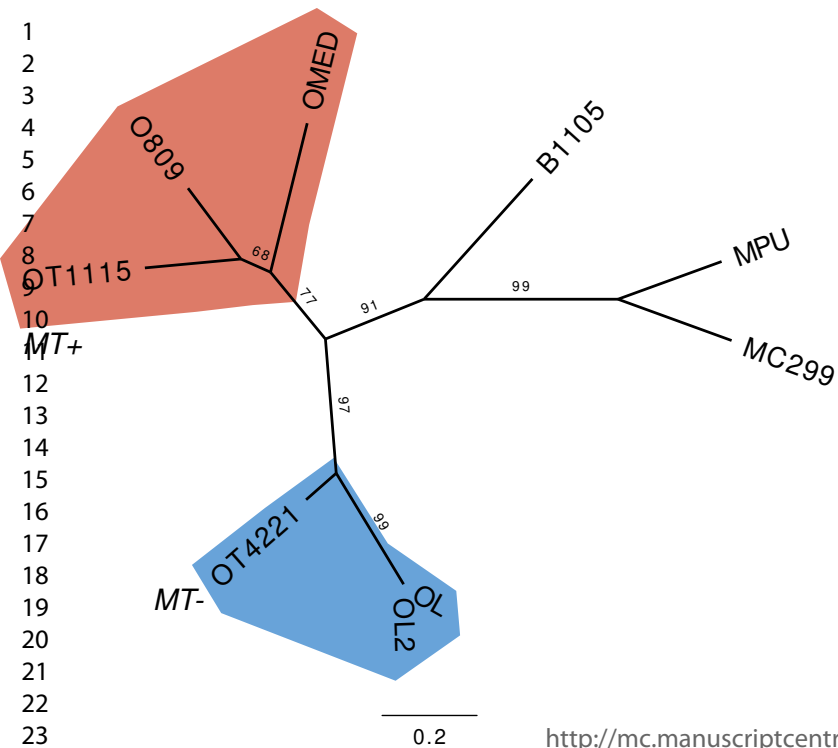
Shared

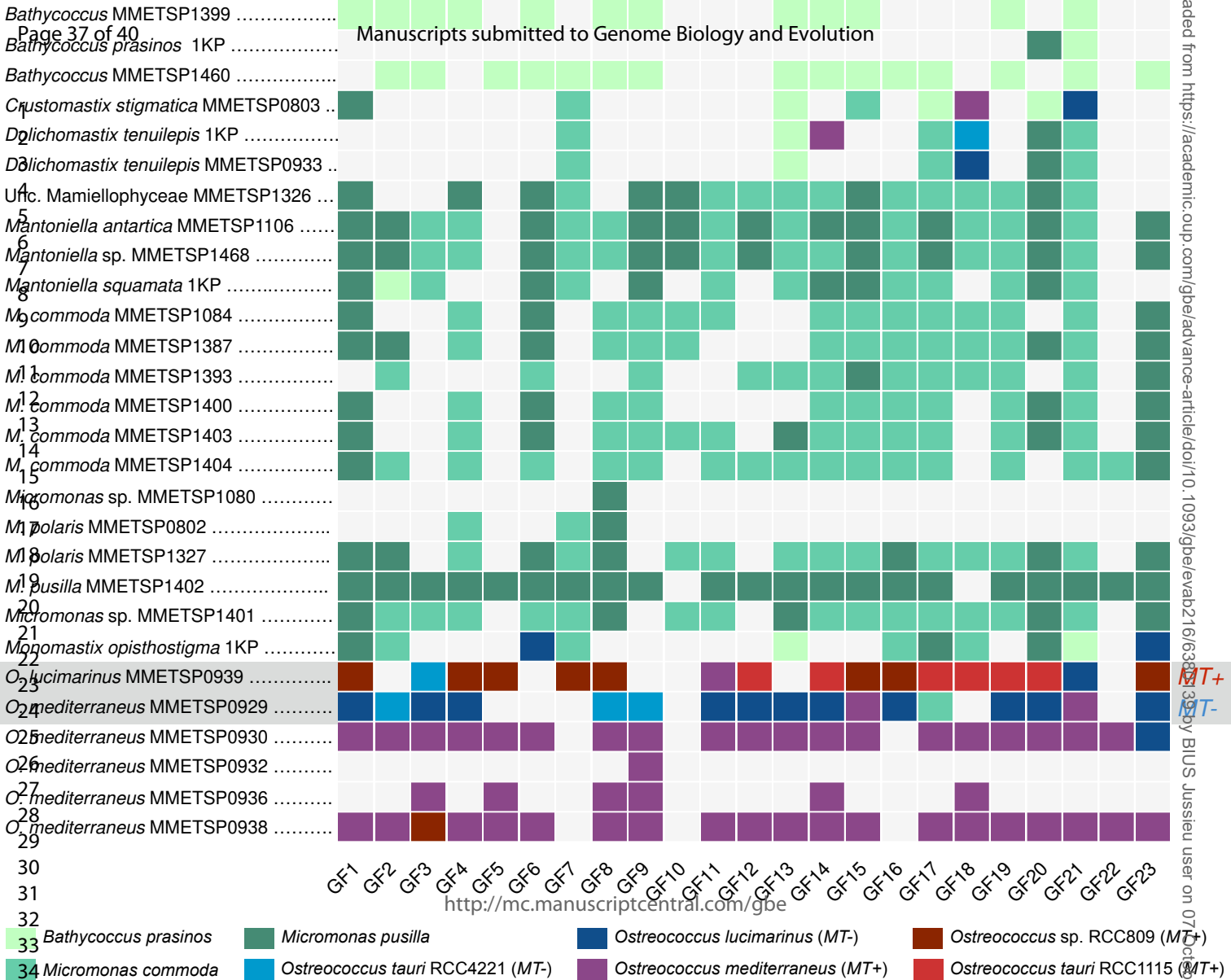
Extending outside

Core

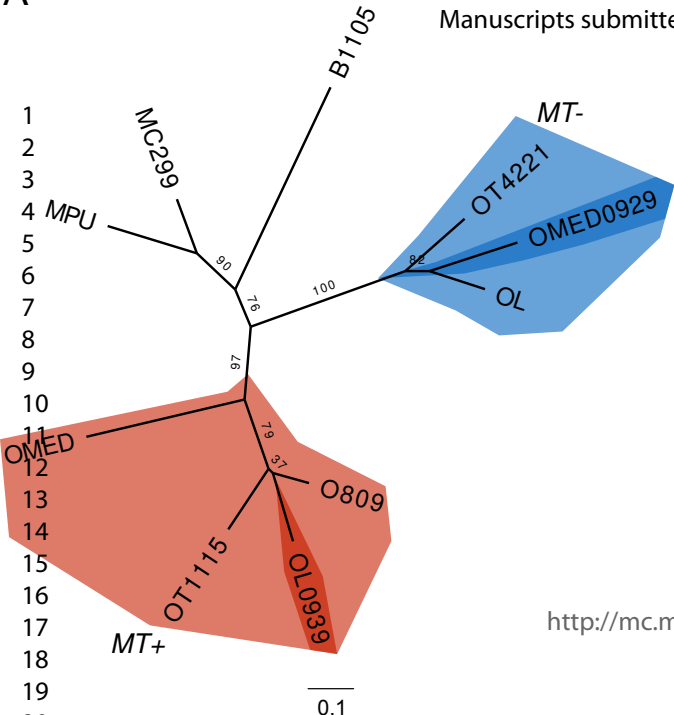
Manuscripts submitted to Genome Biology and Evolution

1
2
3 *O. tauri* MT-4
5
6
7
8
9
10 *O. tauri* MT+





A



B

Manuscripts submitted to Genome Biology and Evolution

