Extreme genome diversity and cryptic speciation in a harmful algal bloom forming eukaryote

Jennifer H. Wisecaver[1,2], Robert P. Auber[1,2]*, Amanda L. Pendleton[1,2]*, Nathan F. Watervoort[1,2], Timothy R. Fallon[3], Olivia L. Riedling[1,2], Bradley S. Moore[3,4], and William W. Driscoll[5]

[1]Department of Biochemistry, Purdue University, West Lafayette, Indiana, USA
[2]Purdue Center for Plant Biology, Purdue University, West Lafayette, Indiana, USA
[3]Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California, San Diego, CA, USA
[4]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, CA, USA
[5]Department of Biology, Penn State Harrisburg, Middletown, PA, USA
* These authors contributed equally

Correspondence: Jennifer Wisecaver, 175 S University St., West Lafayette, IN 47907;
e: jwisecav@purdue.edu; p: 765-494-4157

ORCID IDS: 0000-0001-6843-5906 (J.H.W.), 0000-0002-1442-7473 (R.P.A.), 0000-0002-7201-0596 (A.L.P.), 0000-0002-5227-4818 (N.F.W.), 0000-0002-3048-7679 (T.R.F.), 0000-0003-2735-0513 (O.L.R.), 0000-0002-4652-1253 (B.S.M.)

ABSTRACT

Harmful algal blooms (HABs) of the toxic haptophyte *Prymnesium parvum* are a recurrent problem in many inland and estuarine waters around the world. Strains of *P. parvum* vary in the toxins they produce and in other physiological traits associated with HABs, but the genetic basis for this variation is unknown. To investigate genome diversity in this morphospecies, we generated genome assemblies for fifteen phylogenetically and geographically diverse strains of *P. parvum* including Hi-C guided, near-chromosome level assemblies for two strains. Comparative analysis revealed considerable DNA content variation between strains, ranging from 115 Mbp to 845 Mbp. Strains included haploids, diploids, and polyploids, but not all differences in DNA content were due to variation in genome copy number. Haploid genome size between strains of different chemotypes differed by as much as 243 Mbp. Syntenic and phylogenetic analyses indicate that UTEX 2797, a common laboratory strain from Texas, is a hybrid that retains two phylogenetically distinct haplotypes. Investigation of gene families variably present across strains identified several functional categories associated with metabolism, including candidates for the biosynthesis of toxic metabolites, as well as genome size variation, including recent proliferations of transposable elements. Together, our results indicate that *P. parvum* is comprised of multiple cryptic species. These genomes provide a robust phylogenetic and genomic framework for investigations into the eco-physiological consequences of the intra- and inter-specific genetic variation present in *P. parvum* and demonstrate the need for similar resources for other HAB-forming morphospecies.

22 **SIGNIFICANCE STATEMENT**

23 Harmful algal blooms (HABs) are a global concern. Efforts to understand the genetic basis of
24 traits associated with the success of HAB-forming species are limited by a dearth of genomic
25 resources. In this paper we present genomes for fifteen strains of *Prymnesium parvum*, a toxic
26 alga that causes ecosystem and societally disruptive HABs around the world. We uncover an
27 unprecedented amount of sequence-level, gene family, and genome architecture evolution in *P.*
28 *parvum* and provide evidence for both cryptic speciation and hybridization. These results
29 illustrate how both inter- and intraspecific genetic variation can be dramatically underestimated
30 in a protist morphospecies. More work is needed to understand the eco-physiological
31 consequences of hidden genetic diversity in *P. parvum* and HAB-forming species more
32 generally.
33

INTRODUCTION

Environmentally disruptive harmful algal blooms (HABs) are increasing in frequency and severity in many marine, brackish, and freshwater ecosystems (1–3). Knowledge of genetic variation within HAB populations can aid efforts to predict when HABs will occur and model HAB impacts in a changing climate. However, genome-level data are currently limited for HAB-forming species (4). One such species, *Prymnesium parvum,* is a unicellular microalga belonging to the haptophyte clade of eukaryotes (5). Globally, HABs of *P. parvum* have considerable ecologic and economic impacts. In the southwestern United States, particularly in Texas, blooms are correlated with several troubling environmental trends including increased use of glyphosate herbicides, increased concentrations of atmospheric CO_2, and decreased wetland cover (6, 7).

P. parvum's success as a HAB-forming species likely results from several interrelated aspects of its biology. The species is extremely eurythermal and euryhaline, tolerating temperatures ranging from 2 to 32 °C and salinities ranging from 0.5 to 125% of typical seawater (8). Although *P. parvum* is photosynthetic and maintains permanent chloroplasts, it is also a mixotroph and able to ingest organic material via phagocytosis (9, 10). Despite its small size, *P. parvum* is a ferocious facultative predator capable of swarming and killing various prey including bacteria, other microbial eukaryotes, aquatic invertebrates, and even fish (11–14). The species produces toxins that are thought to assist in this predatory behavior and/or inhibit the activity of competitors and grazers (5). Heterotrophy and toxicity can increase in *P. parvum* cultures grown in sub-optimal conditions, *e.g.*, when nutrients are limiting or water temperature/salinity is at the edge of what a strain can tolerate (8, 10, 15).

P. parvum produces a variety of toxic metabolites (16). Most conspicuous are the prymnesins, massive ladder-frame polyether compounds that are unique to *P. parvum* but structurally similar to the phycotoxins of many HAB-forming dinoflagellates (17). More than 50 different prymnesins have been identified and are grouped into types based on the number of carbons in their aglycone backbone. The backbone of A-type prymnesins contain 91 carbons, while B-type and C-type prymnesins contain 85 and 83 carbons, respectively (18, 19). To date, all strains of *P. parvum* produce just one type of prymnesin, and strains that produce the same prymnesin type (*i.e.*, chemotype) group together in phylogenies of rDNA sequences (19, 20).

65 The strong agreement between phylogeny and chemotype —despite strains of different
66 chemotypes often existing in the same geographical location—suggests that there is limited gene
67 flow between chemotypes of *P. parvum* (19).
68
69 Strains of *P. parvum* may appear morphologically indistinguishable, but the species
70 harbors considerable phenotypic diversity. In addition to the previously discussed diversity in
71 prymnesin chemotypes, the amount of prymnesin produced varies between strains (20) as does
72 overall toxicity (20–23). Predatory behavior also varies between *P. parvum* strains, even between
73 strains taken from the same bloom (24). Moreover, several abiotic factors including temperature,
74 salinity, pH, and irradiance have distinct effects on the relative growth rate and toxicity of
75 different strains (23, 25–27). This demonstrates remarkable genotype by environment
76 interactions and adaptive plasticity in *P. parvum*.
77
78 Extensive physiological and biochemical variation between strains of *P. parvum* indicates
79 either the existence of multiple cryptic species (19) or extreme standing genetic variation in this
80 morphospecies. However, the genetic differences between strains have not yet been quantified.
81 Here, we generated fifteen sequenced genomes of *P. parvum* strains and report considerable
82 genomic variation, both at the nucleotide level and in terms of gene family presence/absence.
83 Our comparative phylogenetic analysis supports monophyletic origins of A-, B-, and C-type
84 prymnesins, with C-type strains having more within-clade sequence diversity compared to A-
85 and B-types. Moreover, we found that the three prymnesin chemotypes of *P. parvum* have
86 dramatically different haploid genome sizes. In addition to haploid strains identified in all three
87 chemotypes, we also identified A- and C-type diploid strains, as well as A- and B-type 4n
88 tetraploids. Lastly, we present evidence that UTEX 2797, a common laboratory strain from
89 Texas, is a hybrid that retains two phylogenetically distinct haplotypes.
90
91 **RESULTS**
92 ***Prymnesium parvum* genome assemblies.** We obtained Hi-C scaffolded, highly contiguous
93 genome assemblies of two *P. parvum* strains from Texas: UTEX 2797 and 12B1. UTEX 2797
94 was selected for sequencing due to its use in numerous toxicology and physiology experiments
95 of *P. parvum* (*e.g.,* 19, 20, 22, 27–32). However, preliminary analysis of k-mer frequencies

96 revealed that UTEX 2797 displays high sequence-level heterozygosity (Fig. 1A), which can
97 complicate genome assembly. Therefore, we also selected strain 12B1 (21), which had no
98 observable heterozygosity to obtain a homozygous reference assembly (Fig. 1A). UTEX 2797
99 produces A-type prymnesins (18, 19), and our chemotyping analysis revealed strain 12B1
100 produces A-type prymnesins as well (Fig. S1).
101
102 The nuclear DNA content of both strains was estimated using flow cytometry (Fig. 1B).
103 The majority of 12B1 nuclei contained an average of 0.12 pg DNA, which corresponds to
104 approximately 115.6 million base pairs (Mbp), assuming a conversion factor of 1 pg = 0.98×10^9
105 base pairs (33). The 12B1 flow cytometry histogram also contained a second, substantially
106 smaller peak corresponding to approximately 0.23 pg DNA (228 Mbp), indicating that some
107 cells may have been in the process of dividing (*i.e.*, the G_2 phase of the cell cycle) (Fig. 1B). The
108 UTEX 2797 flow cytometry histogram contained a single dominant peak corresponding to 0.28
109 pg DNA, or 274.4 Mbp (Fig. 1B, Table S1). A cryptic sexual lifecycle for *P. parvum* has been
110 proposed that alternates between haploid and diploid forms, which are only distinguishable by
111 transmission electron microscopy of the cells' outer scales (34). However, neither syngamy nor
112 meiosis have been observed in *P. parvum*, and its lifecycle remains unknown. Therefore, we
113 must be cautious when inferring the haploid genome size and relative ploidy states of different
114 strains, and we tentatively labeled the dominant peaks in 12B1 and UTEX 2797 as 1C and 2C,
115 respectively (Fig. 1B).
116
117 The Hi-C scaffolded genome assembly for 12B1 consisted of 34 scaffolds spanning 93.6
118 Mbp with an N50 of 3.2 Mbp (Table S2). The UTEX 2797 assembly was over twice the length
119 of the 12B1 assembly at 197.6 Mbp and consisted of 66 scaffolds with an N50 of 3.4 Mbp (Table
120 S2). Compared to nine other haptophyte assemblies (see Table S3 and references therein), the
121 UTEX 2797 and 12B1 assemblies are the second and third most contiguous (Fig. 1C).
122 Chromosomal end-to-end assembly was accomplished for 29.4% and 7.5% of 12B1 and UTEX
123 2797 scaffolds, respectively (Table S4), evidenced by the presence of TTAGGG telomeric
124 sequences observed in other haptophytes (35). Overall, most 12B1 scaffolds (73.5%) had at least
125 one predicted telomere. In contrast, only 42.4% of UTEX 2797 scaffolds had any predicted
126 telomeres (Fig. S2), which suggests that the level of heterozygosity in the UTEX 2797 genome

127 may have impacted the complete assembly of these repetitive regions. Total repetitive sequence
128 makes up 29.4% and 35.5% of the 12B1 and UTEX 2797 genome assemblies, respectively
129 (Table S5), which are within the range of values (22.9% - 64%) reported for other haptophytes
130 (36–38).
131
132 We predicted and annotated 23,802 and 45,535 protein-coding genes in the genomes of
133 12B1 and UTEX 2797, respectively (Table S6). In total, 86.2% and 86.4% of the 12B1 and
134 UTEX 2797 genes could be assigned predicted functional annotations through InterProScan (39).
135 The completeness of the predicted gene sets were assessed with BUSCO (40); 216 of 255
136 conserved eukaryotic genes (84.7%) were complete within both predicted proteomes (Table S7).
137 This level of BUSCO recovery is the greatest of any currently available haptophyte genome
138 assembly (Fig. 1C, Table S3). While only 5.1% of BUSCOs were duplicated in 12B1, the
139 majority of BUSCOs (75.5%) were duplicated in UTEX 2797 (Table S7).
140
141 Synteny and collinearity analyses further highlight the duplicated state of the UTEX 2797
142 genome assembly. A strong 2:1 synteny pattern is observed between the genes of UTEX 2797
143 and 12B1; 93% of UTEX 2797 genes are syntenic to a single region in 12B1, whereas 84% of
144 12B1 genes are syntenic to two regions in the UTEX 2797 genome (Fig. 1D). Most 12B1
145 scaffolds are collinear with two syntenic regions in the UTEX 2797 assembly, but an abundance
146 of large structural variants (*e.g.*, inversions, indels, translocations) are apparent both between the
147 two reference strains as well as between syntenic scaffolds of UTEX 2797 (Fig. 1E). Together,
148 these data indicate that the assembly of UTEX 2797 consists of two largely non-collapsed
149 haplotypes, which is consistent with the strain being a highly heterozygous diploid. Due to the
150 extreme heterozygosity present in UTEX 2797, we suspected that the strain may have arisen via
151 hybridization of two divergent parents. However, further investigation required data from
152 additional *P. parvum* strains.
153
154 To enable phylogenetic analysis of *P. parvum* and investigate the origin of the divergent
155 haplotypes in UTEX 2797, we selected thirteen additional *P. parvum* strains for genomic
156 characterization that represent different geographical locations and prymnesin chemotypes
157 (Table S8). In addition to 12B1 and UTEX 2797, we included four other strains that produce A-

158 type prymnesins: 12A1, CCMP2941, CCMP3037, and RCC3703 (Fig. S1) (19). Three strains in
159 the analysis produce B-type prymnesins (K-0081, K-0374, and KAC-39), and four produce C-
160 type prymnesins (K-0252, RCC191, RCC1433, and RCC1436) (18, 19). Two additional strains
161 in our analysis (RCC3426 and UTEX 995) have not yet been chemotyped. We assembled and
162 annotated the genomes of these thirteen additional strains using Illumina short reads (Table S2).
163 Genome assembly lengths ranged from 77.0 Mbp (CCMP3707) to 92.9 Mbp (RCC1436). As
164 expected of short-read assemblies, genome contiguity was low, with contig N50s that ranged
165 from 3414 bp (K-0081) to 8872 bp (RCC1433). Nevertheless, coding gene space was well
166 represented, and the number of protein coding genes ranged from 26,458 (CCMP3037) to 32,339
167 (UTEX 995) (Table S6). BUSCO scores ranged from 65% (UTEX 995) to 81% (RCC1433),
168 comparable to BUSCO scores observed in other Illumina-only haptophyte genome assemblies
169 (Table S7). However, it is notable that 10 unique BUSCOs were not recovered in any of the 24
170 haptophyte genomes investigated (Table S9). This suggests that some BUSCOs may be absent in
171 haptophytes or too divergent to be recovered using the current BUSCO sequence models.
172
173 **High sequence divergence within a unicellular morphospecies.** To determine the relatedness
174 among strains of different chemotypes, we first identified orthologous gene families (*i.e.*,
175 orthogroups) using OrthoFinder (41). We selected 2699 single-copy orthogroups (SCOGs)
176 present in all 15 strains for concatenation- and coalescent-based species tree analyses. Because
177 the most closely related haptophyte species with sequenced genomes were too divergent to serve
178 as outgroups for a nucleotide-based phylogeny, a nonreversible substitution model was used to
179 evaluate support for different root locations on the *P. parvum* species tree (42). The root with the
180 highest likelihood score caused the C-type prymnesin producers to be paraphyletic, which is a
181 phylogenetic pattern also observed in a prior 18S rDNA phylogeny (20). However, 'rootstrap'
182 support for this root location was low (68%), and two alternative root locations, including one in
183 which C-type producers were monophyletic, could not be rejected (AU tree topology test P-
184 values > 0.1; Fig. 2A, Fig. S3).
185
186 Regardless of the root location, both concatenation- and coalescent-based species trees
187 supported the monophyletic origin of A-type and B-type prymnesins (Fig. 2A, Fig. S8). Support
188 for the species tree topology was assessed using gene and site concordance factors (gCF/sCF)

189 (43). Support was generally high across the tree but low within A-type and B-type clades (Fig.
190 2A, Fig. S8). These low support values also correspond to areas of disagreement between the
191 concatenation and coalescent based phylogenies (Fig. S8). Untyped strain RCC3426 grouped
192 with strains that produce B-type prymnesins with strong support (gCF = 95.1; sCF = 96.7),
193 suggesting RCC3426 also produces B-type prymnesins. Similarly, untyped strain UTEX 995
194 grouped with C-type strain RCC1436 (gCF = 92.7; sCF = 94.9), indicating it likely produces C-
195 type prymnesins.
196
197 Divergence among strains was evaluated using two approaches. First, divergence was
198 estimated based on the average number of substitutions per synonymous site (*Ks*) between gene
199 pairs in 15,074 orthogroups. These orthogroups were selected based on their presence in 10 or
200 more strains and robust nucleotide alignments. Average *Ks* within A- and B-type clades was
201 extremely low at 0.003 and 0.000, respectively (Fig. 2A, Fig. S4, Table S10). The C-type strains
202 formed three distinct groups based on *Ks*: clade C1 (RCC191 and RCC1433), clade C2
203 (RCC1436, UTEX 995), and clade C3 (K-0252). Median *Ks* was elevated when C-type strains
204 were compared to each other (*Ks* = 0.0241) with the C3 strain, K-0252, from Australia acting as
205 a significant outlier when compared to other C-types (*Ks* = 0.048) (Fig. 2A, Fig. S4, Table S10).
206 The largest *Ks* values occurred when C-type strains were compared to A- and B-types (*Ks* =
207 0.093). These results suggest that C-type strains form a monophyletic group and that the best
208 root location for the *P. parvum* species tree is along the branch separating C-type strains from A-
209 and B-types (Fig. 2A).
210
211 A second approach, complimentary to sequence divergence and based on breadth of
212 coverage (BOC) of each strain's Illumina reads, was used to explore strain relatedness. Reads
213 were aligned to the 12B1 and UTEX 2797 reference genomes, and BOC was defined as the
214 percentage of reference sequence spanned by greater than ten reads. Average BOC of the haploid
215 12B1 genome varied dramatically between chemotypes, ranging from 91.7% in A-types, 72.7%
216 in B-types, and 63.4% in C-type strains (Fig. 2B, Table S11). A-type strains from the U.S.A had
217 a higher average BOC (95.4%) compared to the two European strains (84.2%). BOC data from
218 all strains plotted across the 34 scaffolds of the 12B1 reference assembly generally depicted
219 decreased coverage as phylogenetic distance increased but also identified several regions of the

220 12B1 assembly that were strain- or subclade-specific (Fig. S5). All strains maintained high
221 sequence coverage across 12B1's genic space, ranging from 97.5% in other A-type strains to
222 91.2% in strain UTEX 995 (Fig. 2B, Table S11). This is pattern is similar to that seen in
223 interspecific comparisons of *Arabidopsis* spp. (Streptophyta), in which protein-coding genes are
224 conserved while intergenic sequences show significant divergence (44).
225
226 **Phylogenetically divergent haplotypes within a single strain.** Average breadth of coverage
227 was strikingly low for most strains when aligned to the UTEX 2797 assembly (Fig. 2B, Table
228 S11). The two exceptions were UTEX 2797 (*i.e.*, aligned to itself, BOC = 86%) and another
229 strain from Texas, 12A1 (BOC = 91.7%). This pattern of low coverage across the UTEX 2797
230 assembly is consistent with UTEX 2797 having two divergent haplotypes. Strain 12A1 likely
231 shares these or closely related haplotypes as its reads map well to both haplotypes while reads
232 from other strains primarily map to one haplotype or the other (Fig. 2B, Table S11). To
233 investigate the evolutionary history of the UTEX 2797 haplotypes, we first plotted BOC across
234 the 66 scaffolds of the UTEX 2797 reference genome. Clear haplotype blocks were identified
235 that showed either high coverage in A-type strains from the U.S.A. or high coverage in the two
236 A-type strains from Europe: CCMP2941 from Russia and RCC3703 from the United Kingdom
237 (Fig. S6). Similarly, manual inspection of individual gene trees revealed that UTEX 2797
238 frequently grouped with these two European strains (*e.g.*, Fig. S7) despite grouping with other
239 strains from the U.S.A. in both the concatenation and coalescent species trees (Fig. 2A).
240
241 The BOC analysis suggested that the two haplotypes of UTEX 2797 have divergent
242 evolutionary histories. This could lead to high levels of incongruence between the input gene
243 trees (in which UTEX 2797 may group in two locations reflecting the evolutionary histories of
244 its two haplotypes) and the inferred species tree (which forces each strain to appear only once).
245 Therefore, we performed a gene tree-species tree reconciliation analysis using GRAMPA (45) to
246 test whether a multi-labeled (MUL) species tree in which UTEX 2797 appeared twice was a
247 more parsimonious representation of the input gene trees compared to the original single-labeled
248 species tree. Seven of 29 possible MUL phylogenies had a smaller parsimony score, *i.e.*, required
249 fewer total gene duplication and loss events, compared to the single-labeled species tree (Fig. S8,
250 Table S12). In the most parsimonious MUL species tree, UTEX 2797 grouped in two locations:

251 with other A-type strains from the U.S.A (hereafter clade A1) and with CCMP2941 and
252 RCC3703 from Europe (hereafter clade A2) (Fig. 3). Plotting the genomic distribution of UTEX
253 2797 genes that group with either clade A1 or A2 revealed a striking pattern. Most scaffolds
254 formed syntenic pairs in the which one scaffold had a high proportion of genes that grouped with
255 subclade A1 and the second scaffold contained genes that grouped with subclade A2 (Fig. 3,
256 Table S13). To quantify the divergence between the UTEX 2797 haplotypes, we compared the
257 distribution of synonymous substitutions of UTEX 2797 genes that grouped with either the A1 or
258 A2 clades. This analysis revealed a bimodal distribution of *Ks* values in which A1 genes in
259 UTEX 2797 were essentially identical to A1 strains 12B1 and CCMP3037 (*Ks* = 0) and more
260 divergent to A2 strains RCC3703 and CCMP2941 (*Ks* = 0.009) (Fig. S9). This pattern was
261 reversed in A2 genes (Fig. S9). From these results we can infer that UTEX 2797 was formed via
262 the hybridization of an A1-like and an A2-like parent.
263
264 **Inter-strain variation in DNA content and haploid genome size.** Flow cytometry revealed that
265 DNA content varied dramatically among *P. parvum* strains (Fig. 4A). Strain 12B1 had the
266 smallest amount of nuclear DNA at 115.6 Mbp, while K-0081 (a B-type strain) had the largest at
267 845.6 Mbp. Large variation in DNA content between strains of *P. parvum* has previously been
268 interpreted as differences in ploidy (34). However, in our analysis, the relative DNA content
269 between strains did not discretely cluster around integer fold changes (Fig. 4A), as would be
270 expected if all variation in nuclear DNA content was due to differences in genome copy number
271 (*i.e.*, ploidy). Given the sequence-level divergence between strains in our analysis, it is possible
272 that differences between strains could extend to the size of their haploid genome as well.
273 Moreover, due to the limited understanding of the *P. parvum* lifecycle, the ploidy of different
274 strains was not immediately clear. This creates a circular puzzle, as we cannot infer the haploid
275 genome size of a strain from flow cytometry without *a priori* knowing the strain's ploidy.
276
277 To address this limitation, we used two sequence-based approaches to estimate the
278 relative haploid genome size between strains. First, k-mer frequency analyses were performed
279 using the Illumina reads from each strain. Reads were filtered to exclude possible contamination
280 and depth normalized to permit inter-strain comparison of k-mer coverage. The coverage of
281 maximal unique k-mers (CMUK) of the homozygous peak in k-mer frequency plots was used as

282 a proxy for haploid genome size, with larger CMUKs indicative of smaller haploid genomes and
283 vice versa. CMUK varied dramatically between strains with different prymnesin chemotypes,
284 ranging from an average of 55.8 in B-type producers to 120.3 in A-types (Fig. 4B, Table S14).
285 Average CMUK also varied between C-type strains with strains in the C1 clade having an
286 average CMUK of 104.5 compared to strains in C2 and C3 clades, which had and average
287 CMUK of 77.3 (Fig. 4B, Table S14). This pattern indicates considerable differences in haploid
288 genome size between clades with the A-type clade having the smallest haploid genome size and
289 the B-type clade having the largest.
290
291 The amount of heterozygosity also varied considerably between strains (Fig. 4B). In
292 addition to 12B1, six strains (A1 strain CCMP3037; B strains K-0374, KAC-39, RCC3426; and
293 C1 strains RCC191, RCC1433) showed little to no evidence of heterozygosity, which could
294 indicate that, like 12B1, these strains are haploid. However, in the case of A1 strain CCMP3037,
295 its flow cytometry estimated DNA content was twice that of 12B1 despite the two strains having
296 similar CMUKs peaks (Fig. 4A,B). This indicates that CCMP3037 is a diploid strain with low
297 levels of heterozygosity. For all other low heterozygosity strains in B and C1 clades, decreased
298 CMUK relative to 12B1 corresponds with increased DNA content, suggesting that these strains
299 are haploids. All other strains displayed comparatively moderate to high levels of heterozygosity,
300 indicating that they are either diploids (2n) or polyploid (3n or greater). Notably, strain 12A1 had
301 a pronounced heterozygous peak indicating extremely high levels of heterozygosity, similar to
302 that observed in UTEX 2797 (Fig. 4B). This, coupled with 12A1's high BOC against the UTEX
303 2797 diploid assembly, indicates that 12A1 is also a hybrid strain.
304
305 Average read coverage across 2699 single-copy orthogroups (SCOGs) was used as a
306 second sequence-based estimate of haploid genome size (Fig. 4C, Table S14). SCOG and
307 CMUK coverage estimates were in agreement for most strains, except for UTEX 2797, which
308 had significantly lower average SCOG coverage compared to the location of its CMUK peak
309 (Fig. 4B,C). This discrepancy is due to the nature of its hybrid genome and Hi-C scaffolded
310 assembly with resolved haplotypes. To be assigned to a SCOG, genes in UTEX 2797 have likely
311 returned to single-copy post hybridization, resulting in lower coverage compared to genes that
312 have been retained in duplicate. We used the Lander-Waterman equation to estimate the haploid

genome size for all strains based on the total length of the Illumina sequencing library divided by the median SCOG coverage, which served as an estimate of coverage for the entire genome (46, 47). This sequence-based estimate of haploid genome size ranged from 91 Mbp in A-type strain CCMP3031 to 203 Mbp in B-type strain K-0374 (Fig. 4D).

Cross-referencing total DNA content with the sequence-based haploid genome size, we were able to assign a predicted ploidy level for each strain (Fig. 4E, Table S14). Six strains were determined to be haploids, and seven (including hybrids 12A1 and UTEX 2797) were diploids. Two strains, A-type CCMP2941 and B-type K-0081 appear tetraploid. Using this predicted ploidy, we also determined flow cytometry-based estimates of haploid genome size (Fig. 4F, Table S14). The A clade has the smallest genome size (average = 130 Mbp), followed by the C1 clade (153 Mbp). The C3 strain (K-0252) had an estimated genome size of 204 Mbp, and the C2 clade had an average genome size of 225 Mbp. Lastly, B clade strains had the largest genomes (average = 281 Mbp), while also having the largest amount of intra-clade variation in genome size (Fig. 4F).

We investigated the relationship between DNA content and average cell volume and did not identify a clear trend (Fig. S10). When comparing between haploid strains, we found a weak association between cell size and prymnesin type, with B-types having the largest genomes and largest cells and the single A-type haploid (12B1) having both the smallest genome and cell size (Fig. S10B). However, we did not find a similar correlation when comparing between diploid strains. When looking within prymnesin A-type producing strains, we found a weak positive correlation between cell size and ploidy state, but we did not find a similar association within B- or C- type strains (Fig. S10C).

Accessory gene families and horizontal gene transfer. To investigate the impact of genome variation on gene family evolution in *P. parvum,* we performed a pangenome analysis of OrthoFinder-predicted gene families and identified a total of 47,043 orthogroups including 16,453 core orthogroups present in all strains; 20,738 accessory orthogroups present in 2-14 strains; and 9852 singleton orthogroups unique to a single strain (Table S15). The number of shared orthogroups generally clustered by prymnesin type, with some exceptions (Fig. 5A; Table

344 S16). Strains 12B1 and UTEX 2797 did not cluster with other A-types, likely in part due to the
345 differences between their Hi-C scaffolded assemblies and the other Illumina-only assemblies.
346 The C-types also clustered into two groups based on orthogroup membership, with C2 strains
347 (RCC1436 and UTEX 995) clustering separate from C1 and C3 strains (Fig. 5A).
348
349 A significant percentage of all orthogroups (44.1%) were variably present in 2-14 strains.
350 These accessory orthogroups belong to several enriched functional categories that could be
351 associated with metabolic and genome size variation in *P. parvum.* Several KEGG specialized
352 metabolic pathways, including those for the biosynthesis of type I polyketides, macrolides, and
353 nonribosomal peptides, were enriched in accessory orthogroups compared to core and singleton
354 orthogroups (BH adjusted p-value < 0.01; Fig. 5B; Table S17). Moreover, 31 Gene Ontology
355 (GO) categories were enriched in the accessory orthogroups, the most significantly enriched of
356 which was GO:0015074, DNA integration (BH adjusted p-value = 3.68E-73; Fig. 5C; Table
357 S17). Most orthogroups assigned to this GO category were annotated as integrase-like enzymes,
358 common components of lysogenic viruses and transposable elements. We investigated the 60
359 genes annotated with the DNA integration GO term (GO:0015074) in 12B1 and UTEX 2797 and
360 found that all genes fell within 400 bp of a predicted repeat. Many of these genes were found
361 within Ngaro LTRs, which intersected 31.3% of integrase-like genes in 12B1. This pattern
362 reveals that enrichment of the DNA integration GO term is driven by transposable element
363 integrase genes that have been incorporated into the gene model predictions. We investigated the
364 phylogenetic distribution of this pattern by performing functional enrichment on gene families
365 uniquely present in A-, B-, and C-type strains. We found that the DNA integration GO term was
366 enriched in the B- and C-type specific orthogroups (BH adjusted p-values = $3.94e^{-15}$ and $1.72e^{-4}$,
367 respectively) but was not significantly enriched in A-type specific orthogroups (Table S17),
368 suggesting that an expansion of transposon copy number may contribute to the greater haploid
369 genome sizes of the B- and C-type clades.
370
371 One potential source of accessory genes in a pangenome is horizontal gene transfer
372 (HGT) (48). To investigate the role of HGT in the genome evolution of *P. parvum,* we
373 performed an Alien Index (AI) analysis (49) and focused our investigation on HGTs into
374 *Prymnesium* after the genus diverged from other haptophytes. The AI screen flagged 95

375 orthogroups as candidate HGTs (Table S18). We used a custom phylogenetic pipeline to
376 manually evaluate all candidate HGTs and found that most (n = 61) were phylogenetically
377 inconclusive or lacked support for HGT. Of the 34 HGT candidates that passed manual
378 inspection, we focused on eleven HGTs with clear donor lineages with strong node support (Fig.
379 S11; Table S19). Two HGTs were likely acquired from eukaryotic donors: HGT01, a
380 bicarbonate transporter of diatom origin, and HGT02, a gene of unknown function that grouped
381 with pelagophytes. Eight HGTs (HGT03 - HGT10) were likely acquired from bacteria. All the
382 bacterially derived HGTs were enzyme-coding with diverse metabolic activities (Table S19). For
383 example, HGT09 grouped phylogenetically with sequences from marine bacteria (Fig. 5D),
384 including *bmp5*, a decarboxylating flavin-dependent halogenase involved in the biosynthesis of
385 polybrominated natural products in *Pseudoalteromonas* spp. (50). Lastly, HGT11 was a Clp
386 protease that grouped with large dsDNA megaviruses that infect eukaryotic algae (Fig. 5E).
387
388 Several horizontally transferred genes were part of the accessory genome and variably
389 present across strains (Table S19). Some of this variation is likely due to ancestral HGTs being
390 lost in some clades; *e.g.*, HGT07 and HGT10 both appear to have been gained in an ancestor of
391 *Prymnesium* and subsequently lost in the B-type clade (Fig. S11). HGT11, the Clp protease gene
392 of viral origin, was notable due to its presence in only four disparate strains: three A-type strains
393 (UTEX 2797, 12A1, and CCMP2941) and one C-type (RCC1433) (Fig. 5E). If this gene was
394 acquired in a shared ancestor of all four strains, at least six independent loss events would be
395 required to explain this presence/absence pattern. An alternative explanation could be that
396 multiple independent acquisitions of a viral Clp protease have occurred in *P. parvum.* The
397 genomic neighborhood of HGT11 in UTEX 2797 contains six additional genes with a top
398 BLAST hit to EhV-86, a dsDNA megavirus that infects the haptophyte *Emiliania huxleyi* (51).
399 These additional genes were not recovered in our primary screen for HGT due to a limited
400 number of hits to proteins in the NCBI RefSeq database, which were insufficient for
401 phylogenetic analysis (see SI Materials and Methods). However, their sequence similarity
402 suggests that a block of viral genes was likely acquired in a single HGT event in the lineage
403 leading to UTEX 2797 (Fig. S12). We checked for shared synteny between UTEX 2797 and the
404 three additional strains that contain HGT11. In 12A1, HGT11 is located on a 20-gene scaffold
405 (Scaf652979) which shares nearly perfect synteny with UTEX 2797, including five genes within

406 the proposed viral fragment (Fig. S13), which suggests that UTEX 2797 and 12A1 share the
407 same HGT event. In contrast, the scaffolds that contain HGT11 in CCMP2941 and RCC1433 are
408 syntenic with each other, but neither are syntenic with UTEX 2797 (Fig. S13). This suggests that
409 the HGT that gave rise to the Clp protease in CCMP2941 and RCC1433 was likely independent
410 from that of UTEX 2797 and 12A1. Long-read assemblies of additional strains including 12A1,
411 CCMP2941, and RCC1433 are needed to confirm this pattern.
412
413 **DISCUSSION**
414 The harmful algal bloom-forming eukaryote, *Prymnesium parvum*, possesses considerable
415 genomic diversity. Combining data from synonymous substitutions, reference genome coverage,
416 phylogenetics, and genome size, the fifteen strains in our analysis can be subdivided into six
417 distinct clades consisting of at least three cryptic species (Fig. 4). Evidence for these cryptic
418 species comes from the fact that prymnesin chemotype is phylogenetically segregated, which
419 indicates that the *P. parvum* A-, B-, and C-type clades are reproductively isolated. Here, we
420 provide further support for these clades being separate cryptic species based on the extreme
421 variation in their genome size, as excessive chromosome-level differences will likely inhibit
422 proper chromosome pairing during meiosis. The most dramatic difference in genome size
423 occurred between the sister clades of A- and B- type strains (Fig. 4F). These clades show limited
424 substitutions at synonymous sites (*Ks* = 0.02; Fig. 2A), which suggests that the changes to their
425 genome sizes have been relatively recent. Possible mechanisms for genome expansion in B-type
426 strains includes whole-genome duplication (WGD) and/or proliferation of transposable elements.
427 Our analysis of *Ks* distributions showed no evidence of recent WGD in the last common ancestor
428 of the B-clade (Fig. S4). Instead, gene functions associated with transposable elements were
429 enriched in gene families unique to B-type strains (Table S17), suggesting that the increase in
430 their genome size could be due to an increase in transposition activity of these elements. A high-
431 quality reference genome of a B-type strain would enable further investigation of repeat diversity
432 and expansion across these two clades and clarify the mechanism(s) of genome size variation in
433 *P. parvum*.
434
435 Although *P. parvum* is currently considered a single morphospecies, earlier taxonomic
436 descriptions split the species in two (*P. parvum* and *P. patelliferum*) based on slight differences

437 in the morphology of the organic scales that cover their cell surface (52). When the rDNA ITS1
438 region was found to be identical in strains of *P. parvum* and *P. patelliferum* isolated from the
439 same geographic location, the two species were merged into one (53). Later, two strains
440 originally identified as *P. parvum* (K-0081 and strain RL10parv93, not included in our analysis)
441 were found to have higher amounts of DNA compared to three strains labeled *P. patelliferum* (K-
442 0252, RCC191, and a third strain not assessed here, RLpat93) (34). This observation led Larsen
443 and Edvardsen (1998) to propose a cryptic lifecycle for *P. parvum* that alternates between
444 flagellated haploid and diploid forms. Our analysis identified strains of different ploidy states,
445 supporting the existence of a cryptic sexual lifecycle. Moreover, the existence of a hybrid strain
446 (UTEX 2797) with two phylogenetically divergent parental genotypes indicates that syngamy
447 can occur in *P. parvum*. Our assessment of ploidy across these taxa, which includes flow
448 cytometry and read-based estimates of haploid genome size, indicates that K-0081 is a tetraploid,
449 not a diploid as previously thought, and that K-0252 is a diploid and not a haploid (Fig. 4). Given
450 the inclusion of these strains in the hypothesis about scale morphology and ploidy, our new
451 results suggest that scale morphology is not diagnostic of ploidy state. Analysis of the fifteen
452 strains reported here places variation in ploidy and genome size in a phylogenetic framework,
453 facilitating future investigation into cellular morphology and sexual reproduction in *P. parvum*.
454
455 Our phylogenomic analysis also reveals that UTEX 2797, a Texas *P. parvum* strain
456 frequently used in growth and toxicity experiments, is a hybrid of A1 and A2 parents (Fig. 4A).
457 It is intriguing that the A1 strains in our analysis are from the U.S.A while the A2 strains are
458 from Europe, and it is tempting to speculate that hybridization was a result of a recent
459 introduction of the A2 lineage in Texas. Our analyses suggest that the high heterozygosity strain
460 12A1 is a hybrid as well, but it is unclear from our analysis if 12A1 arose from the same
461 hybridization event as UTEX 2797. Further investigation into the origin of these strains requires
462 additional taxon sampling of both A1 and A2 lineages as well as long-read based assemblies of
463 12A1 to resolve its two haplotypes. The evolutionary outcome of hybridization in *P. parvum* is
464 also unknown. Are these hybrid strains capable of sexual reproduction? Genome divergence
465 between A1 and A2 clades may be such that homologous chromosomes are unable to pair
466 correctly during meiosis. If so, these hybrids may be effectively trapped as diploids and only able
467 to reproduce asexually. Alternatively, whole-genome duplication following hybridization

468 (allopolyploidy) has been shown to restore fertility in yeast (54), which could make these *P.*
469 *parvum* hybrids reproductively haploids. Characterizing the sexual lifecycle of *P. parvum* would
470 allow researchers to address several outstanding questions regarding the ecological and
471 evolutionary consequences of extreme genetic variation in these toxic bloom forming eukaryotes.
472 Knowledge of the *P. parvum* lifecycle would enable the design of mating tests to determine the
473 reproductive status of hybrid strains. Mating tests could also be used to assess reproductive
474 isolation between divergent populations, as has been done in other morphologically
475 indistinguishable cryptic species complexes (55–57).
476
477 Previous work indicates that HABs of *P. parvum* are comprised of multiple genotypes
478 (21, 24, 58), and our results reveal that the genome-level differences between these genotypes
479 can be dramatic. For example, strains 12A1 and 12B1 were isolated from the same Texas bloom
480 in 2010 and consistently show different cell-level behaviors involved in toxicity and predation of
481 microalgal prey (21, 24). Here, we show that characteristic differences between 12A1 and 12B1
482 extend to large differences in their genomes as well, with 12A1 being a hybrid diploid and 12B1
483 having a streamlined haploid genome. Moreover, significant gene-level differences exist between
484 these two strains, with 7% (n = 1485) of 12B1 orthogroups absent in 12A1. An even larger
485 percentage, 22% (n = 5709), of 12A1 orthogroups are absent in 12B1 (Table S16), including one
486 that was horizontally acquired (HGT11; Fig. 5E). Gene families that are variably present across
487 strains in our analysis include those for the biosynthesis of type I polyketides, (Fig. 5B), the class
488 of specialized metabolites that includes prymnesins (17). This is unsurprising given the structural
489 diversity in prymnesins that have already been characterized in *P. parvum* (18, 19). However, it
490 is notable that 12A1 and 12B1, both A-type strains, show variable representation of eighteen
491 orthogroups assigned to the type I polyketide KEGG pathway (map01052; Table S17, Table
492 S15). If any of these orthogroups are involved in toxin biosynthesis, phenotypic differences
493 between these two strains could extend to their toxin profiles as well. The coding capacity of *P.*
494 *parvum* has also been expanded by HGT. Prymnesins are halogenated metabolites, and though
495 we have not functionally characterized it, it is tantalizing to note that a halogenase is among the
496 HGT candidates in the *P. parvum* genomes (HGT09; Fig. 5D). HGT is often a source of
497 metabolic innovation (*e.g.,* 49, 59–61) and elucidating the prymnesin pathway will determine if
498 this gene is relevant to the production of these metabolites. More work is needed to identify the

499 genetic factors associated with toxin production and the phenotypic differences between these
500 diverse strains of *P. parvum.* Additionally, work is needed to identify the selective advantage of
501 different phenotypes and whether such advantages fluctuate between bloom and non-bloom
502 conditions. The genome assemblies and phylogenomic analysis reported here are thus essential
503 resources that will enable future investigation into the eco-physiological consequences of hidden
504 genetic diversity in this HAB-forming morphospecies.
505

506 **MATERIALS AND METHODS**

507 Strains and their respective media types and culturing conditions are summarized in Table S8.
508 Flow cytometry for genome size estimation and LC-MS for prymnesin chemotyping were
509 performed as described in SI Materials and Methods. Details on genome sequencing and
510 assembly, repeat and gene prediction, and assessment of assembly completeness are described in
511 SI Materials and Methods. Analysis of synteny within the UTEX 2797 Hi-C scaffolded genome
512 assembly and between assemblies of UTEX 2797 and other *P. parvum* strains was performed
513 using the online Comparative Genomics Platform (CoGe) as described in SI Materials and
514 Methods. Orthologous gene families, individual gene trees, concatenation and coalescent-based
515 species trees, and gene tree-species tree reconciliation was performed using the phylogenomics
516 workflow summarized in Fig. S14 and described in SI Materials and Methods. Breadth of
517 coverage (BOC) was calculated from alignments of subsampled, filtered Illumina reads to the
518 12B1 and UTEX 2797 scaffolded assemblies. See SI Materials and Methods for details on read
519 filtration to remove contamination and subsampling to control for library differences between
520 strains. Depth of coverage was calculated on a per-base level and BOC was calculated as the
521 proportion of base-pairs in the final assembly that had coverage greater than N coverage. BOC
522 using coverage N = 0, 5, 10, 20, and 50 were all evaluated. Final BOC tables for each assembly
523 can be found in Tables S11, and additional details are provided in in SI Materials and Methods.
524 Heterozygosity, coverage of maximal unique k-mers, and SCOG read coverage were estimated
525 using the subsampled, filtered Illumina gDNA reads as described in SI Materials and Methods.
526 The haploid genome size (G) was calculated using the Lander-Waterman equation (46): $G =$
527 LN/C where LN is the total combined length (in base pairs) of the input Illumina reads and C is
528 coverage as estimated by SCOG read depth. All additional method details including cell size

529 estimation, functional annotation, tests for functional enrichment, and detection of horizontal
530 gene transfer are described in SI Materials and Methods.
531
532 **DATA AVAILABILITY**
533 All genome assemblies, predicted CDS and protein sequences, multiple sequence alignments,
534 tree files, and other related data files are available through FigShare
535 (https://doi.org/10.6084/m9.figshare.21376500). Scripts are available through GitHub
536 (https://github.com/WisecaverLab/Pparvum_genome_diversity). Raw sequencing reads have
537 been deposited in the Sequence Read Archive database under accession number PRJNA807128.
538 LC-MS data from chemotyping is available in the EBI MetaboLights database under accession
539 MTBLS5893. Strains 12B1 and 12A1 have been deposited in the UTEX Culture Collection of
540 Algae at UT-Austin under UTEX accessions XXXX and XXXX).
541
542 **ACKNOWLEDGEMENTS**
543 We thank members of the Wisecaver lab, Dr. Jody Banks, and Dr. Brian Dilkes for helpful
544 discussions as well as Dr. Clint Chapple and Dr. Scott McAdam for plant tissue for flow
545 cytometry. This work was conducted in part using the resources of the Rosen Center for
546 Advanced Computing at Purdue University and well as through use of the Flow Cytometry and
547 Cell Separation Facility at the Bindley Bioscience Center at Purdue, a core facility of the NIH-
548 funded Indiana Clinical and Translational Sciences Institute. This work was supported by the
549 National Institute for Environmental Health Sciences under grants F32-ES032276 to TRF and
550 R21-ES032056 to BSM and the National Science Foundation under grant DEB-1831493 to JHW
551 and WWD.
552

REFERENCES

1. D. L. Roelke, *et al.*, A chronicle of a killer alga in the west: ecology, assessment, and management of Prymnesium parvum blooms. *Hydrobiologia* **764**, 29–50 (2016).

2. C. J. Gobler, *et al.*, Ocean warming since 1982 has expanded the niche of toxic algal blooms in the North Atlantic and North Pacific oceans. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 4975–4980 (2017).

3. G. M. Hallegraeff, *et al.*, Perceived global increase in algal blooms is attributable to intensified monitoring and emerging bloom impacts. *Commun. Earth Environ.* **2**, 1–10 (2021).

4. M. L. Wells, *et al.*, Future HAB science: Directions and challenges in a changing climate. *Harmful Algae* **91**, 101632 (2020).

5. D. L. Roelke, S. R. Manning, "Harmful algal species fact sheet: Prymnesium parvum (Carter) 'golden algae'" in *Harmful Algal Blooms*, (John Wiley & Sons, Ltd, 2018), pp. 629–632.

6. G. M. Southard, L. T. Fries, A. Barkoh, Prymnesium parvum: The Texas experience. *JAWRA J. Am. Water Resour. Assoc.* **46**, 14–23 (2010).

7. S. Tábora-Sarmiento, R. Patiño, C. Portillo-Quintero, C. Coldren, Air, land, and water variables associated with the first appearance and current spatial distribution of toxic Prymnesium parvum blooms in reservoirs of the Southern Great Plains, USA. *Sci. Total Environ.* **836**, 155567 (2022).

8. J. W. Baker, *et al.*, Growth and toxicity of Prymnesium parvum (Haptophyta) as a function of salinity, light, and temperature. *J. Phycol.* **43**, 219–227 (2007).

9. A. Brutemark, E. Granéli, Role of mixotrophy and light for growth and survival of the toxic haptophyte Prymnesium parvum. *Harmful Algae* **10**, 388–394 (2011).

10. E. Granéli, B. Edvardsen, D. L. Roelke, J. A. Hagström, The ecophysiology and bloom dynamics of Prymnesium spp. *Harmful Algae* **14**, 260–270 (2012).

11. K. Nygaard, A. Tobiesen, Bacterivory in algae: A survival strategy during nutrient limitation. *Limnol. Oceanogr.* **38**, 273–279 (1993).

12. U. Tillmann, Phagotrophy by a plastidic haptophyte, Prymnesium patelliferum. *Aquat. Microb. Ecol.* **14**, 155–160 (1998).

13. U. Tillmann, Kill and eat your predator: a winning strategy of the planktonic flagellate Prymnesium parvum. *Aquat. Microb. Ecol.* **32**, 73–84 (2003).

14. E. J. Remmel, K. D. Hambright, Toxin-assisted micropredation: experimental evidence shows that contact micropredation rather than exotoxicity is the role of Prymnesium toxins. *Ecol. Lett.* **15**, 126–132 (2012).

15. W. F. Carvalho, E. Granéli, Contribution of phagotrophy versus autotrophy to Prymnesium parvum growth under nitrogen and phosphorus sufficiency and deficiency. *Harmful Algae* **9**, 105–115 (2010).

16. S. R. Manning, J. W. La Claire, Prymnesins: toxic metabolites of the golden alga, Prymnesium parvum Carter (Haptophyta). *Mar. Drugs* **8**, 678–704 (2010).

17. S. A. Rasmussen, *et al.*, Chemical diversity, origin, and analysis of phycotoxins. *J. Nat. Prod.* **79**, 662–673 (2016).

18. S. A. Rasmussen, *et al.*, Chemodiversity of ladder-frame prymnesin polyethers in Prymnesium parvum. *J. Nat. Prod.* **79**, 2250–2256 (2016).

19. S. B. Binzer, *et al.*, A-, B- and C-type prymnesins are clade specific compounds and chemotaxonomic markers in Prymnesium parvum. *Harmful Algae* **81**, 10–17 (2019).

20. K. Anestis, *et al.*, Polyketide synthase genes and molecular trade-offs in the ichthyotoxic species Prymnesium parvum. *Sci. Total Environ.* **795**, 148878 (2021).

21. W. W. Driscoll, N. J. Espinosa, O. T. Eldakar, J. D. Hackett, Allelopathy as an emergent, exploitable public good in the bloom-forming microalga *Prymnesium parvum*. *Evolution* **67**, 1582–1590 (2013).

22. H. E. Blossom, *et al.*, Prymnesium parvum revisited: relationship between allelopathy, ichthyotoxicity, and chemical profiles in 5 strains. *Aquat. Toxicol. Amst. Neth.* **157**, 159–166 (2014).

23. N. Medić, E. Varga, D. B. V. de Waal, T. O. Larsen, P. J. Hansen, The coupling between irradiance, growth, photosynthesis and prymnesin cell quota and production in two strains of the bloom-forming haptophyte, Prymnesium parvum. *Harmful Algae* **112**, 102173 (2022).

24. W. W. Driscoll, Behavioral differences underlie toxicity and predation variation in blooms of *Prymnesium parvum* (2022).

25. A. Larsen, W. Eikrem, E. Paasche, Growth and toxicity in *Prymnesium patelliferum* (Prymnesiophyceae) isolated from Norwegian waters. *Can. J. Bot.* **71**, 1357–1362 (1993).

26. A. Larsen, S. Bryant, Growth rate and toxicity of *Prymnesium parvum* and *Prymnesium patelliferum* (haptophyta) in response to changes in salinity, light and temperature. *Sarsia* **83**, 409–418 (1998).

27. M. L. Lysgaard, L. Eckford-Soper, N. Daugbjerg, Growth rates of three geographically separated strains of the ichthyotoxic Prymnesium parvum (Prymnesiophyceae) in response to six different pH levels. *Estuar. Coast. Shelf Sci.* **204**, 98–102 (2018).

621 28. A. Talarski, S. R. Manning, J. W. La Claire, Transcriptome analysis of the euryhaline alga,
622 Prymnesium parvum (Prymnesiophyceae): Effects of salinity on differential gene
623 expression. *Phycologia* **55**, 33–44 (2016).

624 29. R. H. Rashel, R. Patiño, Influence of genetic background, salinity, and inoculum size on
625 growth of the ichthyotoxic golden alga (Prymnesium parvum). *Harmful Algae* **66**, 97–104
626 (2017).

627 30. R. B. Taylor, *et al.*, Suspect and non-target screening of acutely toxic Prymnesium parvum.
628 *Sci. Total Environ.* **715**, 136835 (2020).

629 31. E. T. Richardson, R. Patiño, Growth of the harmful alga, Prymnesium parvum
630 (Prymnesiophyceae), after gradual and abrupt increases in salinity. *J. Phycol.* **57**, 1335–
631 1344 (2021).

632 32. R. B. Taylor, B. N. Hill, L. M. Langan, C. K. Chambliss, B. W. Brooks, Sunlight
633 concurrently reduces Prymnesium parvum elicited acute toxicity to fish and prymnesins.
634 *Chemosphere* **263**, 127927 (2021).

635 33. T. Cavaller-Smith, *The evolution of genome size* (John Wiley and Sons Inc.,New York, NY,
636 1985) (March 21, 2022).

637 34. A. Larsen, B. Edvardsen, Relative ploidy levels in Prymnesium parvum and P. patelliferum
638 (Haptophyta) analyzed by flow cytometry. *Phycologia* **37**, 412–424 (1998).

639 35. J. Fulnecková, *et al.*, A broad phylogenetic survey unveils the diversity and evolution of
640 telomeres in eukaryotes. *Genome Biol. Evol.* **5**, 468–483 (2013).

641 36. B. A. Read, *et al.*, Pan genome of the phytoplankton Emiliania underpins its global
642 distribution. *Nature* **499**, 209–213 (2013).

643 37. C. J. Hulatt, R. H. Wijffels, M. C. Posewitz, The genome of the Haptophyte Diacronema
644 lutheri (Pavlova lutheri, Pavlovales): A model for lipid biosynthesis in eukaryotic algae.
645 *Genome Biol. Evol.* **13**, evab178 (2021).

646 38. D. Chen, *et al.*, Multi-omics analyses provide insight into the biosynthesis pathways of
647 fucoxanthin in Isochrysis galbana. *Genomics Proteomics Bioinformatics* (2022)
648 https:/doi.org/10.1016/j.gpb.2022.05.010 (August 16, 2022).

649 39. P. Jones, *et al.*, InterProScan 5: genome-scale protein function classification. *Bioinformatics*
650 **30**, 1236–1240 (2014).

651 40. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO:
652 assessing genome assembly and annotation completeness with single-copy orthologs.
653 *Bioinformatics* **31**, 3210–3212 (2015).

654 41. D. M. Emms, S. Kelly, OrthoFinder: solving fundamental biases in whole genome
655 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157
656 (2015).

657 42. S. Naser-Khdour, B. Q. Minh, R. Lanfear, Assessing confidence in root placement on
658 phylogenies: An empirical study using nonreversible models for mammals. *Syst. Biol.*,
659 syab067 (2021).

660 Minh BQ, Hahn MW, Lanfear R. 2020. New Methods to Calculate Concordance Factors for
661 Phylogenomic Datasets. *Mol. Biol. Evol.* 37:2727–2733.

662 44. P. Y. Novikova, *et al.*, Sequencing of the genus Arabidopsis identifies a complex history of
663 nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.* **48**, 1077–
664 1082 (2016).

665 45. G. W. C. Thomas, S. H. Ather, M. W. Hahn, Gene-tree reconciliation with MUL-trees to
666 resolve polyploidy events. *Syst. Biol.* **66**, 1007–1018 (2017).

667 46. E. S. Lander, M. S. Waterman, Genomic mapping by fingerprinting random clones: A
668 mathematical analysis. *Genomics* **2**, 231–239 (1988).

669 47. J. M. Pflug, V. R. Holmes, C. Burrus, J. S. Johnston, D. R. Maddison, Measuring genome
670 sizes using read-depth, k-mers, and flow cytometry: methodological comparisons in beetles
671 (Coleoptera). *G3 Genes Genomes Genet.* **10**, 3047–3060 (2020).

672 48. S. J. Sibbald, L. Eme, J. M. Archibald, A. J. Roger, Lateral gene transfer mechanisms and
673 pan-genomes in eukaryotes. *Trends Parasitol.* **36**, 927–941 (2020).

674 49. J. H. Wisecaver, W. G. Alexander, S. B. King, C. Todd Hittinger, A. Rokas, Dynamic
675 evolution of nitric oxide detoxifying flavohemoglobins, a family of single-protein
676 metabolic modules in bacteria and eukaryotes. *Mol. Biol. Evol.*, msw073 (2016).

677 50. V. Agarwal, *et al.*, Biosynthesis of polybrominated aromatic organic compounds by marine
678 bacteria. *Nat. Chem. Biol.* **10**, 640–647 (2014).

679 51. W. H. Wilson, *et al.*, Complete genome sequence and lytic phase transcription profile of a
680 Coccolithovirus. *Science* **309**, 1090–1092 (2005).

681 52. J. C. Green, D. J. Hibberd, R. N. Pienaar, The taxonomy of Prymnesium
682 (Prymnesiophyceae) including a description of a new cosmopolitan species, P. patellifera
683 sp. nov., and further observations on P. parvum N. carter. *Br. Phycol. J.* **17**, 363–382
684 (1982).

685 53. A. Larsen, L. K. Medlin, Inter- and intraspecific genetic variation in twelve Prymnesium
686 (Haptophyceae) clones. *J. Phycol.* **33**, 1007–1015 (1997).

687 54. R. A. Ortiz-Merino, *et al.*, Evolutionary restoration of fertility in an interspecies hybrid
688 yeast, by whole-genome duplication after a failed mating-type switch. *PLOS Biol.* **15**,
689 e2002128 (2017).

690 55. T. M. Sonneborn, The Paramecium aurelia complex of fourteen sibling species. *Trans. Am.*
691 *Microsc. Soc.* **94**, 155–178 (1975).

692 56. E. M. Simon, D. L. Nanney, F. P. Doerder, The "Tetrahymena pyriformis" complex of
693 cryptic species. *Biodivers. Conserv.* **17**, 365–380 (2008).

694 57. U. John, *et al.*, Formal revision of the Alexandrium tamarense species complex
695 (Dinophyceae) taxonomy: The introduction of five species with emphasis on molecular-
696 based (rDNA) classification. *Protist* **165**, 779–804 (2014).

697 58. F. S. Barreto, C. R. Tomas, M. A. McCartney, AFLP fingerprinting shows that a single
698 Prymnesium parvum harmful algal bloom consists of multiple clones. *J. Hered.* **102**, 747–
699 752 (2011).

700 59. K. I. Verster, *et al.*, Horizontal Transfer of Bacterial Cytolethal Distending Toxin B Genes
701 to Insects. *Mol. Biol. Evol.* **36** (2019).

702 60. T. A. Richards, *et al.*, Horizontal gene transfer facilitated the evolution of plant parasitic
703 mechanisms in the oomycetes. *Proc. Natl. Acad. Sci.* **108**, 15258–15263 (2011).

704 61. J. C. Slot, D. S. Hibbett, Horizontal transfer of a nitrate assimilation gene cluster and
705 ecological transitions in fungi: a phylogenetic study. *PLoS ONE* **2**, e1097 (2007).

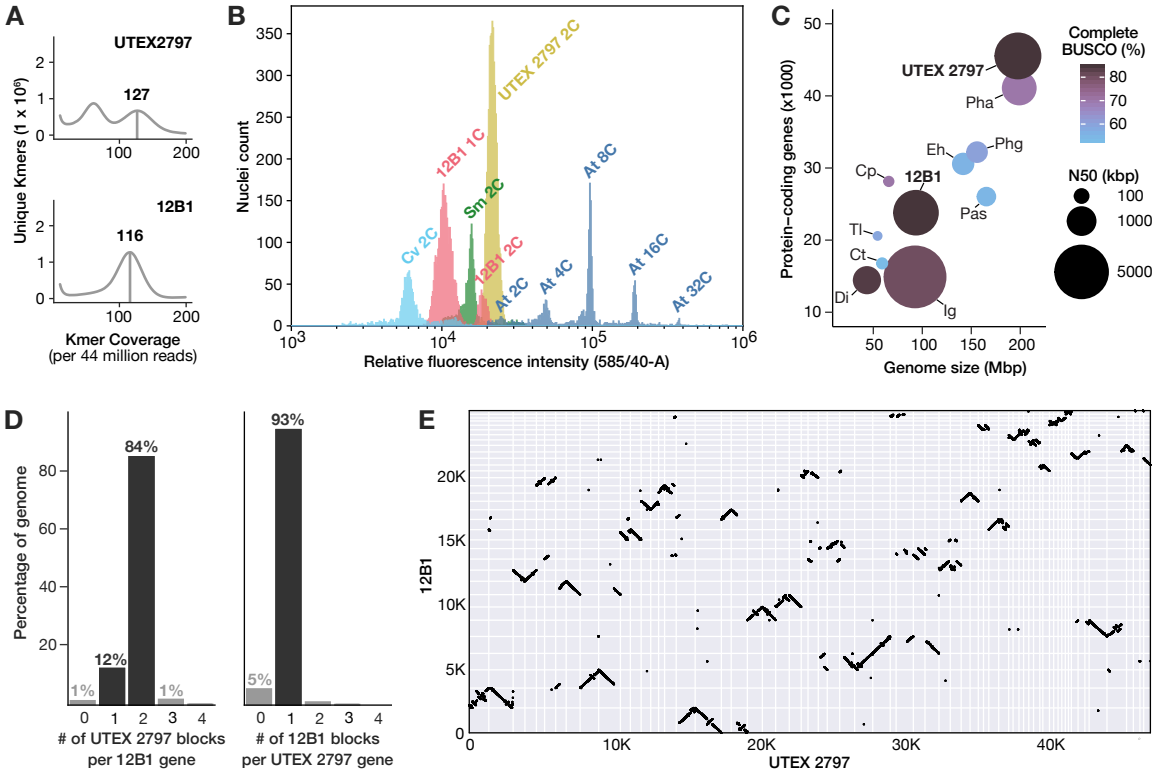
706 62. D. W. Galbraith, Endoreduplicative standards for calibration of flow cytometric C-Value
707 measurements. *Cytometry A* **85**, 368–374 (2014).

708 63. F. Supek, M. Bošnjak, N. Škunca, T. Šmuc, REVIGO summarizes and visualizes long lists
709 of gene ontology terms. *PLOS ONE* **6**, e21800 (2011).

710

711 **FIGURES**

712



713

714

715 **Figure 1. Genome metrics for *P. parvum* Hi-C scaffolded long-read assemblies.** A) K-mer
716 frequency plots showing estimated heterozygosity for strains UTEX 2797 and 12B1**.** The
717 homozygous k-mer peaks are indicated by vertical bars, with the number above the homozygous
718 peaks indicating the coverage of maximal unique k-mers (CMUKs). UTEX 2797 has high
719 heterozygosity as indicated by the large second peak at half the k-mer coverage of the
720 homozygous peak. B) Histogram indicating the relative fluorescent intensity of propidium iodide
721 stained nuclei for strains 12B1 (pink) and UTEX 2797 (yellow) relative to genome standards:
722 Cv, *Chlorella vulgaris* (light blue); Sm, *Selaginella moellendorfii* (green); At, *Arabidopsis*
723 *thaliana* Col-0 (dark blue). The endoreduplicative nature of the *A. thaliana* tissue enabled
724 identification of 2C, 4C, 8C, 16C, and 32C nuclei in this species (62). C) Comparison of
725 assembly completeness and contiguity across eleven haptophyte genome assemblies. Cp,
726 *Chrysochromulina parva*; Ct, *Chrysochromulina tobinii*; Dl, *Diacronema lutheri*; Eh, *Emiliania*
727 *huxleyi*; Ig, *Isochrysis galbana*; Pas, Pavlovales sp.; Pha, *Phaeocystis antarctica*; Phg,
728 *Phaeocystis globosa*; Tl, *Tisochrysis lutea*. D) Ratio of syntenic block counts between *P. parvum*
729 genome assemblies for strains 12B1 and UTEX 2797. Syntenic blocks of UTEX 2797 per 12B1
730 gene (left) and syntenic blocks of 12B1 per UTEX 2797 gene (right) are shown, indicating a
731 clear 1:2 pattern of 12B1 to UTEX 2797. E) Macrosynteny of the 12B1 and UTEX 2797
732 genomes. Syntenic gene pairs are denoted by black points and positionally oriented by scaffold
733 (grid squares).

734

735

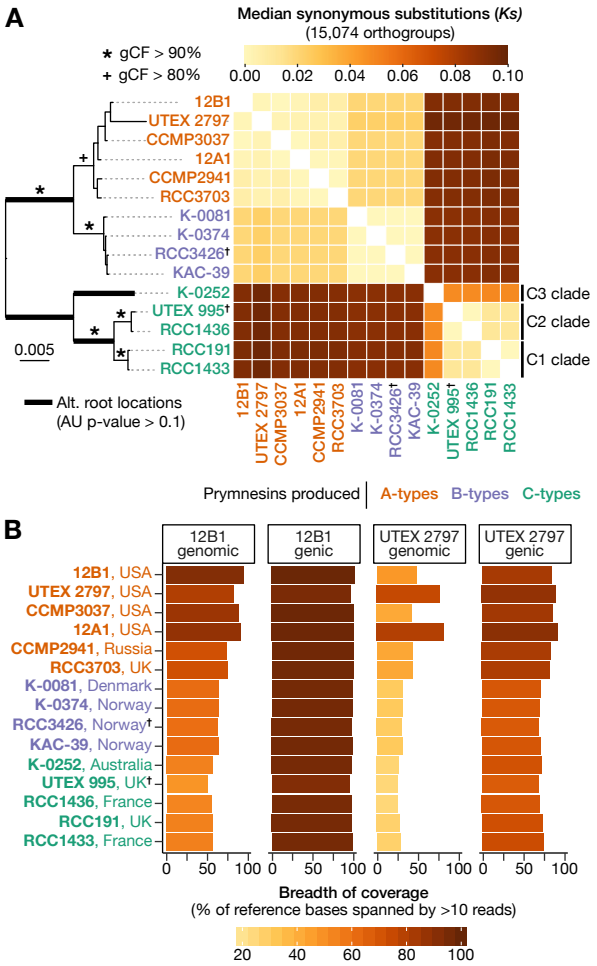


Figure 2. Phylogenomic and breadth of coverage analyses in *P. parvum.* A) Concatenation-based ML species phylogeny showing three possible root locations (bold branches). The heatmap shows the median synonymous substitutions per synonymous site (*Ks*) between all strains. C) Breadth of coverage (BOC) bar plots showing the percent of 12B1 and UTEX 2797 reference bases spanned by >10 Illumina reads from each strain (See Tables S11). BOC is provided for all genomic positions and genic space as separate statistics. The chemotypes of strains RCC3426 and UTEX 995 (†) are inferred based on phylogenetic placement.

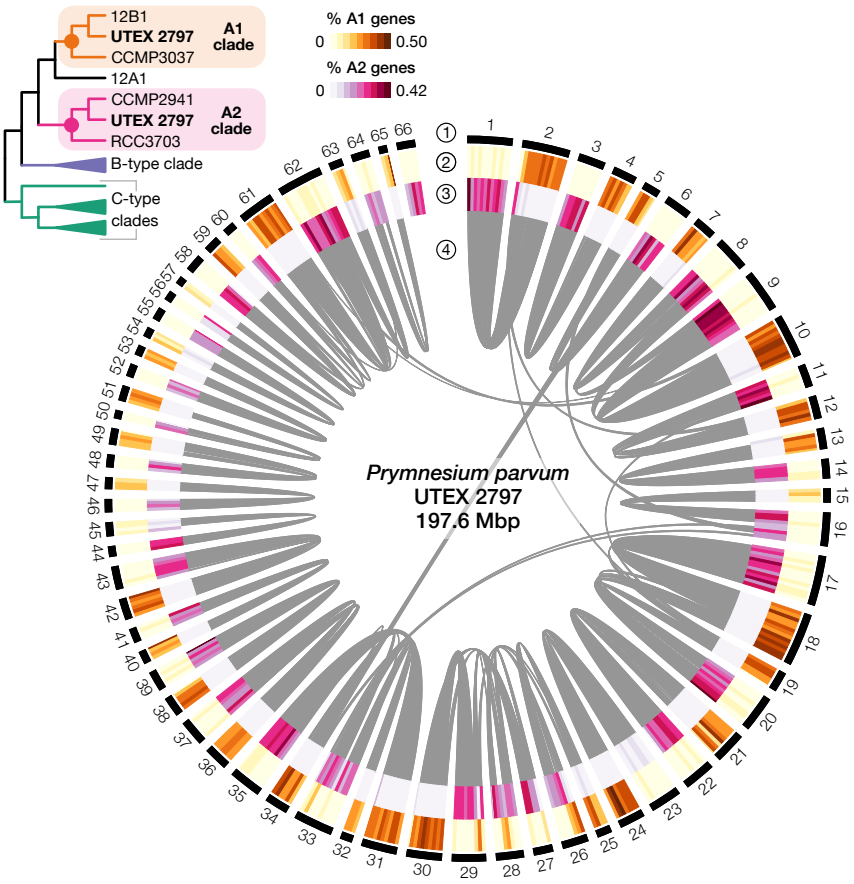


Figure 3. Phylogenetically distinct haplotypes of *P. parvum* strain UTEX 2797. Circos plot showing the 66 scaffolds of UTEX 2797 with four tracks (1) outer black track indicates scaffolds, (2) orange heatmap illustrates the percentage of genes in 50 kbp windows that group within the A1 clade, (3) pink heatmap indicates the percentage of genes in the same 50 kbp windows that group within the A2 clade, and (4) syntenic blocks (≥ 15 syntenic genes per block) are designated as gray bands. No branch support threshold was imposed for A1/A2 clade calls; the general pattern was the same but less vibrant when an ultrafast branch support threshold of 90 was imposed (Fig. S15). Multi-labeled (MUL) species tree (top left) shows the topology of the A1 and A2 clades as determined by the GRAMPA gene tree-species tree reconciliation analysis.

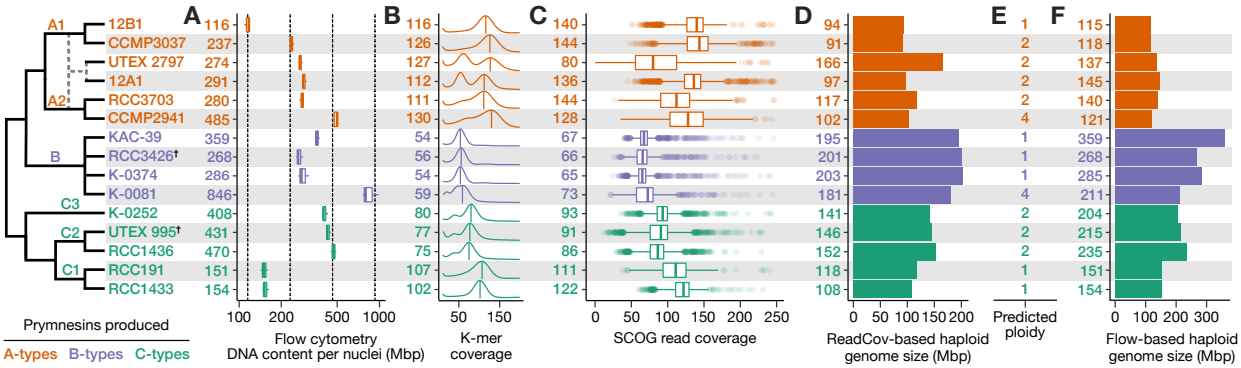


Figure 4. Summary of ploidy, heterozygosity, and genome size diversity in *P. parvum.* Evolutionary model (left) depicts strain relationships, including a predicted hybridization (dashed lines) giving rise to strains UTEX 2797 and 12A1. The chemotypes of strains RCC3426 and UTEX 995 (†) are inferred based on phylogenetic placement. A) Boxplots depicting total DNA content per nuclei (Mbp) based on flow cytometry. Numbers (left) indicate mean Mbp per strain. Vertical dashed lines indicate 1x, 2x, 4x, 8x Mbp relative to 12B1. B) K-mer coverage plots depict estimated heterozygosity; numbers indicate CMUKs, *i.e.*, the homozygous k-mer peaks labeled by the labeled by vertical bars. Heterozygosity can be qualitatively assessed by the presence and relative height of a second peak at half the k-mer coverage of the homozygous peak. C) Boxplots depicting the distribution of read coverage for 2699 SCOGs; numbers indicate median read coverage. D) Haploid genome size was estimated using read coverage (total read length divided by median Illumina read coverage of SCOGs). E) Predicted ploidy was determined by cross-referencing DNA content (A) with sequencing-based estimated of genome size (B-D). F) Haploid genome size was estimated using flow cytometry (total DNA content divided by proposed ploidy).

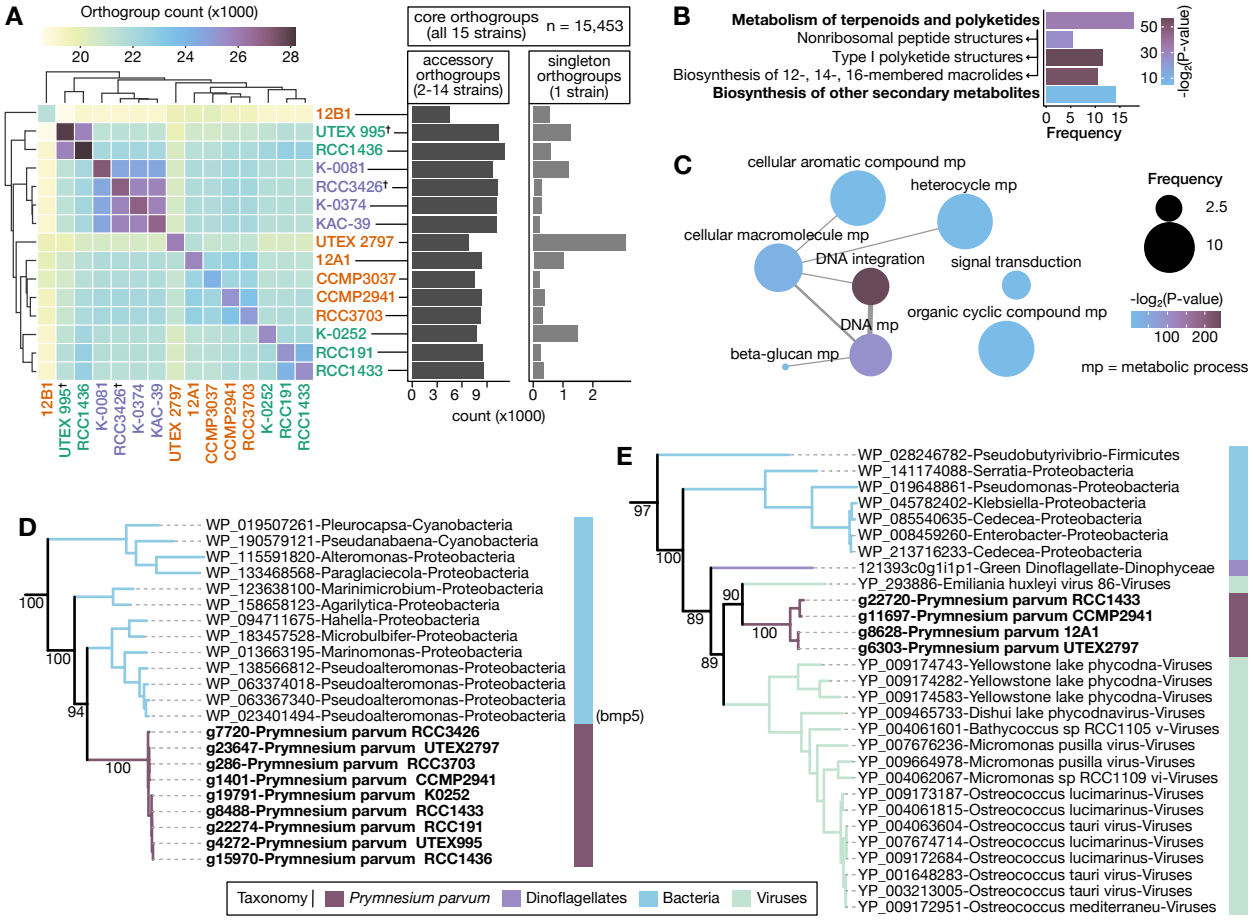


Figure 5. Pan genome analysis and horizontal gene transfer in *P. parvum.* A) Hierarchically clustered heatmap showing the number of orthogroups shared by each strain pair. Strains are colored based on the prymnesin type produced (as in Fig. 2). The chemotypes of strains RCC3426 and UTEX 995 (†) are inferred based on phylogenetic placement. Center diagonal indicates the total number of orthogroups, including singletons, present in each strain. Bar charts indicate the number of accessory orthogroups and singleton orthogroups in each strain. B) Significantly enriched KEGG pathways (unbolded) and pathway categories (bolded); arrows point to KEGG pathway parent category. Bar height indicates frequency of the annotation in accessory orthogroups with one or more KEGG annotations. C) Significantly enriched GO categories in accessory orthogroups. Width of network edges indicate the degree of similarity between GO terms as calculated by REVIGO (63). Bubble size indicates frequency of the annotation in accessory orthogroups with one or more GO annotations. See Table S17 for list of all functional enrichment tests performed. D) Flavin-dependent halogenase phylogeny showing HGT from marine bacteria. The *bmp5* gene functionally characterized in *Pseudoalteromonas* is labeled. E) Clp protease phylogeny showing HGT from viruses. For both phylogenies, numbers along select branches indicate ultrafast bootstrap support values for the descendant nodes. Support for all branches and expanded phylogenies are available in Fig. S11.