

1 LETTER

2 EVIDENCE THAT INCONSISTENT GENE PREDICTION CAN MISLEAD  
3 ANALYSIS OF ALGAL GENOMES<sup>1</sup>

4 Yibi Chen

5 Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia

6 School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane, QLD  
7 4072, Australia

8 Raúl A. González-Pech

9 Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia

10 Timothy G. Stephens

11 Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072,  
12 Australia

13 Debashish Bhattacharya

14 Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ 08901,  
15 USA

16 Cheong Xin Chan<sup>2</sup>

17 Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia

18 School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane,  
19 Queensland 4072, Australia

20 Running head: Methodological biases in predicting algal genes

21 <sup>1</sup>Received XXXXXX. Accepted XXXXXX

<sup>2</sup>Author for correspondence: e-mail [c.chan1@uq.edu.au](mailto:c.chan1@uq.edu.au), phone number +61-7-33462617, fax number +61-7-33462101.

22 **Abstract**

23 Comparative algal genomics often relies on predicted gene models from *de novo* assembled  
24 genomes. However, the artifacts introduced by different gene-prediction approaches, and their  
25 impact on comparative genomic analysis, remain poorly understood. Here, using available  
26 genome data from six dinoflagellate species in Symbiodiniaceae, we identified potential  
27 methodological biases in the published gene models that were predicted using different  
28 approaches. We developed and applied a comprehensive customized workflow to predict genes  
29 from these genomes. The observed variation among predicted gene models resulting from our  
30 workflow agreed with current understanding of phylogenetic relationships among these taxa,  
31 whereas those published earlier were largely biased by the distinct approaches used in each  
32 instance. Importantly, these biases mislead the inference of homologous gene families and  
33 synteny among genomes, thus impacting biological interpretation of these data. Our results  
34 demonstrate that a consistent gene-prediction approach is critical for comparative genomics,  
35 particularly for non-model algal genomes.

36 We implemented a customized, comprehensive workflow to predict protein-coding genes in six  
37 published draft Symbiodiniaceae genomes: *Breviolum minutum* (Shoguchi et al. 2013),  
38 *Symbiodinium tridacnidorum*, *Cladocopium* C92 (Shoguchi et al. 2018), *Symbiodinium*  
39 *microadriaticum* (Aranda et al. 2016), *Cladocopium goreau* and *Fugacium kawagutii* (Liu et al.  
40 2018). These draft genomes, generated largely using short-read sequence data, remain  
41 fragmented (e.g. N50 lengths range from 98.0 Kb for *C. goreau* to 573.5 Kb for *S.*  
42 *microadriaticum*); we treated these genome assemblies independently as is standard practice. The  
43 published gene models from these four studies were predicted using three different approaches:  
44 (a) *ab initio* using AUGUSTUS (Stanke et al. 2006) guided by transcriptome data (Shoguchi et

45 al. 2013, Shoguchi et al. 2018), (b) *ab initio* using AUGUSTUS guided by a more-stringent  
46 selection of genes (Aranda et al. 2016), and (c) a more-thorough approach incorporating evidence  
47 from transcriptomes, machine learning tools, homology to known sequences and *ab initio*  
48 methods (Liu et al. 2018). Because repetitive regions are commonly removed prior to gene  
49 prediction, multi-copy genes are sometimes mis-identified as repeats and excluded from the final  
50 gene models. To address this issue, we adapted the workflow from Liu et al. (2018) to ignore  
51 inferred repeats in the final step that integrates multiple evidence sources using  
52 EVidenceModeler (Haas et al. 2008). To minimize potential contaminants in the published draft  
53 genomes and their impact on gene prediction, we identified and removed genome scaffolds that  
54 share high similarity (BLASTn,  $E \leq 10^{-20}$ , bit-score  $\geq 1000$ , query cover  $\geq 5\%$ ) to bacterial,  
55 archaeal and viral genome sequences in the RefSeq database (release 88), adopting a similar  
56 approach to Liu et al. (2018). We then compared, for each genome, the published gene models in  
57 the remaining scaffolds against the predicted gene models in these same scaffolds using our  
58 approach. Specifically, we assessed metrics of gene models, and the inference of homologous  
59 gene families and conserved synteny within a phylogenetic context.

60 For simplicity, hereinafter we refer to the published gene models as  $\alpha$  genes, and those predicted  
61 in this study as  $\beta$  genes. Compared to  $\alpha$  genes, the structure of  $\beta$  genes (based on the distribution  
62 of intron lengths) resembles more closely the structure of dinoflagellate genes inferred using  
63 transcriptome data (Figure S1). These results suggest that  $\beta$  genes are likely more biologically  
64 realistic. Variation between  $\alpha$  and  $\beta$  genes was assessed using ten metrics: number of predicted  
65 genes per genome, average gene length, number of exons per genome, average exon length,  
66 number of introns per genome, average intron length, proportion of splice-donor site motifs (GT,  
67 GC or GA), number of intergenic regions, and average length of intergenic regions.

68 As shown in Table S1, the metrics for  $\alpha$  and  $\beta$  genes differed substantially. The number of  $\alpha$   
69 genes per genome was much higher in some cases and showed greater variation (mean 48,050;  
70 standard deviation 16,741) than that of  $\beta$  genes (mean 32,819; standard deviation 7567). This is  
71 likely due to the more-stringent criteria used by our workflow to delineate protein-coding genes.  
72 The larger variation in the number of  $\alpha$  genes is likely due to biases arising from the distinct  
73 prediction methods and not assembly artifacts, because the same genome assembly for each  
74 species was used to independently derive  $\alpha$  and  $\beta$  genes. Most predicted genes (>60% genes in  
75 each genome) were supported by transcriptome evidence (BLASTn,  $E \leq 10^{-10}$ ). In some cases,  $\beta$   
76 genes have stronger transcriptome support than  $\alpha$  genes; e.g. 82.6% compared to 66.9% in *S.*  
77 *tridacnidorum*, and 78.4% compared to 61.9% in *Cladocopium* C92 (Table S1).

78 Variation in the ten observed metrics among  $\alpha$  and  $\beta$  genes was also assessed using PCA (Fig.  
79 1a). The  $\alpha$  genes are more widespread along principal component 1 (PC1, between -0.54 and  
80 0.46), with those based on AUGUSTUS-predominant workflows distinctly separated (PC1 <  
81 -0.19; Fig. 1a). The  $\beta$  genes are distributed more narrowly on PC1 (between 0 and 0.27) and  
82 more widely along principal component 2 (PC2; between -0.55 and 0.20). Interestingly, the  
83 distribution of genes along PC2 exhibits a pattern that is consistent with our current  
84 understanding of the phylogeny of these six species (Fig. 1b). Specifically, the *Symbiodinium*  
85 species are clearly separated from the others along PC2 (Fig. 1a) and the two *Cladocopium*  
86 species are clustered more closely based on  $\beta$ , rather than  $\alpha$  genes. Therefore, PC1 (explaining  
87 51.46% of the variance) largely reflects the variation introduced by distinct gene prediction  
88 methods, whereas the distribution along PC2 (explaining 25.91% of the variance) is likely  
89 attributable to the phylogeny of these species. This result suggests that variation among  $\alpha$  genes is  
90 predominantly due to methodological biases, and that these biases are larger compared to those of

91  $\beta$  genes. Variation in the latter appears to be more biologically relevant and consistent with  
92 Symbiodiniaceae evolution.

93 Genomes that are phylogenetically closely related are expected to share greater synteny than  
94 those that are more distantly related. Here, we defined a collinear syntenic gene block as a region  
95 common to two genomes in which five or more genes are coded in the same order and  
96 orientation. These gene blocks were identified using SynChro (Drillon et al. 2014) at  $\Delta = 4$ .  
97 Overall, 421 collinear syntenic blocks (implicating 2454 genes) between any genome-pairs were  
98 identified among  $\alpha$  genes, compared to 450 blocks (implicating 2728 genes) among  $\beta$  genes  
99 (Figs. 2a and 2b). Based on the  $\alpha$  genes comparison (Fig. 2a), *S. microadriaticum* and *S.*  
100 *tridacnidorum* shared the largest number of syntenic blocks (130; 760 genes), whereas *S.*  
101 *microadriaticum* and *F. kawagutii* shared the fewest (1; 6 genes). Surprisingly, *S. tridacnidorum*  
102 and *Cladocopium* C92 shared 38 blocks (222 genes). This close relationship is not evident  
103 between any other pair of genomes from these two genera (e.g. only 3 blocks implicating 15  
104 genes between *S. microadriaticum* and *C. goreau*), and is even closer than the relationship  
105 between the two *Cladocopium* species (i.e. *C. goreau* and C92: 33 blocks, 187 genes). In an  
106 independent analysis, the unexpectedly high conserved synteny between *S. tridacnidorum* and  
107 *Cladocopium* C92 was attributed to inflated evidence support from isoforms of similar  $\alpha$  genes  
108 (as predicted by AUGUSTUS), and the structural configuration (i.e. combination of exons)  
109 among  $\alpha$  genes that is distinct from that among  $\beta$  genes. This observation may be explained by  
110 the fact that  $\alpha$  genes from these two genomes were predicted using the same method (Shoguchi et  
111 al. 2018). In contrast, based on the  $\beta$  genes comparison (Fig. 2b), the number of syntenic blocks  
112 shared between any *Symbiodinium* and *Cladocopium* species did not vary to the same extent; e.g.  
113 7 blocks (38 genes) between *S. tridacnidorum* and *Cladocopium* C92, and 10 blocks (55 genes)

114 between *S. microadriaticum* and *C. goreau*. The number of  $\beta$  genes implicated in blocks shared  
115 by these two genera is also smaller than those between the two *Cladocopium* species (263 genes  
116 in 48 blocks), consistent with their closer phylogenetic relationship.

117 To assess the impact of methodological biases on the delineation of homologous gene families,  
118 Orthofinder v2.3.1 (Emms & Kelly 2018) was used to infer “orthogroups” from protein  
119 sequences (i.e. homologous protein sets) encoded by the  $\alpha$  and  $\beta$  genes (Figs 2c and 2d). More  
120 homologous sets were inferred among the  $\alpha$  genes (33,580) than among the  $\beta$  genes (26,924),  
121 likely due to the higher number of  $\alpha$  genes in all genomes. Genomes from closely related taxa are  
122 expected to share more homologous sequences (and therefore more sets) than those that are  
123 phylogenetically distant. Most of the identified homologous sets (6431 from  $\alpha$  genes, 5217 from  
124  $\beta$  genes) contained sequences from all analyzed taxa; these represent core gene families of  
125 Symbiodiniaceae. Similar to the results of the synteny analysis described above, the pattern of  
126 homologous sets shared between members from *Symbiodinium* and *Cladocopium* varies among  
127 the  $\alpha$  genes (Fig. 2c). For instance, 638 homologous sets are shared only between *S.*  
128 *tridacnidorum* and *Cladocopium* C92, compared to 89 between *C. goreau* and *S. tridacnidorum*.  
129 In contrast, the corresponding number of homologous sets inferred based on  $\beta$  genes are closer to  
130 each other (Fig. 2d); i.e. 92 between *S. tridacnidorum* and *Cladocopium* C92, and 123 between  
131 *C. goreau* and *S. tridacnidorum*.

132 Our results indicate that comparative genomics using the  $\alpha$  genes (i.e. simply based on published  
133 gene models) could lead to the inference that *S. tridacnidorum* and *Cladocopium* C92 are more  
134 closely related with each other than is each of them with other isolates in their corresponding  
135 genus. The bias introduced by different gene-prediction approaches can significantly impact

136 downstream comparative genomic analyses and lead to incorrect biological interpretations. We  
137 therefore urge the research community to consider a consistent gene-prediction workflow when  
138 pursuing comparative genomics, particularly among highly divergent, non-model algal genomes.  
139 Although we only considered dinoflagellate genomes from a single family in this study, the  
140 implication of our results can be applied more broadly to all other non-model eukaryote genomes.

#### 141 **Acknowledgements**

142 RAGP is supported by an International Postgraduate Research Scholarship and a University of  
143 Queensland Centenary Scholarship. TGS was supported by an Australian Government Research  
144 Training Program Scholarship. This work was supported by two Australian Research Council  
145 grants (DP150101875 awarded to Mark Ragan, CXC and DB, and DP190102474 awarded to  
146 CXC and DB), and the computational resources of the Australian National Computational  
147 Infrastructure (NCI) Facility through the NCI Merit Allocation Scheme (project d85) awarded to  
148 CXC.

#### 149 **Competing interests**

150 The authors declare no competing interests.

#### 151 **Data accessibility**

152 All genome data (after removal of microbial contaminants), and all predicted gene models from  
153 this study are available at: <https://cloudstor.aarnet.edu.au/plus/s/JXALPndBKLNYgF9>

#### 154 **Author contribution**

155 YC, RAGP and CXC conceived the study and designed the experiments. YC conducted all  
156 computational analyses. All authors analyzed and interpreted the results. YC and RAGP prepared

157 all figures, tables, and the first draft of this manuscript. YC, TGS and RAGP provided analytical  
158 tools and scripts. All authors wrote, reviewed, commented on and approved the final manuscript.

### 159 **Competing interests**

160 The authors declare no competing interests.

### 161 **References**

- 162 Aranda, M., Li, Y., Liew, Y. J., Baumgarten, S., Simakov, O., Wilson, M. C., Piel, J., Ashoor, H.,  
163 Bougouffa, S., Bajic, V. B., Ryu, T., Ravasi, T., Bayer, T., Micklem, G., Kim, H., Bhak, J.,  
164 LaJeunesse, T. C. & Voolstra, C. R. 2016. Genomes of coral dinoflagellate symbionts  
165 highlight evolutionary adaptations conducive to a symbiotic lifestyle. *Sci Rep* **6**:39734.
- 166 Drillon, G., Carbone, A. & Fischer, G. 2014. SynChro: a fast and easy tool to reconstruct and  
167 visualize synteny blocks along eukaryotic chromosomes. *PLoS ONE* **9**:e92621.
- 168 Emms, D. M. & Kelly, S. 2018. OrthoFinder2: fast and accurate phylogenomic orthology  
169 analysis from gene sequences. *bioRxiv*:466201v1.
- 170 Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R. &  
171 Wortman, J. R. 2008. Automated eukaryotic gene structure annotation using  
172 EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**:R7.
- 173 LaJeunesse, T. C., Parkinson, J. E., Gabrielson, P. W., Jeong, H. J., Reimer, J. D., Voolstra, C. R.  
174 & Santos, S. R. 2018. Systematic revision of Symbiodiniaceae highlights the antiquity and  
175 diversity of coral endosymbionts. *Curr Biol* **28**:2570-80.

- 176 Liu, H., Stephens, T. G., Gonzalez-Pech, R. A., Beltran, V. H., Lapeyre, B., Bongaerts, P.,  
177 Cooke, I., Aranda, M., Bourne, D. G., Foret, S., Miller, D. J., van Oppen, M. J. H.,  
178 Voolstra, C. R., Ragan, M. A. & Chan, C. X. 2018. *Symbiodinium* genomes reveal adaptive  
179 evolution of functions related to coral-dinoflagellate symbiosis. *Commun Biol* **1**:95.
- 180 Shoguchi, E., Beedessee, G., Tada, I., Hisata, K., Kawashima, T., Takeuchi, T., Arakaki, N.,  
181 Fujie, M., Koyanagi, R., Roy, M. C., Kawachi, M., Hidaka, M., Satoh, N. & Shinzato, C.  
182 2018. Two divergent *Symbiodinium* genomes reveal conservation of a gene cluster for  
183 sunscreen biosynthesis and recently lost genes. *BMC Genomics* **19**:458.
- 184 Shoguchi, E., Shinzato, C., Kawashima, T., Gyoja, F., Mungpakdee, S., Koyanagi, R., Takeuchi,  
185 T., Hisata, K., Tanaka, M., Fujiwara, M., Hamada, M., Seidi, A., Fujie, M., Usami, T.,  
186 Goto, H., Yamasaki, S., Arakaki, N., Suzuki, Y., Sugano, S., Toyoda, A., Kuroki, Y.,  
187 Fujiyama, A., Medina, M., Coffroth, M. A., Bhattacharya, D. & Satoh, N. 2013. Draft  
188 assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene  
189 structure. *Curr Biol* **23**:1399-408.
- 190 Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. & Morgenstern, B. 2006. AUGUSTUS:  
191 *ab initio* prediction of alternative transcripts. *Nucleic Acids Res* **34**:W435-9.
- 192
- 193

194 **Figure legends**

195 **Fig. 1. Variation among  $\alpha$  and  $\beta$  genes from six Symbiodiniaceae genomes.** (a) PCA plot  
196 based on ten metrics of the predicted gene models, shown for the  $\alpha$  genes in orange, and the  $\beta$   
197 genes in purple, for each of the six genomes (noted in different symbols) as indicated in the  
198 legend. The two *Cladocopium* and the two *Symbiodinium* species were highlighted for clarity. (b)  
199 Tree topology depicting the phylogenetic relationship among the six taxa, based on LaJeunesse et  
200 al. (2018).

201 **Fig. 2. Conserved synteny and homologous sets among six Symbiodiniaceae genomes.** The  
202 number of collinear syntenic gene blocks between each genome-pair is shown for those inferred  
203 based on (a)  $\alpha$  and (b)  $\beta$  genes; the upper bar chart shows the number of blocks, the lower bar  
204 chart shows the number of implicated genes in these blocks, and the middle panel shows the  
205 genome-pairs corresponding to each bar with a line joining the dots that represent the implicated  
206 taxa. The number of homologous sets inferred from (c)  $\alpha$  and (d)  $\beta$  genes is shown, in which the  
207 taxa represented in the set corresponding to each bar are indicated in the bottom panel. The most  
208 remarkable differences between (a) and (b), and (c) and (d), focusing on *Symbiodinium* and  
209 *Cladocopium* species, are highlighted in red.

210

211 **Supplementary Information**

212 **Fig. S1. Distribution of intron lengths in predicted genes from six Symbiodiniaceae**

213 **genomes.** In each graph, the distribution of intron lengths among  $\alpha$  genes (orange line), among  $\beta$   
214 genes (purple line), and among transcript-based gene models (predicted using PASA v2.3.3 and  
215 TransDecoder v5.2.0; red dashed line) are shown. The transcript-based gene models were  
216 considered as a proxy for true gene structure.

217 **Table S1. Metrics of predicted gene models in genomes of Symbiodiniaceae.**

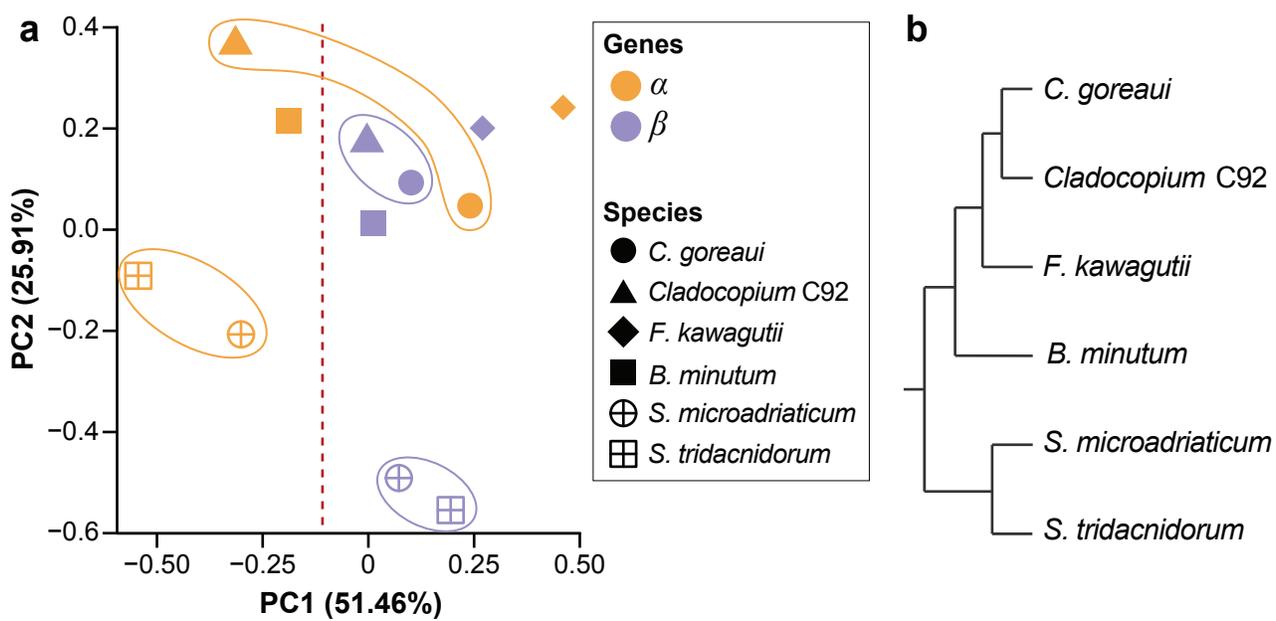


Figure 1

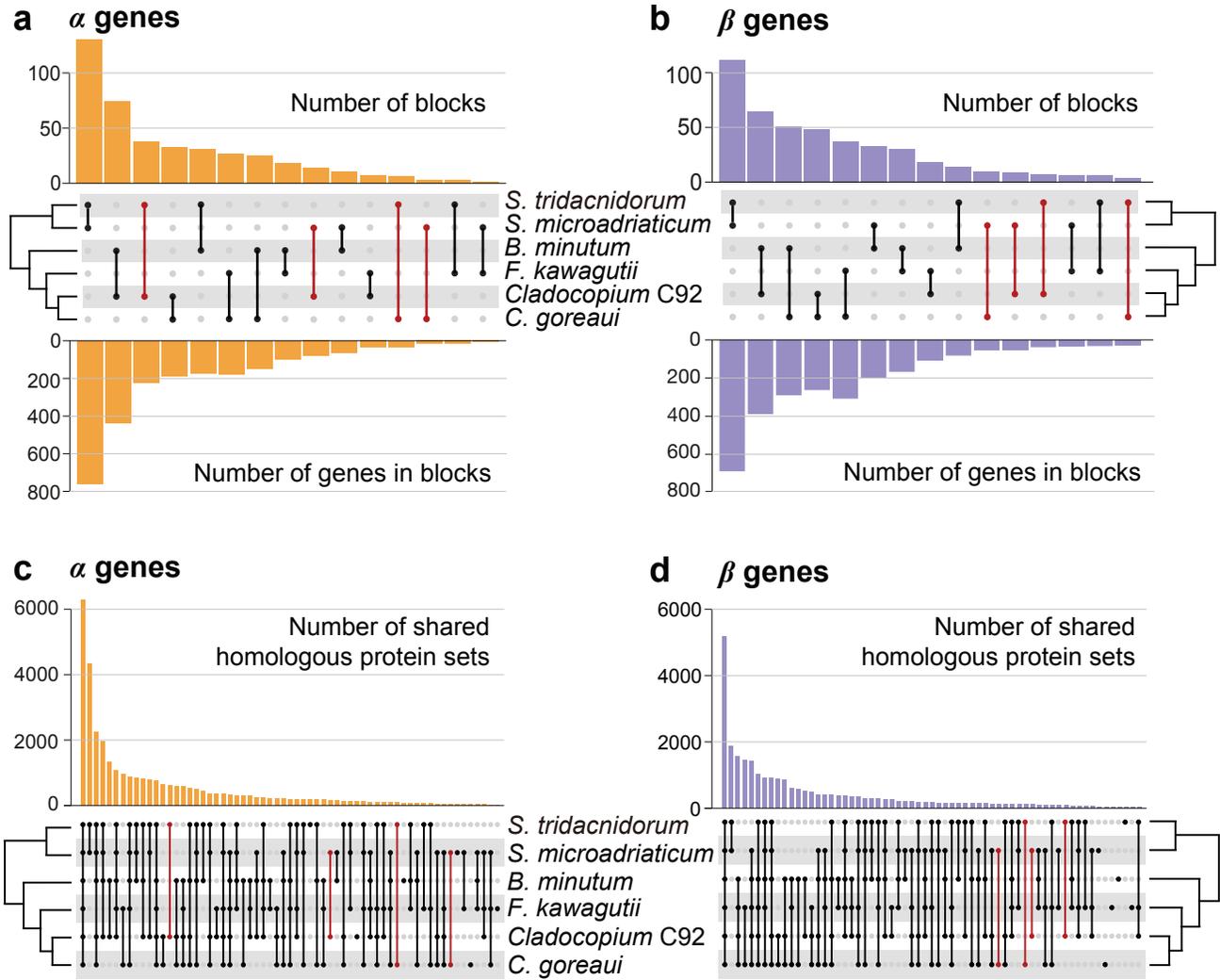


Figure 2