

1 LETTER

2 EVIDENCE THAT INCONSISTENT GENE PREDICTION CAN MISLEAD ANALYSIS  
3 OF ALGAL GENOMES

4

5 Yibi Chen

6 Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072,  
7 Australia; School of Chemistry and Molecular Biosciences, The University of Queensland,  
8 Brisbane, QLD 4072, Australia

9

10 Timothy G. Stephens

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

13

14 Debashish Bhattacharya

15 Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ  
16 08901, USA

17

18 Raúl A. González-Pech

19 Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072,  
20 Australia

21

22 Cheong Xin Chan<sup>1</sup>

23 Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072,  
24 Australia; School of Chemistry and Molecular Biosciences, The University of Queensland,  
25 Brisbane, Queensland 4072, Australia

26

27

<sup>1</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

## 28 **Abstract**

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

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

<sup>1</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

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

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

<sup>1</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

75 As shown in Table S1, the metrics for  $\alpha$  and  $\beta$  genes differed substantially. The number of  $\alpha$   
76 genes per genome was much higher in some cases and showed greater variation (mean  
77 48,050; standard deviation 16,741) than that of  $\beta$  genes (mean 32,819; standard deviation  
78 7567). This is likely due to the more-stringent criteria used by our workflow to delineate  
79 protein-coding genes. The larger variation in the number of  $\alpha$  genes is likely due to biases  
80 arising from the distinct prediction methods and not assembly artifacts, because the same  
81 genome assembly for each species was used to independently derive  $\alpha$  and  $\beta$  genes. In  
82 general, most predicted genes (>60% genes in each genome) were supported by  
83 transcriptome evidence (BLASTn,  $E \leq 10^{-10}$ ). In some cases,  $\beta$  genes have stronger  
84 transcriptome support than  $\alpha$  genes; e.g. 82.6% compared to 66.9% in *S. tridacnidorum*, and  
85 78.4% compared to 61.9% in *Cladocopium* sp. C92 (Table S1).

86 Variation in the ten observed metrics among  $\alpha$  and  $\beta$  genes was also assessed using Principal  
87 Component Analysis (Fig. 1a). The  $\alpha$  genes are more widespread along principal component  
88 1 (PC1, between -0.54 and 0.44), with those based on AUGUSTUS-predominant workflows  
89 distinctly separated (PC1 < -0.18; Fig. 1a). The  $\beta$  genes are distributed more narrowly on  
90 PC1 (between -0.01 and 0.36) and more widely along principal component 2 (PC2; between  
91 -0.56 and 0.20). Interestingly, the distribution of genes along PC2 exhibits a pattern that is  
92 consistent with our current understanding of the phylogeny of these six species (Fig. 1b).  
93 Specifically, the *Symbiodinium* species are clearly separated from the others along PC2 (Fig.  
94 1a) and the two *Cladocopium* species are clustered more closely based on  $\beta$ , rather than  $\alpha$   
95 genes. Therefore, PC1 (explaining 52.47% of the variance) largely reflects the variation  
96 introduced by distinct gene prediction methods, whereas the distribution along PC2  
97 (explaining 25.83% of the variance) is likely attributable to the phylogeny of these species.  
98 This result suggests that variation among  $\alpha$  genes is predominantly due to methodological

<sup>1</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

99 biases, and that these biases are larger compared to those of  $\beta$  genes. Variation in the latter  
100 appears to be more biologically relevant and consistent with Symbiodiniaceae evolution.

101 Genomes that are phylogenetically closely related are expected to share greater synteny than  
102 those that are more distantly related. Here, we defined a collinear syntenic gene block as a  
103 region common to two genomes in which five or more genes are coded in the same order and  
104 orientation. These gene blocks were identified using SynChro (Drillon et al. 2014) at  $\Delta$  =  
105 4. Overall, 421 collinear syntenic blocks (implicating 2454 genes) between any genome-pairs  
106 were identified among  $\alpha$  genes, compared to 426 blocks (implicating 2553 genes) among  $\beta$   
107 genes (Figs. 2a and 2b). Based on  $\alpha$  genes comparison (Fig. 2a), *S. microadriaticum* and *S.*  
108 *tridacnidorum* shared the largest number of syntenic blocks (130; 760 genes), whereas *S.*  
109 *microadriaticum* and *F. kawagutii* shared the fewest (1; 6 genes). Surprisingly, *S.*  
110 *tridacnidorum* and *Cladocopium* sp. C92 shared 38 blocks (222 genes). This close  
111 relationship is not evident between any other pair of genomes from these two genera (e.g.  
112 only 3 blocks implicating 15 genes between *S. microadriaticum* and *C. goreau*), and is even  
113 closer than the relationship between the two *Cladocopium* species (i.e. *C. goreau* and C92:  
114 33 blocks, 187 genes). The unexpectedly high conserved synteny between *S. tridacnidorum*  
115 and *Cladocopium* sp. C92 may be explained by the fact that these genes were predicted with  
116 the same method (Shoguchi et al. 2018). In contrast, based on the  $\beta$  genes comparison (Fig.  
117 2b), the number of syntenic blocks shared between any *Symbiodinium* and *Cladocopium*  
118 species did not vary to the same extent; e.g. 7 blocks (38 genes) between *S. tridacnidorum*  
119 and *Cladocopium* sp. C92, and 10 blocks (55 genes) between *S. microadriaticum* and *C.*  
120 *goreau*. The number of  $\beta$  genes implicated in blocks shared by these two genera is also  
121 smaller than those between the two *Cladocopium* species (263 genes in 48 blocks), consistent  
122 with their closer phylogenetic relationship.

<sup>1</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

123 To assess the impact of methodological biases on the delineation of homologous gene  
124 families, Orthofinder v2.3.1 (Emms and Kelly 2018) was used to infer “orthogroups” from  
125 protein sequences (i.e. homologous protein sets) encoded by the  $\alpha$  and  $\beta$  genes (Figs 2c and  
126 2d). More homologous sets were inferred among the  $\alpha$  genes (33,580) than among the  $\beta$   
127 genes (26,380), likely due to the higher number of  $\alpha$  genes in all genomes. Genomes from  
128 closely related taxa are expected to share more homologous sequences (and therefore more  
129 sets) than those that are phylogenetically distant. Most of the identified homologous sets  
130 (6431 from  $\alpha$  genes, 4941 from  $\beta$  genes) contained sequences from all analyzed taxa; these  
131 represent core gene families of Symbiodiniaceae. Similar to the results of the synteny  
132 analysis described above, the pattern of homologous sets shared between members from  
133 *Symbiodinium* and *Cladocopium* varies among the  $\alpha$  genes (Fig. 2c). For instance, 638  
134 homologous sets are shared only between *S. tridacnidorum* and *Cladocopium* sp. C92,  
135 compared to 89 between *C. goreau* and *S. tridacnidorum*. In contrast, the corresponding  
136 number of homologous sets inferred based on  $\beta$  genes are closer to each other (Fig. 2d); i.e.  
137 100 between *S. tridacnidorum* and *Cladocopium* sp. C92, and 132 between *C. goreau* and *S.*  
138 *tridacnidorum*.

139 Our results indicate that comparative genomics using the  $\alpha$  genes (i.e. simply based on  
140 published gene models) could lead to the inference that *S. tridacnidorum* and *Cladocopium*  
141 sp. C92 are more closely related with each other than is each of them with other isolates in  
142 their corresponding genus. The bias introduced by different gene-prediction approaches can  
143 significantly impact downstream comparative genomic analyses and lead to incorrect  
144 biological interpretations. We therefore urge the research community to consider a consistent  
145 gene-prediction workflow when pursuing comparative genomics, particularly among highly  
146 divergent, non-model algal genomes. Although we only considered dinoflagellate genomes

<sup>1</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

147 from a single family in this study, the implication of our results can be applied more broadly  
148 to all other non-model eukaryote genomes.

## 149 **Acknowledgements**

150 TGS is supported by an Australian Government Research Training Program Scholarship.  
151 RAGP is supported by an International Postgraduate Research Scholarship and a University  
152 of Queensland Centenary Scholarship. This work was supported by Australian Research  
153 Council grants (DP150101875 and DP190102474) awarded to CXC and DB, and the  
154 computational resources of the Australian National Computational Infrastructure (NCI)  
155 Facility through the NCI Merit Allocation Scheme (project d85) awarded to CXC.

## 156 **References**

- 157 Aranda, M., Li, Y., Liew, Y. J., Baumgarten, S., Simakov, O., Wilson, M. C., Piel, J.,  
158 Ashoor, H., Bougouffa, S., Bajic, V. B., Ryu, T., Ravasi, T., Bayer, T., Micklem, G.,  
159 Kim, H., Bhak, J., LaJeunesse, T. C. & Voolstra, C. R. 2016. Genomes of coral  
160 dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic  
161 lifestyle. *Scientific Reports* **6**:39734.
- 162 Drillon, G., Carbone, A. & Fischer, G. 2014. SynChro: a fast and easy tool to reconstruct and  
163 visualize synteny blocks along eukaryotic chromosomes. *PLoS ONE* **9**:e92621.
- 164 Emms, D. M. & Kelly, S. 2018. OrthoFinder2: fast and accurate phylogenomic orthology  
165 analysis from gene sequences. *bioRxiv*:466201.
- 166 Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C.  
167 R. & Wortman, J. R. 2008. Automated eukaryotic gene structure annotation using  
168 EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biol.*  
169 **9**:1.

<sup>1</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

- 170 LaJeunesse, T. C., Parkinson, J. E., Gabrielson, P. W., Jeong, H. J., Reimer, J. D., Voolstra,  
171 C. R. & Santos, S. R. 2018. Systematic revision of Symbiodiniaceae highlights the  
172 antiquity and diversity of coral endosymbionts. *Curr. Biol.* **28**:2570-80.
- 173 Liu, H., Stephens, T. G., González-Pech, R. A., Beltran, V. H., Lapeyre, B., Bongaerts, P.,  
174 Cooke, I., Aranda, M., Bourne, D. G., Forêt, S., Miller, D. J., van Oppen, M. J. H.,  
175 Voolstra, C. R., Ragan, M. A. & Chan, C. X. 2018. *Symbiodinium* genomes reveal  
176 adaptive evolution of functions related to coral-dinoflagellate symbiosis. *Commun.*  
177 *Biol.* **1**:95.
- 178 Shoguchi, E., Beedessee, G., Tada, I., Hisata, K., Kawashima, T., Takeuchi, T., Arakaki, N.,  
179 Fujie, M., Koyanagi, R., Roy, M. C., Kawachi, M., Hidaka, M., Satoh, N. & Shinzato,  
180 C. 2018. Two divergent *Symbiodinium* genomes reveal conservation of a gene cluster  
181 for sunscreen biosynthesis and recently lost genes. *BMC Genomics* **19**:458.
- 182 Shoguchi, E., Shinzato, C., Kawashima, T., Gyoja, F., Mungpakdee, S., Koyanagi, R.,  
183 Takeuchi, T., Hisata, K., Tanaka, M. & Fujiwara, M. 2013. Draft assembly of the  
184 *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene structure. *Curr.*  
185 *Biol.* **23**:1399-408.
- 186 Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. & Morgenstern, B. 2006.  
187 AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res.*  
188 **34**:W435-W39.

## 189 **Data accessibility**

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

192

<sup>1</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



193 **Author contribution**

194 YC, RAGP and CXC conceived the study and designed the experiments. YC conducted all  
195 computational analyses. YC, TGS, RAGP and CXC analyzed and interpreted the results. YC  
196 and RAGP prepared all figures, tables, and the first draft of this manuscript. YC, TGS and  
197 RAGP provided analytical tools and scripts. All authors wrote, reviewed, commented on and  
198 approved the final manuscript.

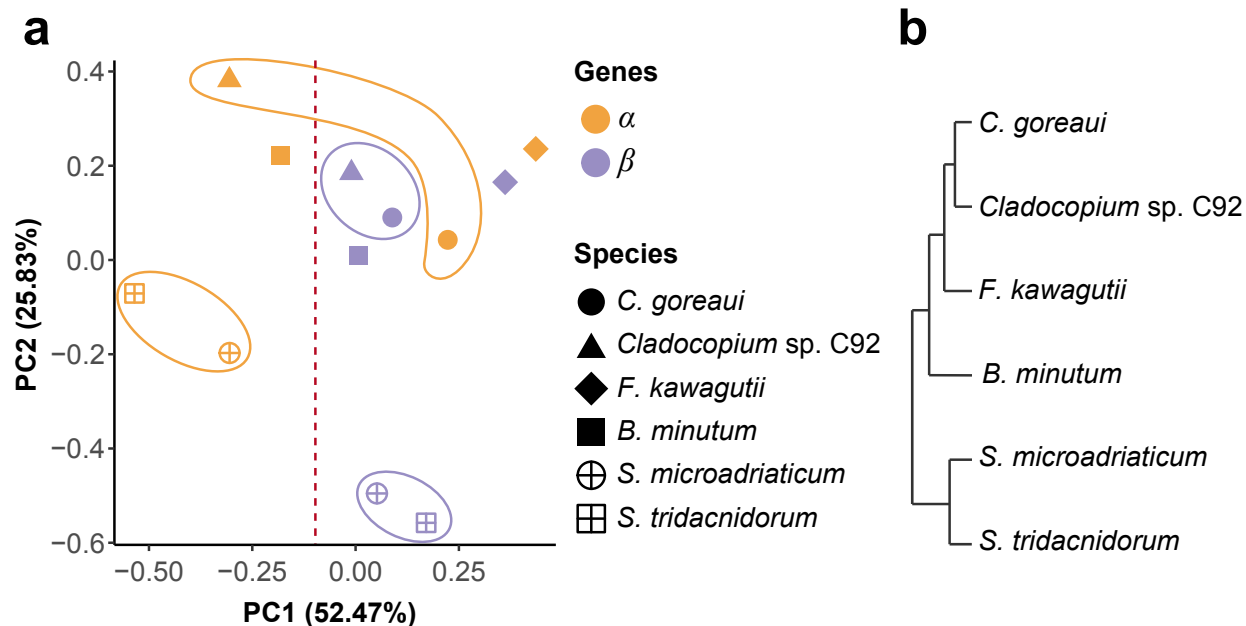
199 **Competing interests**

200 The authors declare no competing interests.

201

<sup>1</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

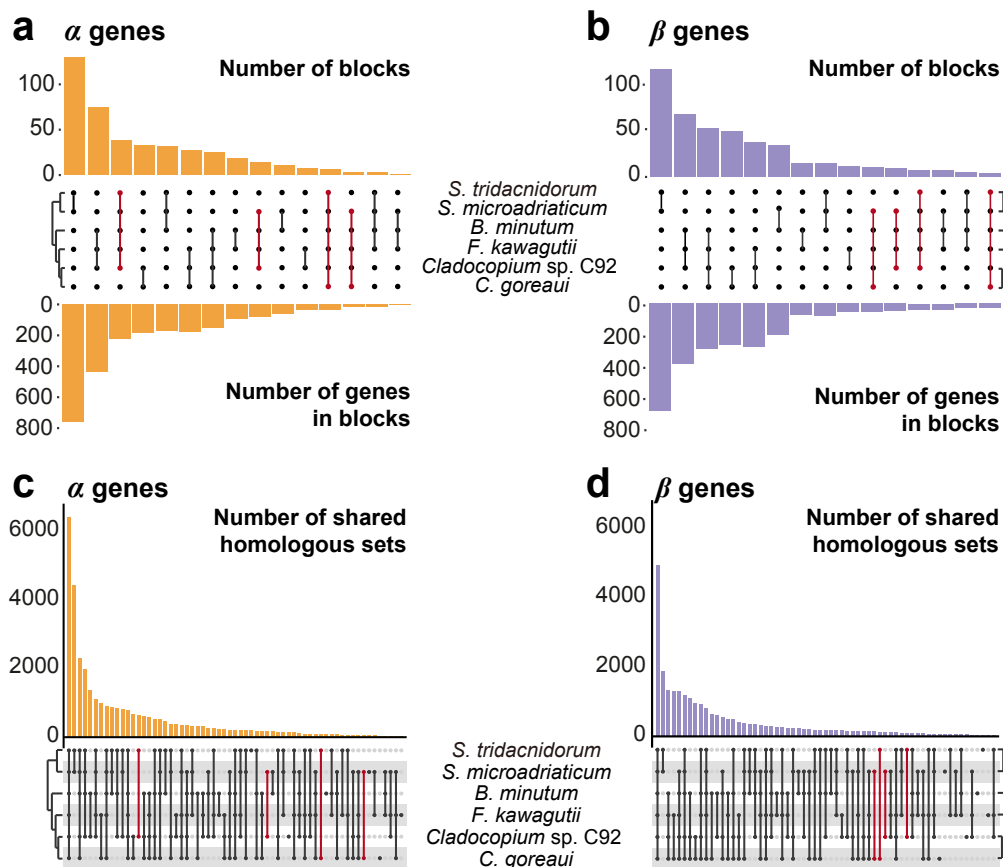
202 **Figures**



203

204 **Fig. 1. Gene metrics of  $\alpha$  and  $\beta$  genes from six Symbiodiniaceae genomes.** (a) Principal  
205 Component Analysis plot based on ten metrics of the predicted gene models, shown for the  $\alpha$   
206 genes in orange, and the  $\beta$  genes in purple, for each of the six genomes (noted in different  
207 symbols) as indicated in the legend. The two *Cladocopium* species and the two *Symbiodinium*  
208 species were highlighted for clarity. (b) A tree topology depicting the phylogenetic  
209 relationship among the six taxa, based on (LaJeunesse et al. 2018).

210



211

212 **Fig. 2. Conserved synteny and homologous sets among six Symbiodiniaceae genomes.**

213 The number of collinear syntenic gene blocks between each genome-pair is shown for those  
214 inferred based on (a)  $\alpha$  and (b)  $\beta$  genes; the upper bar chart shows the number of blocks, the  
215 lower bar chart shows the number of implicated genes in these blocks, and the middle panel  
216 shows the genome-pairs corresponding to each bar with a line joining the dots that represent  
217 the implicated taxa. The number of homologous sets inferred from (c)  $\alpha$  and (d)  $\beta$  genes is  
218 shown, in which the taxa represented in the set corresponding to each bar are indicated in the  
219 bottom panel. The most remarkable differences between (a) and (b), and (c) and (d), focusing  
220 on *Symbiodinium* and *Cladocopium* species are highlighted in red.

<sup>1</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