



**HAL**  
open science

## Use of RAD sequencing for delimiting species

Eric Pante, Jawad Abdelkrim, Amélia Viricel, Delphine Gey, Scott France,  
Marie-Catherine Boisselier, Sarah Samadi

► **To cite this version:**

Eric Pante, Jawad Abdelkrim, Amélia Viricel, Delphine Gey, Scott France, et al.. Use of RAD sequencing for delimiting species. *Heredity*, 2015, 114, pp.450-459. 10.1038/hdy.2014.105 . hal-01064545

**HAL Id: hal-01064545**

**<https://hal.science/hal-01064545>**

Submitted on 17 Sep 2014

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

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

1 Heredity – Original Article

2

3 **Use of RAD sequencing for delimiting species**

4

5 Pante E<sup>1</sup>, Abdelkrim J<sup>2\*</sup>, Viricel A<sup>1\*</sup>, Gey D<sup>2</sup>, France S<sup>3</sup>, Boisselier MC<sup>2,4</sup>, Samadi S<sup>4</sup>

6

7 <sup>1</sup> Laboratoire LIENSs, UMR 7266 CNRS - Université de La Rochelle, La Rochelle, France

8 <sup>2</sup> Département Systématique et Evolution, UMS 2700 MNHN-CNRS, SSM, Muséum  
9 national d'Histoire naturelle, Paris, France

10 <sup>3</sup> Department of Biology, University of Louisiana at Lafayette, Lafayette LA, USA

11 <sup>4</sup> ISYEB – UMR 7205 – CNRS, MNHN, UPMC, EPHE, Muséum national d'Histoire  
12 naturelle, Sorbonne Universités, Paris

13 \* authors contributed equally to the work

14

15 Corresponding author:

16 Eric Pante, Laboratoire LIENSs, UMR 7266 CNRS - Université de La Rochelle, La

17 Rochelle, France. Tel: +33 05 46 50 76 37; Fax: +33 05 46 50 76 63; Email:

18 [pante.eric@gmail.com](mailto:pante.eric@gmail.com)

19

20 Running title: Use of RAD sequencing for delimiting species

21 Word count: 6 999 words

22

23

24 **Abstract**

25

26 RAD-tag sequencing is a promising method for conducting genome-wide evolutionary  
27 studies. However, to date, only a handful of studies empirically tested its applicability  
28 above the species level. In this communication, we use RAD-tags to contribute to the  
29 delimitation of species within a diverse genus of deep-sea octocorals, *Chrysogorgia*, for  
30 which few classical genetic markers have proved informative. Previous studies have  
31 hypothesized that single mitochondrial haplotypes can be used to delimit  
32 *Chrysogorgia* species. Based on two lanes of Illumina sequencing, we inferred  
33 phylogenetic relationships among twelve putative species that were delimited using  
34 mitochondrial data, comparing two RAD analysis pipelines (Stacks and pyRAD). The  
35 number of homologous RAD loci decreased dramatically with increasing divergence,  
36 as >70% of loci are lost when comparing specimens separated by two mutations on  
37 the 700 nt long mitochondrial phylogeny. Species delimitation hypotheses based on  
38 the mitochondrial *mtMutS* gene are largely supported, as six out of nine putative  
39 species represented by more than one colony were recovered as discrete, well-  
40 supported clades. Significant genetic structure (correlating with geography) was  
41 detected within one putative species, suggesting that individuals characterized by the  
42 same *mtMutS* haplotype may belong to distinct species. Conversely, three *mtMutS*  
43 haplotypes formed one well-supported clade within which no population structure  
44 was detected, also suggesting that intra-specific variation exists at *mtMutS* in  
45 *Chrysogorgia*. Despite an impressive decrease in the number of homologous loci  
46 across clades, RAD data helped us to fine-tune our interpretations of classical  
47 mitochondrial markers used in octocoral species delimitation, and discover previously  
48 undetected diversity.

49

50 **Keywords (3-6):** phylogenomics, octocoral, Stacks, PyRAD, SNP, species delimitation

## 51 **Introduction**

52 The advent of next-generation sequencing tools has permitted significant advances in  
53 our understanding of evolutionary processes such as speciation (e.g. Ekblom and  
54 Galindo 2011), but some other practical applications of genomic data have been less  
55 explored, including phylogenomics and species delimitation. Among genomic  
56 approaches that are applicable to these fields, the usefulness of restriction-site-  
57 associated DNA tag (RAD-tag; Baird *et al.*, 2008) sequencing has been investigated in  
58 few studies to date. This methodology typically provides short sequences (~ 100-150  
59 bp) flanking the cut sites of a restriction enzyme (or several enzymes), generally  
60 yielding thousands of loci distributed throughout the genome. This approach does not  
61 require a reference genome, and can therefore be applied to non-model organisms.  
62 However, some technical difficulties remain for groups where very little genomic  
63 knowledge is available (see Davey *et al.*, 2011). For instance, the choice of restriction  
64 enzyme(s) and methodology (single-digest versus double-digest RAD) is key to  
65 estimating the number of expected cut sites and coverage, but relies on prior  
66 knowledge of genome size and GC content.

67 Despite these difficulties, RAD-tag sequencing constitutes one of the reduced  
68 genomic approaches that are suitable for investigating inter-specific evolutionary  
69 questions. Published RAD-tag sequencing research beyond the species level  
70 includes *in silico* studies (*Drosophila*, mammals, and yeasts in Rubin *et al.*, 2012;  
71 *Drosophila* in Cariou *et al.*, 2013) and empirical work (e.g. Restionaceae flowering  
72 plants in Lexer *et al.*, 2013; cetaceans in Viricel *et al.*, 2014), which both suggest this  
73 approach is promising for taxa having diverged up to 60 million years ago. For

74 instance, RAD-tag sequencing has proven useful in species delimitation and  
75 phylogenies within recently and rapidly diverged groups (e.g. Orobanchaceae  
76 flowering plants in Eaton and Ree 2013; swordtails in Jones *et al.*, 2013; *Heliconius*  
77 butterflies in Nadeau *et al.*, 2013; cichlids in Wagner *et al.*, 2013; geckos in Leaché *et*  
78 *al.*, 2014). Comparatively, reconstructing the phylogeny of more distantly related taxa  
79 has been the topic of a single study (*Carabus* beetles, Cruaud *et al.*, 2014), to the best  
80 of our knowledge. Herein we use this approach on a group of deep-sea octocorals for  
81 which little genomic data are available. Thus, our contribution constitutes one of the  
82 first studies investigating the use of RAD-tag sequencing for practical species  
83 delimitation within a taxonomic group composed of divergent species (up to 16  
84 million years ago).

85           Deep-sea octocorals are one of the groups for which RAD-tag sequencing can  
86 significantly advance our understanding of evolutionary patterns. As for shallow-  
87 water octocorals, deep-water octocorals present significant challenges for  
88 taxonomists, with few morphological characters being available for species  
89 delimitation (e.g., McFadden *et al.*, 2010). In addition, several studies have shown  
90 conflicting patterns of morphological and molecular data (France 2007; Dueñas and  
91 Sánchez 2009; Pante and France 2010), suggesting that an integrative approach to  
92 taxon delimitation must be applied in this group (e.g. Schlick-Steiner *et al.*, 2010).  
93 Octocorals, as with other anthozoans (e.g. scleractinians and sea anemones), are also  
94 plagued with remarkably low levels of mitochondrial genome evolution that renders  
95 the use of classical barcoding gene regions such as *cox1* of limited use (McFadden *et*  
96 *al.*, 2011). Comparatively, a few studies have successfully used nuclear markers within  
97 octocoral species (e.g. Concepcion *et al.*, 2008; Mokhtar-Jamaï *et al.*, 2011), but these  
98 are either not widely useable across octocorals (e.g. SRP54; France and Pante  
99 unpublished observations), or not informative at multiple phylogenetic scales (e.g.

100 microsatellites). Multi-copy markers have been employed (e.g. Herrera *et al.*, 2010),  
101 however their use implies that lack of concerted evolution within and across genomes  
102 will not blur the phylogenetic signal (Vollmer and Palumbi 2004; Calderón et al 2006).  
103 In this group, RAD-tag genotyping may therefore offer a panel of markers to help  
104 describe patterns of population structure, delimit species, and investigate  
105 phylogenetic relationships. This technique may however be difficult to implement in  
106 this group. Indeed, the composition of the deep-sea octocoral genome is unknown  
107 (size, GC content, prevalence of cut sites for restriction enzymes, etc.); the size of  
108 known cnidarian genomes, for instance, varies between 224 Mb and 1.8 Tb (Animal  
109 Genome Size Database; Gregory, 2014). In addition, sampling of deep-sea animals can  
110 be associated with a loss of quality of genomic DNA samples, particularly when  
111 sampling in tropical waters using trawls or dredges.

112           The genus *Chrysogorgia* (Calcaxonia: Chrysogorgiidae) is a noteworthy model  
113 for testing the utility of RAD sequencing for delimiting octocoral species, as it is  
114 diverse (62 nominal species described, 93% of which were based solely on  
115 morphology), widely distributed, and can be locally abundant (Watling *et al.*, 2011).  
116 The large geographic, bathymetric, and ecological distributions of  
117 some *Chrysogorgia* species (Pante *et al.*, 2012b) question whether taxa are  
118 appropriately delimited, and whether cryptic diversity is important in the group. In  
119 the northwestern Atlantic, congruence exists between morphological and genetic data,  
120 suggesting that a relatively short fragment of the mitochondrial *mtMutS* gene can be  
121 used to formulate “Primary Species Hypotheses” (Pante and Watling 2012). It is  
122 suspected that little to no intra-specific variation exists for this marker within the  
123 group (McFadden *et al.*, 2011), but the null hypothesis that single mutations  
124 at *mtMutS* are diagnostic of species limits must be evaluated using genetic data from

125 markers informative within and above the species level. RAD loci allow to test  
126 whether lineages that putatively belong to different species do not exchange genes.

127           In this communication we test the utility of RAD-tag genotyping for delimiting  
128 species in *Chrysogorgia* using the genealogical criterion defined by Taylor *et al.*,  
129 (2001). More specifically, we test whether single mutations on the mitochondrial  
130 *mtMutS* gene can be used as a criterion for grouping *Chrysogorgia* colonies into  
131 separate, putative species (or, more specifically, “Primary Species Delimitation  
132 hypotheses” as in Puillandre *et al.*, 2012). We compare the results from two analysis  
133 pipelines, Stacks (Catchen *et al.*, 2013) and PyRAD (Eaton, 2014), which significantly  
134 differ in the method employed for detecting homologous loci.

135

## 136 **Material and methods**

### 137 *Specimen collection and mtDNA typing*

138 *Chrysogorgia* specimens were collected from the SE slope of New Caledonia (NC) and  
139 adjacent seamounts of the Norfolk Ridge (82 colonies; *Terrasses* cruise, 2008), from  
140 Papua New Guinea (PNG; 8 colonies; *BioPapua* cruise, 2010), and from the  
141 northwestern Atlantic (1 colony, *Extreme Coral 2010* cruise; Tables 1 and S1). Pacific  
142 specimens were retrieved from dredges and trawls (details on cruises of the Tropical  
143 Deep Sea Benthos research program: Bouchet *et al.*, 2008; details on the *BioPapua*  
144 cruise: Pante *et al.*, 2012a); the Atlantic specimen was collected using the Jason II ROV  
145 (Woods Hole Oceanographic Institution). Specimens were fixed in 80% ethanol as  
146 soon as possible after collection. Genomic DNA was extracted using a CTAB protocol  
147 according to France *et al.* (1996). A 700-bp fragment of the mitochondrial *mtMutS*

148 gene (identified as more informative than *cox1* or 18S in chrysogorgiids, Pante *et al.*,  
149 2012b) was amplified using the ND4L2475F – MUT3458R primer pair and sequenced  
150 using an ABI PRISM (R) 3100 or 3130xl Genetic Analyzer (primer information, PCR  
151 and sequencing conditions: Pante *et al.*, 2012b). Sequences were checked for quality  
152 and edited in Sequencher (TM) 4.7 (Gene Codes), aligned by eye (a single, 3 bp indel  
153 was present in the alignment), and haplotypes were submitted to GenBank (Table S1).  
154 Divergence times among putative species were estimated using the molecular clock  
155 from Lepard (2003), which was calculated for the shallow-water octocoral genus  
156 *Lepogorgia* based on *mtMutS* genetic distances for clades located on either sides of the  
157 Isthmus of Panama (0.14–0.25%/million years).

#### 158 *Library construction, RAD sequencing, and quality control*

159 Genomic DNA quality was evaluated by 1% agarose gel electrophoresis, and  
160 quantified using a Thermo Scientific Nanodrop ND-1 000 spectrophotometer. DNA  
161 was sent to Eurofins Genomics (Ebersberg, Germany) for RAD-tag library preparation  
162 and sequencing. Libraries were constructed from 1-2 µg of DNA per colony using the  
163 *SbfI* restriction enzyme. This enzyme was chosen because it was successfully used in  
164 RADseq experiments with marine invertebrates (sea-anemones, Reitzel *et al.*, 2013;  
165 abalone, Gruenthal *et al.*, 2014), and was expected to allow an acceptable compromise  
166 between prevalence of cut sites and depth of coverage, based on RADcounter (the  
167 University of Edinburgh, <https://www.wiki.ed.ac.uk/display/RADSequencing/Home>).  
168 As the genome size and GC content of *Chrysogorgia* (or other octocorals, to the best of  
169 our knowledge) are not known, we estimated the prevalence of *SbfI* cut sites based on  
170 a range of genome sizes and GC content, based on information from the Animal  
171 Genome Size Database (see Introduction) and with a GC content of 40% (e.g. Soza-  
172 Ried *et al.*, 2009). Barcodes 6-9 nucleotides long and differing by at least 2 nucleotides



173 were used to differentiate multiplexed samples (Table S1). Sequencing was performed  
174 on two lanes of the Illumina (R) HiSeq (TM) 2 000 instrument (Illumina Inc., San Diego  
175 CA, USA) using the single read, 100 nucleotide configuration. Raw HiSeq output was  
176 processed using the CASAVA v1.8.2 software pipeline (Illumina Inc., San Diego CA,  
177 USA), and de-multiplexed and quality filtered using the process\_radtags.pl module  
178 (default quality settings) of the Stacks v.0.99994 pipeline (Catchen *et al.*, 2013). A  
179 single sequencing error was tolerated in the barcode. Reads were truncated to 91  
180 nucleotides. Quality (as measured by phred scores and percentage of sequence  
181 overrepresentation) was checked before and after treatment by process\_radtags using  
182 FastQC v.0.10.1 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>).

### 183 *Exploration of the divergence parameter space*

184 Two main pipelines specifically designed for analysis of RADseq data are currently  
185 available. The most used to date is the Stacks pipeline. It constructs a catalog of loci for  
186 a set of samples mainly based on three parameters: the minimum stack depth  
187 parameter  $m$  (i.e. the minimum number of reads allowed per allele), the intra-  
188 individual divergence parameter  $M$  (i.e. the maximum number of mutations that can  
189 be observed between stacks within a sample), and the inter-individual divergence  
190 parameter  $n$  (i.e. the maximum number of mutations that can be observed between  
191 loci across samples).

192 PyRAD (Eaton 2014) is a more recently developed pipeline and differs from  
193 Stacks in several ways, the most important one being that it allows the presence of  
194 indels, since the clustering process of reads into loci uses alignment tools. This is  
195 anticipated to be an advantage compared to the first pipeline when considering more  
196 phylogenetically distant species. PyRAD relies on a large number of parameters used  
197 at different steps of the process. Most of them are related to reads quality control,

198 detection of homology and filtering of paralogs. Two main parameters are of  
199 particular importance: the minimum depth coverage Mindepth (minimum depth  
200 necessary to make a statistical base call at each position of a cluster) and the similarity  
201 threshold Wclust (similarity value to be used for the alignment during both the within  
202 and across-sample clustering).

203           For both pipelines, these parameter settings are expected to influence greatly  
204 the number of markers available for intra- and inter-specific comparisons and it is  
205 necessary to explore which parameter combinations maximize the number of  
206 orthologous loci (Viricel *et al.*, 2014). To explore the effect of these parameters at  
207 different phylogenetic depths, we randomly selected pairs of specimens that (1) were  
208 separated by 0 to 16 mutations at *mtMutS* (representing different levels of  
209 phylogenetic divergence), and (2) were characterized by 1 to 1.5 million reads (to  
210 alleviate potential effects of depth of coverage on the number of assembled loci). For  
211 each level of divergence, we used three replicate pairs of specimens. We refer to  
212 specimens with *mtMutS* haplotypes differing by few mutations as pairs of closely-  
213 related colonies, and those with haplotypes differing by many mutations as distantly-  
214 related colonies.

215           In Stacks, *m* was kept to 3 (the default value); *M* was incremented from 1 to  
216 10 in two cases (specimens separated by 0 and 12 mutations at *mtMutS*), and from 1  
217 to 7 in all other cases. Similarly, *n* was incremented from 1 to 10 (0 and 12 mutations  
218 cases) and from 1 to 8 (all cases). All combinations of *M* and *n* were not tested: only  
219 similar values of *M* and *n* were used together (two settings were used:  $M=n$  and  
220  $M+1=n$ ), as to (1) keep maximum levels of intra- and inter-individual divergence  
221 levels close, and (2) keep the number of Stacks analyses to a reasonable number. A  
222 total of 408 Stacks catalog construction tests were therefore performed using the

223 denovo\_map.pl script available in Stacks. Catalogs were parsed with the populations.pl  
224 script, where each sample was considered as a separate population, no missing data  
225 were allowed, and a minimum of 10 reads per SNP was set.

226 In PyRAD v. 2.0, combinations of two values for Mindepth (3 and 6) and 3  
227 values for Wclust (0.89, 0.93 and 0.96) were tested, resulting in 156 analyses. For  
228 these analyses, the maximum number of sites per read with a quality < 20 (NQual)  
229 was set to 4, the minimum number of samples in a final locus (MinCov) was set to 1  
230 and the maximum proportion of shared polymorphic sites in a locus (MaxSH) was set  
231 to 10%. For this last parameter, which aims at detecting paralogs, preliminary tests  
232 showed that in our case, changing this value did not drastically affect the number of  
233 loci and SNPs detected. Finally, optional parameters were kept to default values.

#### 234 *Comparison of Stacks and PyRAD*

235 To evaluate what proportion of loci was detected by both PyRAD and Stacks, a custom  
236 BLASTN search was performed (BLAST toolkit v. 2.2.25; Zhang *et al.*, 2000). Local  
237 BLAST databases were constructed using PyRAD sequences (locus file containing  
238 consensus sequences for each individual; PyRAD parameters m=6 and Wclust=93%  
239 and 89%) for three groups of specimens with different numbers of reads (Table 2).  
240 Stacks loci for these specimens (based on the locus file produced by the populations  
241 script, for which a single allele was retained per locus; denovo\_map parameters m=3,  
242 M=4, n=4, and m=3, M=10, n=12) were then compared to the PyRAD database using  
243 BLASTN (percent identity set to 93% and 89%, word size 80 and 84 nt, ungapped  
244 alignments). The XML output of BLASTN searches was then parsed in bash using grep.

#### 245 *Phylogenetic reconstruction and species delimitation*

246 RAxML v. 8.0.9 (Stamatakis 2006; Stamatakis *et al.*, 2008) was used on the

247 CIPRES Portal (Miller *et al.*, 2010) to infer phylogenetic relationships among  
248 *Chrysogorgia* colonies, based on mitochondrial and nuclear sequences, using the  
249 GTRCATI model and automating boot-stopping. The mitochondrial phylogeny was  
250 inferred from the first 700 nt of the *mtMutS* gene (see above); the nuclear phylogeny  
251 was inferred using concatenated RAD loci obtained based on two parameter sets in  
252 Stacks, and one parameter set in PyRAD. The first Stacks set (“m3M4n4”, denovo\_map  
253 parameters m=3, M=4, n=4; populations script parameters m=6, p=2, r=0.5)  
254 corresponds to parameters that maximize the total number of loci detected while  
255 minimizing divergence parameters (see “Exploration of the divergence parameter  
256 space” section above). For this analysis, each *mtMutS* haplotype was considered as a  
257 separate population. The Stacks populations script parameters that were used signify  
258 that 50% missing data were allowed within each population, a locus had to be present  
259 in at least two populations to be included in the output and a minimum of 6 reads per  
260 SNP was required. The second Stacks set (“m3M10n12”, Stacks script denovo\_map  
261 parameters m=3, M=10, n=12; populations script parameters m=6, p=2, r=0.5)  
262 allowed more divergence between loci. The PyRAD dataset (“m6s93”) was  
263 constructed with m=6 and Wclust=93% (details above). In all analyses, the Atlantic  
264 colony JAC1018 was used as the outgroup.

265           Once clades were delimited with RAxML, a Discriminant Analysis on  
266 Principal Components (DAPC, Jombart *et al.*, 2010) was used to explore genetic  
267 structure within three clades represented by 18 to 31 colonies (see below). This  
268 method takes into account the multilocus genotype of each individual and forms  
269 clusters based on genetic similarity without considering a model of evolution. We also  
270 used TESS (Durand *et al.*, 2009) to investigate population structure using the  
271 conditional auto-correlative (CAR) admixture model with a spatially explicit, Bayesian  
272 framework. In TESS, the Deviance Information Criterion (DIC) was used to compare

273 population structure in the presence of different numbers of clusters (the maximum  
274 number of cluster K was set to the total number of individual in the tested clade; for  
275 example, K was set from 2 to 18 for clade 1). Five replicate runs were used per K, with  
276 1 200 MCMC steps and a 200-step burnin. The best K was determined by minimizing  
277 DIC and its variance; once the best K determined, a longer analysis with 12 000 steps  
278 and a 2 000-step burnin was run to obtain reliable individual assignments. The  
279 populations script in Stacks was re-run to keep only one SNP per locus, in order to  
280 minimize the probability of co-analyzing linked markers. The Stacks m3M4n4 dataset  
281 was chosen for these analyses for two reasons: (1) the DAPC and TESS analyses are  
282 run within clades at shallow phylogenetic depths, and (2) as only one SNP / locus is  
283 retained, divergence level should be kept minimal to prevent the inclusion of non-  
284 homologous loci. The DAPC analysis was run using adegenet in R (Jombart 2008; R  
285 Development Core Team 2014).

286

## 287 **Results**

288

### 289 *Mitochondrial typing and RAD-tag sequencing*

290

291 A total of 12 *mtMutS* haplotypes were detected among the 91 colonies investigated, 10  
292 of which were from NC, 3 from PNG, 1 from the northwestern Atlantic, and 2 being  
293 shared between NC and PNG. The biogeography of these mitochondrial haplotypes at  
294 these locations is further discussed in Pante *et al.* (2012ab). A total of 236 million raw  
295 reads, corresponding to 35 463 Mbp were produced on two HiSeq2000 lanes. The  
296 number of quality-filtered reads (in millions) per colony varied between 0.04  
297 (TER11108) and 5.82 (TER2044), with a median of 1.6. There was a significant  
298 correlation between the number of quality-filtered reads per colony and haplotypes

299 (Kruskal-Wallis chi-squared = 25.11, df = 13, p-value = 0.02), haplotypes 6 and 10, for  
300 instance, yielded fewer reads than other haplotypes (haplotype 10 colonies were  
301 sampled from depths down to 880 m, and haplotype 6 colonies had remarkably small  
302 polyps that may have been particularly sensitive to prolonged times to preservation).

303

#### 304 *Loci, SNPs, and indel cataloguing using Stacks and PyRAD*

305

306 Results from both pipelines (Stacks and PyRAD) show variations in the number of loci  
307 and SNPs depending on the set of parameters used (Figure 1a-e, 1g-k), as well as the  
308 mitochondrial genetic distance between samples (Figure 1f). For Stacks, as the  
309 mitochondrial genetic distance among included samples decreases, both the total  
310 number of loci and the number of polymorphic loci increases (Figure 1ab). The former  
311 ranges from a few loci to more than 2 000, whereas the latter ranges from a few loci to  
312 ~1 000, depending on the set of parameters used. When related to time of divergence  
313 (in MY, based on mtDNA), the total number of loci obtained decreases exponentially  
314 (Figure 1f). Inversely, the percentage of polymorphic loci is lower for more closely-  
315 related colonies (~40%) than for distantly related-colonies (~90%; Figure 1c). These  
316 three measures (number of loci, number of polymorphic loci and percentage of  
317 polymorphic loci) show the same response to an increase in divergence parameters M  
318 and n, namely a rapid increase followed by a plateau. This plateau is reached for the  
319 m3M4n4 set of parameters. Conversely, the number of SNPs increases drastically  
320 without reaching a plateau, from a few SNPs for the most stringent set of parameters  
321 and the most distantly-related colonies to around 3 000 for the most closely-related  
322 colonies and the most relaxed set of parameters (Figure 1d). Thus, the effect of  
323 increasing mitochondrial genetic distance among samples or decreasing stringency of

324 parameters is to increase SNPs densities, from one SNP every 250 bp to one SNP every  
325 20 bp (Figure 1e).

326

327 Results of the PyRAD analyses follow the general trends observed for the Stacks  
328 pipeline. These trends are an increase in total number of loci and polymorphic loci  
329 (Figure 1gh) for more relaxed parameters sets, as well as for more closely-related  
330 colonies. As for Stacks, more distantly-related specimen pairs have fewer loci than for  
331 closely-related ones, but a larger proportion of those is polymorphic (Figure 1i). While  
332 the percentage of polymorphic loci shows similar ranges of values for Stacks and  
333 PyRAD, the total number of loci as well as the number of polymorphic loci are almost  
334 doubled (from 2 000 to almost 4 000 and from 1 000 to almost 2 000, respectively).

335 The same pattern is observed for the number of SNPs and SNP densities (Figure 1jk):  
336 PyRAD output differs from Stacks output by a factor of almost two, resulting in SNPs  
337 densities twice as high (from one SNP every 130 bp to one SNP every 20bp). Finally,  
338 unlike Stacks, PyRAD allows for indels within loci. The percentage of loci containing  
339 indels increases with less stringent sets of parameters (Figure 1l). Depending on the  
340 pair of samples considered, this measure varies from a few percent to almost 40 %.  
341 For PyRAD, the number of catalogued loci decreased rapidly with the number of  
342 specimens included in the analysis (with significant drops corresponding to the  
343 number of individuals in the haplotype clades revealed by the phylogenetic  
344 reconstruction, see below) (Figure 2). Most loci bore <3 SNPs even when 10  
345 polymorphisms were allowed on a single RAD locus (Figure 2).

346 We measured the proportion of loci catalogued by Stacks that was also  
347 detected by PyRAD using custom BLASTN database searches. Overall, 0.6 to 42.7% of  
348 loci detected by Stacks were present in the PyRAD catalog. This pattern is partly  
349 explained by the proportion of PyRAD loci with indels (see above), but might also be

350 influenced by the differential detection of repeated regions (i.e. deleveraging  
351 algorithm in Stacks), or the number of reads per individual (the proportion of loci in  
352 common between Stacks and PyRAD was lower for individuals with fewer reads;  
353 Table 2).

354

### 355 *Phylogenetic reconstruction and species delimitation*

356

357 The automatic boot-stopping method implemented in RAxML yielded 1 000 bootstrap  
358 replicates for the mitochondrial phylogeny (91 taxa x 700 nt), 500 replicates for the  
359 Stacks RAD phylogenies (91 taxa x 1 080 352 nt, 11 872 loci for the first dataset, and 1  
360 146 054 nt, 12 594 loci for the second dataset), and 200 replicates for the PyRAD  
361 phylogeny (91 taxa x 6 120 523 nt, 69 851 loci). The proportion of gaps and  
362 undetermined characters ranged between 83 and 84% for Stacks and was 92% for  
363 PyRAD. The three RAD phylogenies were similar but not identical, the second Stacks  
364 dataset being better resolved than the first, and the PyRAD dataset being better  
365 resolved than the Stacks sets (nodes with bootstrap >70%: 19% for m3M4n4, 29% for  
366 m3M10n12, 40% for m6s93; Figure 3). Divergence levels were much higher in the  
367 RAD phylogenies compared to the mitochondrial phylogeny. For instance, the groups  
368 composed of haplotypes 9 and 10 were separated by a distance of 0.001  
369 substitution/site on the *mtMutS* tree, while these clades were separated by 0.27 and  
370 0.25 substitutions/sites on the m3M4n4 and m3M10n12 RAD phylogenies,  
371 respectively (Figure 3).

372         Out of nine mitochondrial haplotypes represented by more than one  
373 individual, six formed well-supported monophyletic groups on the RAD phylogenies,  
374 for all datasets. One of these clades (corresponding to haplotype 10) contained  
375 specimens from both NC and PNG. The group formed by mitochondrial haplotype 7



376 was polyphyletic on the RAD phylogenies, with specimens grouping in two well-  
377 supported clades on the PyRAD phylogeny: one composed of five closely-related NC  
378 specimens and one composed of three more divergent PNG colonies (this clade was  
379 split in two on the Stacks phylogenies). Specimens characterized by *mtMutS* haplotype  
380 7 may therefore belong to at least three distinct species. On the other hand, specimens  
381 characterized by three distinct mitochondrial haplotypes (2, 8, 13) clustered into a  
382 single, well-supported clade (with the exception of one individual, TER13034,  
383 haplotype 8, which clusters well outside this clade). These three haplotypes, which  
384 form a paraphyletic group on the mitochondrial phylogeny and are one to two  
385 mutations different from each other, would therefore be considered as one  
386 evolutionary unit based on the RAD phylogenies (and population clustering analyses  
387 with DAPC and TESS failed to detect structure within this clade; see below). Finally,  
388 out of three singleton haplotypes (J, 13, 14), two (J, 14) sit on long branches and are  
389 clearly differentiated from other haplotypes using RAD-tag data.

390         We ran a DAPC on the three clades that contained the most colonies (clade 1:  
391 18 colonies of haplotype 9; clade 2: 20 colonies of haplotype 4; clade 3: 31 colonies of  
392 haplotypes 2, 8, 13). Within these clades, 3 685, 1 470 and 8 201 loci were retained  
393 (with 25, 42 and 55% missing data, respectively). In all three cases, DAPC failed to  
394 detect intra-clade genetic structure, as the most likely number of group (based on BIC,  
395 discounting the scenario in which each sample belongs to its own group), in each case,  
396 was one (Figure S1). The spatially-explicit admixture model implemented in TESS also  
397 failed to detect genetic structure within clades 1 and 3, but suggested the presence of  
398 three clusters in clade 2, these clusters being composed of colonies sampled (1) on the  
399 slope of New Caledonia, (2) Munida Seamount (Norfolk Ridge), and (3) Jumeaux Ouest  
400 Seamount (Norfolk Ridge; Figure S1). The population genetics of *Chrysogorgia* will be  
401 further discussed in a separate study.

402

403 *Detection of environmental contaminants*

404

405           As octocoral DNA was extracted from whole polyps rather than dissected,  
406 internal tissue, some loci may come from environmental contaminants such as  
407 bacteria. To evaluate the prevalence of such loci, we blasted all the loci that were  
408 catalogued for the m3M4n4 Stacks dataset from individual JAC1018 (n = 1 202). The  
409 BLASTN algorithm (Altschul *et al.*, 1997) was used to match RAD loci to the non-  
410 redundant NCBI nucleotide database, using  $10^{-3}$  as a statistical significance threshold  
411 (e-value). Most sequences (92.6%) could not be assigned to a match in the nucleotide  
412 database and 4.5% of loci were similar to bacterial sequences (78-100% similarity  
413 between match and query). A single locus matched human mitochondrial DNA (84%  
414 similarity); other matches (n = 34) included other invertebrates and plant sequences.  
415 Given (1) the small prevalence of potential contaminants, (2) our inability to  
416 determine whether these loci really belong to contaminant DNA or correspond to  
417 coral sequences which closest matches are non-cnidarian taxa, and (3) the large  
418 number of Stacks analyses performed (>400), we decided to run our analyses without  
419 trying to filter loci from exogenous DNA sources.

420

421

## 422 **Discussion**

423

424 A critical decision in RAD analyses is the way the sequencing data are filtered to get to  
425 the final SNP dataset. This process goes through several steps to ensure that the final  
426 loci will correspond to homologous sequences. The main filters involve several quality  
427 filters (sequencing quality, sequencing depth) as well as several similarity thresholds

428 aimed at identifying the different allelic states of homologous loci. Finally, for each  
429 sample, an algorithm is used to tell apart sequencing errors from real mutations in  
430 order to conduct the final SNP calling. Even though the overall process is quite similar  
431 for Stacks and PyRAD analyses pipelines, a strict comparison of their results is not  
432 straightforward since they use sets of parameters that differ to some extent. A main  
433 difference between these two pipelines is in the assessment of similarity of loci: Stacks  
434 uses a strict similarity criterion (maximum number of mutations) in order to cluster  
435 reads into loci, whereas PyRAD uses an overall similarity criterion, after an alignment  
436 step, allowing for the presence of indels within clusters. This should be a critical  
437 difference when comparing genetically more-distant samples as indels are more likely  
438 to occur, and would thus result in sequences being assigned to different loci using  
439 Stacks (which will then be excluded from the final catalog since not present in both  
440 individuals) while PyRAD would theoretically allow these reads to be considered as  
441 homologous loci.

442

443           Our results show that more loci are recovered using the PyRAD pipeline.  
444 Despite these differences, general trends are similar using both pipelines. First, fewer  
445 loci and SNPs are recovered when comparing more genetically distant samples. This  
446 result is expected and has been anticipated through simulation (Cariou *et al.* 2013)  
447 and observed empirically (Cruaud *et al.* 2014). Our data show an exponential decay of  
448 the number of loci recovered as a function of divergence time of samples. Second, the  
449 stringency of the filtering process has a significant effect on the number of loci and  
450 SNPs identified. Indeed, higher minimum depth of sequencing thresholds and higher  
451 similarity threshold lead to fewer loci being identified. This trend is observed  
452 regardless of the level of genetic divergence between samples, but it seems to be  
453 accentuated when samples are more closely related.

454

455           Despite the similarities in general trends, quantitative and qualitative  
456 differences are observed in the outputs of each pipeline. Indeed, whatever the set of  
457 parameters used, almost twice as many loci are identified using PyRAD compared to  
458 Stacks. This difference cannot be solely attributed to the management of indels since  
459 our results show that the percentage of loci containing indels is usually around 5-20%  
460 and never reaches 40% whatever the genetic distance between samples and the  
461 parameters set. Another interesting result is that PyRAD is not simply adding extra  
462 loci to the total loci identified by Stacks: only half of the loci identified using Stacks are  
463 also present in the PyRAD loci catalogs. It is thus necessary to invoke other filtering  
464 processes and differences in algorithm to explain these differences in output. More  
465 thorough analyses would be needed in order to identify precisely what are the main  
466 sources of divergence in the processing of raw data, in addition to the treatment of  
467 indels.

468

469           One major result is the remarkable loss of homologous loci with increasing  
470 divergence among specimens with different mitochondrial haplotypes. For instance,  
471 compared to specimens sharing the same haplotype, specimens two mutations apart  
472 at *mtMutS* (estimated divergence of 1-2 My) had on average 70% fewer homologous  
473 loci (Stacks analysis at m3M7n8). Within the genus, specimens from mitochondrial  
474 clades 16 mutations apart (i.e. the highest divergence level included in our study,  
475 estimated between 9 and 16 My) share 97% fewer loci. This rate of loss of  
476 homologous RAD-tags is far greater than what has been observed in cetaceans (Viricel  
477 *et al.*, 2014), for which 66% of homologous loci were retained at the inter-familial  
478 level (short-beaked common dolphins, *Delphinus delphis*, vs. harbour porpoise,  
479 *Phocoena phocoena*; estimated divergence of 14-19 My) compared to the intra-specific

480 level (within *Delphinus delphis*). Comparisons within cetaceans were performed using  
481 the same custom pipeline as used in the present study, using Stacks parameters  
482 m3M3n3 (the results for corals were similar when comparing m3m3n3 to m3M7n8).

483

484           The differences observed between our study and that of Viricel *et al.* (2014)  
485 may be explained by various factors. For example, the choice of restriction enzyme  
486 was different (*Sbf1* here, *Not1* for Viricel *et al.*), and differences in genome  
487 composition (most importantly GC content and size) are unknown. While both studies  
488 were conducted with two lanes of Illumina HiSeq2000 sequencing (conducted by  
489 Eurofins Genomics in both cases), throughput may have been influenced by the quality  
490 of genomic DNA (trawled deep-sea samples here, stranded animals for Viricel *et al.*).  
491 These various factors may have significantly influenced the number of cut sites. Our  
492 comparisons might also be significantly affected by the precision of the molecular  
493 clocks available. Divergence times between cetacean families were inferred based on  
494 fossil evidence (see references in Viricel *et al.*, 2014), while no such fossil-calibrated  
495 molecular clock exists, to the best of our knowledge, for octocorals. The *mtMutS*  
496 divergence rates estimated by Lepard (2003) are based on a group of shallow-water  
497 octocorals that may evolve faster than the deep-sea *Chrysogorgia* (a long standing  
498 question in deep-sea biology is whether evolutionary process take longer in deeper  
499 water, compared to shallower waters; e.g. Wilson and Hessler 1987), and rely on a  
500 geological event (rising of the Isthmus of Panama), which can introduce further bias.

501

502           The exploration of divergence parameter space, as outlined above, was made  
503 using pairs of specimens, and not allowing any missing data. Stacks and PyRAD can  
504 build catalogs with loci shared by a set proportion of individuals within pre-defined  
505 groups. Hence, our phylogenetic matrix based on over 12K loci (Stacks parameters

506 m3M10n12) resolved most deeper nodes of the tree despite 83 to 84% of missing  
507 data. Similarly, Cruaud *et al.* (2014) constructed a phylogeny of 18 species of the  
508 beetle genus *Carabus*, and found that the deepest node of the tree (17 My divergence  
509 between species) was characterized by 67% of missing data but strong statistical  
510 support. Jones *et al.* (2013) reconstructed phylogenetic relationships among  
511 congeneric species of swordtail and platyfish (*Xiphophorus* sp.) that diverged <3 My,  
512 and estimated up to 70% missing data (ingroup data; their Table S2). They noted,  
513 however, that missing data had little effect on tree topology and branch support. The  
514 rate of loss of homologous loci observed in swordtail and platyfish is more on par with  
515 what we observed for *Chrysogorgia* than what was reported for cetaceans and *Carabus*  
516 beetles, and further emphasizes that (1) the utility of RAD sequencing for phylogenetic  
517 reconstruction may be taxon-dependent, and (2) molecular clocks must be critically  
518 interpreted. It must be underlined, however, that notable differences in tree  
519 topologies were observed between the three inferred RAD phylogenies, such as deep  
520 but well-supported nodes (e.g. relative positions of clade 3 and haplotypes 6, 7 and 8).

521

522           RAD-tag sequencing has also proven very useful in testing the criterion used  
523 for our primary species delimitation hypotheses, namely that single mitochondrial  
524 *mtMutS* haplotypes discriminate species that fit within the General Lineage Concept of  
525 species as defined by de Queiroz (1998). Indeed, a large numbers of variable loci could  
526 be catalogued within and among closely-related colonies (sharing the same *mtMutS*  
527 haplotype, and therefore putatively belonging to the same species) and more  
528 distantly-related colonies (separated by 1-16 mutations at *mtMutS*, putatively  
529 belonging to different species), allowing us (1) to plot our primary delimitation  
530 hypotheses onto well-supported phylogenies, and (2) to explore the spatial structure  
531 of populations. Three patterns were evidenced from the data: (1) in the majority of

532 cases we noted a complete congruence between *mtMutS* haplotypes and RAD clades  
533 (6/9 non-singleton haplotypes and 2/3 singleton haplotypes); (2) in one case  
534 incomplete congruence was noted (with PyRAD, haplotype 7 corresponding to two  
535 RAD clades (one NC, one PNG) that did not form a monophyletic group; (3) in one case  
536 a single RAD clade included specimens with different (but closely-related) haplotypes.  
537 This result is significant for octocoral taxonomy and systematics, as *mtMutS* has been  
538 widely used to assist species delimitation across a large number of families (e.g.  
539 review of McFadden *et al.*, 2010). While morphological, mitochondrial (Pante and  
540 Watling, 2012) and genomic data (this study) all point to the utility of *mtMutS* for  
541 delimiting *Chrysogorgia* species, its resolution should be interpreted in two ways.  
542 First, as we did not find 100% congruence between RAD clades and *mtMutS*  
543 haplotypes, and tested only a restricted set of putative species, *mtMutS* should still be  
544 considered as one of the first steps in an integrative taxonomic loop incorporating  
545 more variable markers (e.g. Schlick-Steiner *et al.*, 2010; Kekkonen and Hebert 2014).  
546 Second, the evolutionary speed of *mtMutS* may well vary among octocorals, and its  
547 resolving power may therefore vary from one group to another (e.g. Baco and Cairns,  
548 2012). Nevertheless, combining mitochondrial markers such as *mtMutS* and RAD-tag  
549 data will without doubt be of tremendous value for testing the large number of  
550 outdated species hypotheses within the Octocorallia.

551

552

### 553 **Acknowledgments**

554

555 The authors thank the participants of the national research group “Génomique  
556 Environnementale” for stimulating discussions on the use of RAD-tags for inferring  
557 phylogenies, in particular R. Debruyne and A. Cruaud. Samples used in this study were

558 collected during the *Terrasses* and *BioPapua* cruises (PIs S. Samadi and L. Corbari) as  
559 part of the MNHN/IRD Tropical Deep-Sea Benthos programme (Bouchet *et al.*, 2008),  
560 and during the *Extreme Corals 2010* cruise in the northwestern Atlantic (PIs SW Ross  
561 and SD Brooke; funding from NOAA's Deep Sea Coral Research and Technology  
562 Program). We warmly thank the participants and crew members on these cruises for  
563 they indispensable help at sea. This study was funded by the "Institut Écologie et  
564 Environnement" of the CNRS ("Appel à Projets en Génomique Environnementale,"  
565 organizers Dominique Joly and Denis Faure), the French Muséum national d'Histoire  
566 naturelle (Action Thématique du Muséum "Taxonomie moléculaire: DNA Barcode et  
567 gestion durable des collections") and by the Agence National de la Recherche (ANR  
568 12-ISV7-0005-01 French-Taiwanese project TF-DeepEvo). Parts of the analysis were  
569 run on the YMIR super-computer (partly funded by the European Union (contract  
570 31031-2008, European Regional Development Fund) of the University of La Rochelle  
571 (many thanks to Mikael Guichard, Marc-Henri Boisis-Delavaud and Frédéric Bret). We  
572 also thank Julio Pedraza (UMS 2700, MNHN) for his help with bioinformatics. Salary  
573 for EP was covered by a grant to the Poitou-Charentes region (Contrat de Projet État-  
574 Région 2007- 2013). Finally, the authors thank the editor and two anonymous  
575 reviewers for their constructive comments.

576

577 **Conflict of Interest:** The authors declare no conflict of interest.

578

579 **Data Archiving:** Mitochondrial haplotypes were deposited on GenBank (Table S1).

580 Phylogenetic data were deposited on Dryad: doi:xxxxx.

581



582 **References**

583

584 Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997).

585 Gapped BLAST and PSI-BLAST: a new generation of protein database search.

586 *Nucleic Acids Res* **25**:3389-3402.

587 Baco AR, Cairns SD (2012). Comparing molecular variation to morphological species

588 designations in the deep-sea coral *Narella* reveals new insights into seamount

589 coral ranges. *PLoS ONE* **7**: e45555.

590 Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, *et al.* (2008). Rapid

591 SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**:

592 e3376.

593 Bouchet P, Héros V, Lozouet P, Maestrati P (2008). A quarter-century of deep-sea

594 malacological exploration in the South and West Pacific: Where do we stand? How

595 far to go? In: Héros V, Cowie RH, Bouchet P (eds), *Tropical Deep-Sea Benthos* 25.

596 vol. **196**, pp. 9–40.

597 Calderón I, Garrabou J, Aurelle D (2006). Evaluation of the utility of COI and ITS

598 markers as tools for population genetic studies of temperate gorgonians. *J Exp*

599 *Mar Biol Ecol* **336**: 184–197.

600 Cariou M, Duret L, Charlat S (2013). Is RAD-Seq suitable for phylogenetic inference? an

601 *in silico* assessment and optimization. *Ecol Evol* **3**: 846–852.

602 Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013). Stacks: an analysis

603 tool set for population genomics. *Mol Ecol* **22**: 3124–3140.

604 Concepcion GT, Crepeau M, Toonen RJ (2008). An alternative to ITS, a hypervariable,

605 single-copy nuclear intron in corals, and its use in detecting cryptic species within

606 the octocoral genus *Carijoa*. *Coral Reefs* **27**: 323–336.

607 Cruaud A, Gautier M, Galan J M Foucaud, Sauné L, Genson G, Dubois E, *et al.* (2014).  
608 Empirical Assessment of RAD sequencing for interspecific phylogeny. *Mol Biol*  
609 *Evol* **31**: 1272–1274.

610 Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011). Genome-  
611 wide genetic marker discovery and genotyping using next-generation sequencing.  
612 *Nat Rev Genet* **12**: 499–510.

613 de Queiroz K (1998). The General Lineage Concept of Species, Species Criteria, and the  
614 Process of Speciation. In: Howard DJ, Berlocher SH (eds) *Endless Forms: Species*  
615 *and Speciation*. Oxford University Press, pp 57-75.

616 Dueñas L, Sánchez J (2009). Character lability in deep-sea bamboo corals  
617 (Octocorallia, Isididae, Keratoisidinae). *Mar Ecol Prog Ser* **397**: 11–23.

618 Durand E, Jay F, Gaggiotti OE, François O (2009). Spatial inference of admixture  
619 proportions and secondary contact zones. *Mol Biol Evol* **26**: 1963–1973.

620 Eaton D (2014). PyRAD: assembly of *de novo* RADseq loci for phylogenetic analyses.  
621 *Bioinformatics* **30**:1844-1849.

622 Eaton DAR, Ree RH (2013). Inferring phylogeny and introgression using RADseq data:  
623 an example from flowering plants (Pedicularis: Orobanchaceae). *Syst Biol* **62**:  
624 689–706.

625 Ekblom R, Galindo J (2011). Applications of next generation sequencing in molecular  
626 ecology of non-model organisms. *Heredity* **107**: 1–15.

627 France SC (2007). Genetic analysis of bamboo corals (Cnidaria: Octocorallia: Isididae):  
628 does lack of colony branching distinguish *Lepidisis* from *Keratoisis*? *Bull Mar Sci*  
629 **81**: 323–333.

630 France SC, Rosel PE, Agenbroad JE, Mullineaux LS, Kocher TD (1996). DNA sequence  
631 variation of mitochondrial large-subunit rRNA provides support for a two-

632 subclass organization of the Anthozoa (Cnidaria). *Mol Mar Biol Biotechnol* **5**: 15–  
633 28.

634 Gregory TR (2014). *Animal genome size database*. URL: <http://www.genomesize.com>

635 Gruenthal KM, Witting DA, Ford T, Neuman MJ, Williams JP, Pondella DJ, Bird A,  
636 Caruso N, Hyde JR, Seeb LW, Larson WA (2014). Development and application of  
637 genomic tools to the restoration of green abalone in southern California. *Conserv*  
638 *Genet* **15**:109–121.

639 Herrera S, Baco A, Sánchez JA (2010). Molecular systematics of the bubblegum coral  
640 genera (Paragorgiidae, Octocorallia) and description of a new deep-sea species.  
641 *Mol Phylogenet Evol* **55**: 123–135.

642 Jombart T (2008). adegenet: a R package for the multivariate analysis of genetic  
643 markers. *Bioinformatics* **24**: 1403–5.

644 Jombart T, Devillard S, Balloux F (2010). Discriminant analysis of principal  
645 components: a new method for the analysis of genetically structured populations.  
646 *BMC Genetics* **11**.

647 Jones JC, Fan S, Franchini P, Schartl M, Meyer A (2013). The evolutionary history of  
648 *Xiphophorus* fish and their sexually selected sword: a genome-wide approach  
649 using restriction site-associated DNA sequencing. *Mol Ecol* **22**: 2986-3001.

650 Kekkonen M, Hebert PD (2014). DNA barcode-based delineation of putative species:  
651 efficient start for taxonomic workflows. *Mol Ecol Res* **14**:706-15.

652 Leaché AD, Fujita MK, Minin VN, Bouckaert RR (2014). Species delimitation using  
653 genome-wide SNP data. *Syst Biol* **63**:534-542.

654 Lepard A (2003). *Analysis of variation in the mitochondrial encoded msh1 in the genus*  
655 *Lepidogorgia (Cnidaria: Octocorallia) and implications for population and*  
656 *systematics studies*. Master's thesis, College of Charleston, Charleston, SC.

657 Lexer C, Mangili S, Bossolini E, Forest F, Stölting KN, Pearman PB, *et al.* (2013). 'Next  
658 generation' biogeography: towards understanding the drivers of species  
659 diversification and persistence. *J Biogeogr* **40**: 1013–1022.

660 McFadden CS, Benayahu Y, Pante E, Thoma JN, Nevarez PA, France SC (2011).  
661 Limitations of mitochondrial gene barcoding in Octocorallia. *Mol Ecol Res* **11**: 19–  
662 31.

663 McFadden CS, Sánchez JA, France SC (2010). Molecular phylogenetic insights into the  
664 evolution of Octocorallia: A review. *Integr Comp Biol* **50**: 389–410.

665 Miller M, Pfeiffer W, Schwartz T (2010). Creating the CIPRES Science Gateway for  
666 inference of large phylogenetic trees. In: *Proceedings of the Gateway Computing*  
667 *Environments Workshop (GCE)*. New Orleans, LA, United States, pp. 1–8.

668 Mokhtar-Jamaï K, Pascual M, Ledoux JB, Coma R, Féral JP, Garrabou J, *et al.* (2011).  
669 From global to local genetic structuring in the red gorgonian *Paramuricea clavata*:  
670 the interplay between oceanographic conditions and limited larval dispersal. *Mol*  
671 *Ecol* **20**: 3291–3305.

672 Nadeau NJ, Martin SH, Kozak KM, Salazar C, Dasmahapatra K, Davey JW, *et al.* (2013).  
673 Genome-wide patterns of divergence and gene flow across a butterfly radiation.  
674 *Mol Ecol* **22**: 814–826.

675 Pante E, Corbari L, Thubaut J, Chan TY, Mana R, Boisselier MC, *et al.* (2012a).  
676 Exploration of the deep-sea fauna of Papua New Guinea. *Oceanography* **25**: 214–  
677 225.

678 Pante E, France S, Couloux A, Cruaud C C McFadden, Samadi S, Watling L (2012b).  
679 Deep-sea origin and in-situ diversification of chrysogorgiid octocorals. *PLoS ONE*  
680 **7**: e38357.

681 Pante E, France SC (2010). *Pseudochrysogorgia bellona* n. gen. n. sp.: a new genus and  
682 species of chrysogorgiid octocoral (Coelenterata: Anthozoa) from the Coral Sea.  
683 *Zoosystema* **32**: 595–612.

684 Pante E, Watling L (2012). *Chrysogorgia* from the New England and Corner  
685 Seamounts: Atlantic – Pacific connections. *J Mar Biol Assoc U.K.* **92**: 911–927.

686 Puillandre N, Modica MV, Zhang Y, Sirovich L, Boisselier MC, Cruaud C, Holford M,  
687 Samadi S (2012). Large-scale species delimitation method for hyperdiverse  
688 groups. *Mol Ecol* **21**:2671-2691.

689 R Development Core Team (2014). *R: A Language and Environment for Statistical*  
690 *Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-  
691 900051-07-0. URL: <http://www.R-project.org>

692 Reitzel AM, Herrera S, Layden MJ, Martindale MQ, Shank TM (2013). Going where  
693 traditional markers have not gone before: Utility of and promise for RAD  
694 sequencing in marine invertebrate phylogeography and population genomics. *Mol*  
695 *Ecol* **22**:2953-2970.

696 Rubin BER, Ree RH, Moreau CS (2012). Inferring phylogenies from RAD sequence  
697 data. *PLoS ONE* **7**: e33394.

698 Schlick-Steiner BC, Steiner FM, Seifert B, Stauffer C, Christian E, Crozier RH (2010).  
699 Integrative taxonomy: a multisource approach to exploring biodiversity. *Annu Rev*  
700 *Entomol* **55**:421-438.

701 Soza-Ried J, Hotz-Wagenblatt A, Glatting K-H, del Val C, Fellenberg K, Bode HR, Frank  
702 U, Hoheisel JD, Frohme M (2010). The transcriptome of the colonial marine  
703 hydroid *Hydractinia echinata*. *FEBS Journal* **277**:197–209.

704 Stamatakis A (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic  
705 analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–2690.

706 Stamatakis A, Hoover P, Rougemont J (2008). A rapid bootstrap algorithm for the  
707 RAxML web servers. *Syst Biol* **57**: 758–771.

708 Taylor JW, Jacobson DJ, Kroken S, Kasuga T, Geiser DM, Hibbett DS, Fisher MC (2000).  
709 Phylogenetic species recognition and species concepts in fungi. *Fungal Genet Biol*  
710 **31**:21-32.

711 Viricel A, Pante E, Dabin W, Simon-Bouhet B (2014). Applicability of RAD-tag  
712 genotyping for interfamilial comparisons: empirical data from two cetaceans. *Mol*  
713 *Ecol Res* **14**: 597–605.

714 Vollmer S, Palumbi S (2004). Testing the utility of internally transcribed spacer  
715 sequences in coral phylogenetics. *Mol Ecol* **13**: 2763–2772.

716 Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, *et al.* (2013). Genome-  
717 wide RAD sequence data provide unprecedented resolution of species boundaries  
718 and relationships in the Lake Victoria cichlid adaptive radiation. *Mol Ecol* **22**:  
719 787–798.

720 Watling L, France SC, Pante E, Simpson A (2011). Biology of deep-water octocorals.  
721 *Adv Mar Biol* **60**: 41–123.

722 Wilson GDF, Hessler R (1987). Speciation in the Deep Sea. *Annu Rev Ecol Evol Syst*  
723 **18**:185-207.

724 Zhang Z, Schwartz S, Wagner L, Miller W (2000). A greedy algorithm for aligning DNA  
725 sequences. *J Comput Biol* **7**: 203–214.

726

727 **Titles and Legends to Figures**

728

729 **Figure 1.** Comparison of locus detection for Stacks (a-f) and PyRAD (g-l). The number  
730 of loci, SNPs and indels detected for specimens separated by 0-16 mutations at the  
731 mitochondrial *mtMutS* gene are shown for the different read coverage (m  
732 parameter) and divergence levels (M and n parameters, see text). In PyRAD  
733 analyses, “s” corresponds to the “Wclust” parameter.

734

735 **Figure 2.** Information content of the locus catalog built by PyRAD for all 91  
736 *Chrysogorgia* specimens. Wclust: percent divergence permitted between loci  
737 within and across specimens; in addition to the 93% Wclust level used to infer the  
738 *Chrysogorgia* phylogeny, the 89% Wclust level was tested here.

739

740 **Figure 3.** Maximum likelihood phylogenetic trees inferred using RAxML for the  
741 mitochondrial *mtMutS* data (a), and RAD loci (b-d). Bootstrap node support (1000  
742 replicates for a, 500 replicates for b-c, 200 for d) is presented only for nodes with  
743  $\geq 70\%$  support. At the tips, colored dots, which represent *mtMutS* haplotype  
744 membership (each color represents a unique haplotype), are followed by  
745 specimen identifiers and haplotype numbers. Each tree was rooted to the Atlantic  
746 specimen (JAC1018, haplotype J). Genetic structure within clades 1, 2 and 3 were  
747 further investigated using a DAPC and TESS (see text and Figure S1). Scale bars:  
748 substitution / site.

749

750 **Figure S1.** Population genetic structure within three clades of the phylogenetic  
751 analysis. a-c: Bayesian Information Criterion (BIC) values for each tested number  
752 of DAPC cluster. For each clade, the maximum number of clusters was set as the

753 number of individuals minus one (a: clade 1, b: clade 2,c: clade 3). d-e: Boxplots of  
754 Deviance Information Criterion (DIC) values for each value of K. g: Longer TESS  
755 analysis (12 000 MCMC steps) performed on clade 2 colonies, for K=3. g: On the  
756 left, the phylogenetic relationships between colonies within clade 2 are  
757 represented based on the PyRAD dataset, and colored squared at the tips  
758 represent geography (orange: Jumeau Ouest Seamount, green: Munida Seamount,  
759 blue: New Caledonia slope). On the right, q values (ancestry proportions inferred  
760 from the CAR admixture model) are given for each individual from clade 2.

761

762 **Table 1.** Summary table of haplotype information (sample size, geographical spread,  
763 depth range, habitat (seamounts vs. slopes) and *mtMutS* vs. RAD delimitation. NC:  
764 New Caledonia, PNG: Papua New Guinea

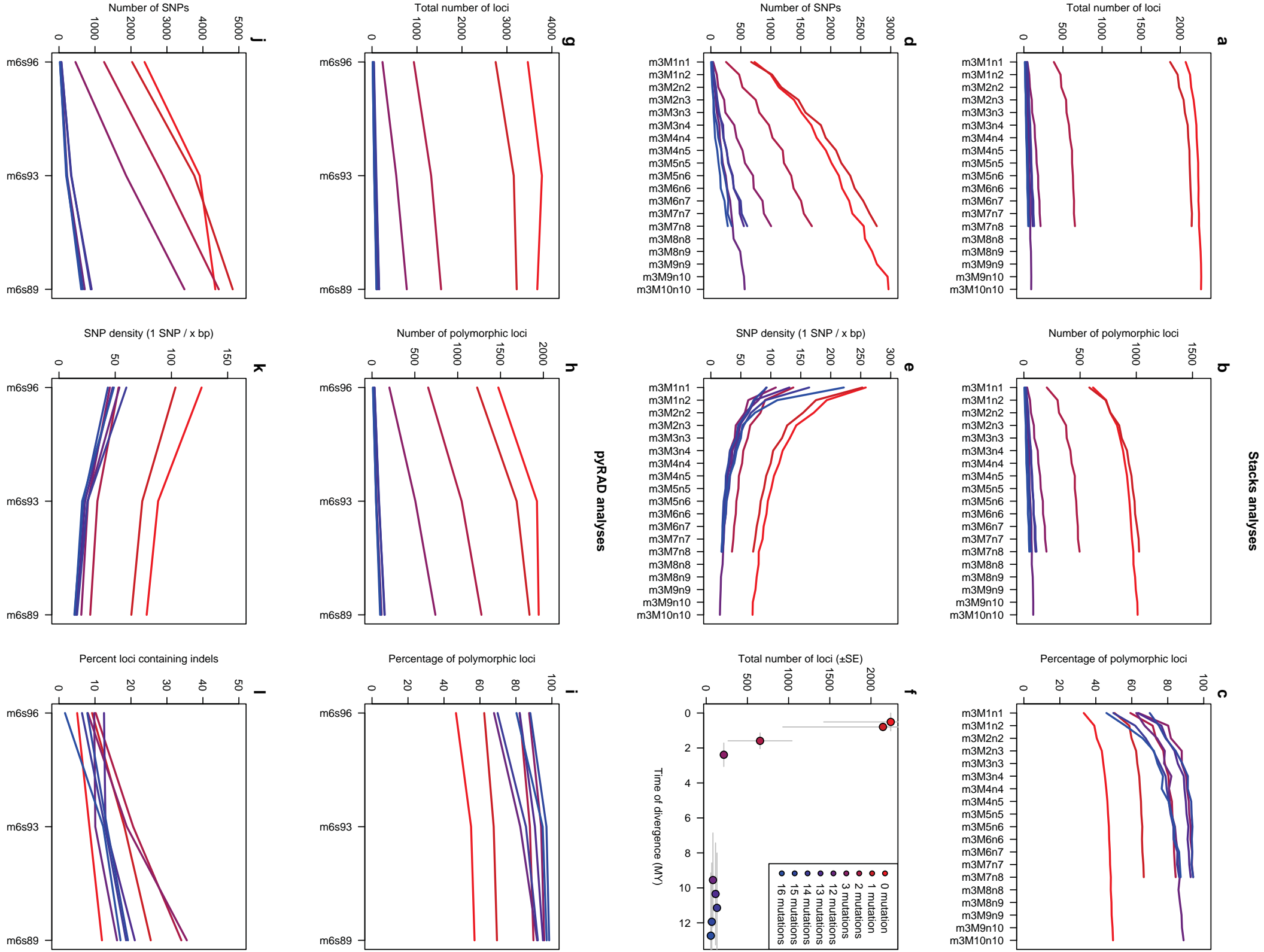
765

766 **Table 2.** Results of the BLASTN alignments performed between Stacks and PyRAD  
767 sequences. The number of loci detected within nine individuals (with high,  
768 medium and low read numbers) is presented for the two analyses performed on  
769 the entire set of 91 specimens. The number of quality-filtered reads is given in  
770 million.

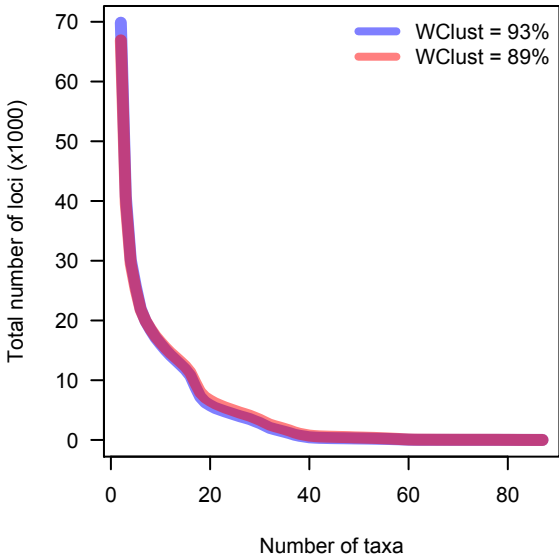
771

772 **Table S1.** Excel table with information on collection (location, date, coordinates,  
773 depth), mitochondrial haplotypes (haplotype number and GenBank accession  
774 number), and number of quality-filtered reads for the 91 *Chrysogorgia* specimens  
775 used in this study. The 6-9 nucleotide barcodes used to distinguish specimens  
776 after Illumina sequencing are also included.

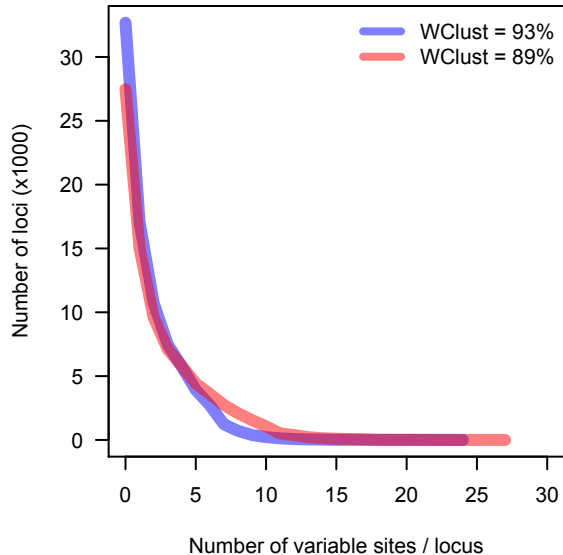




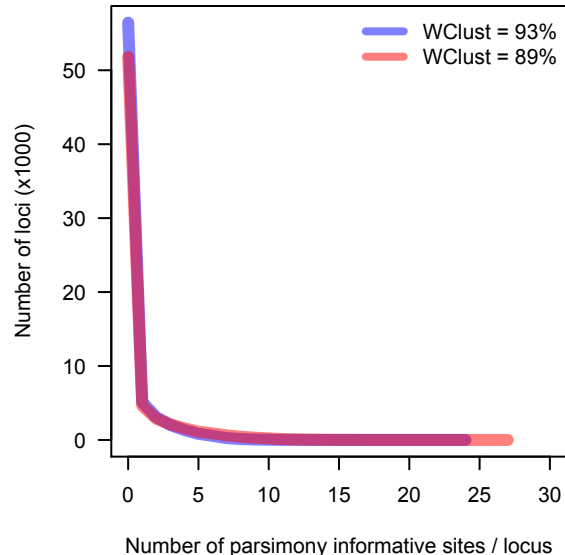
**Number of loci  
across taxa**



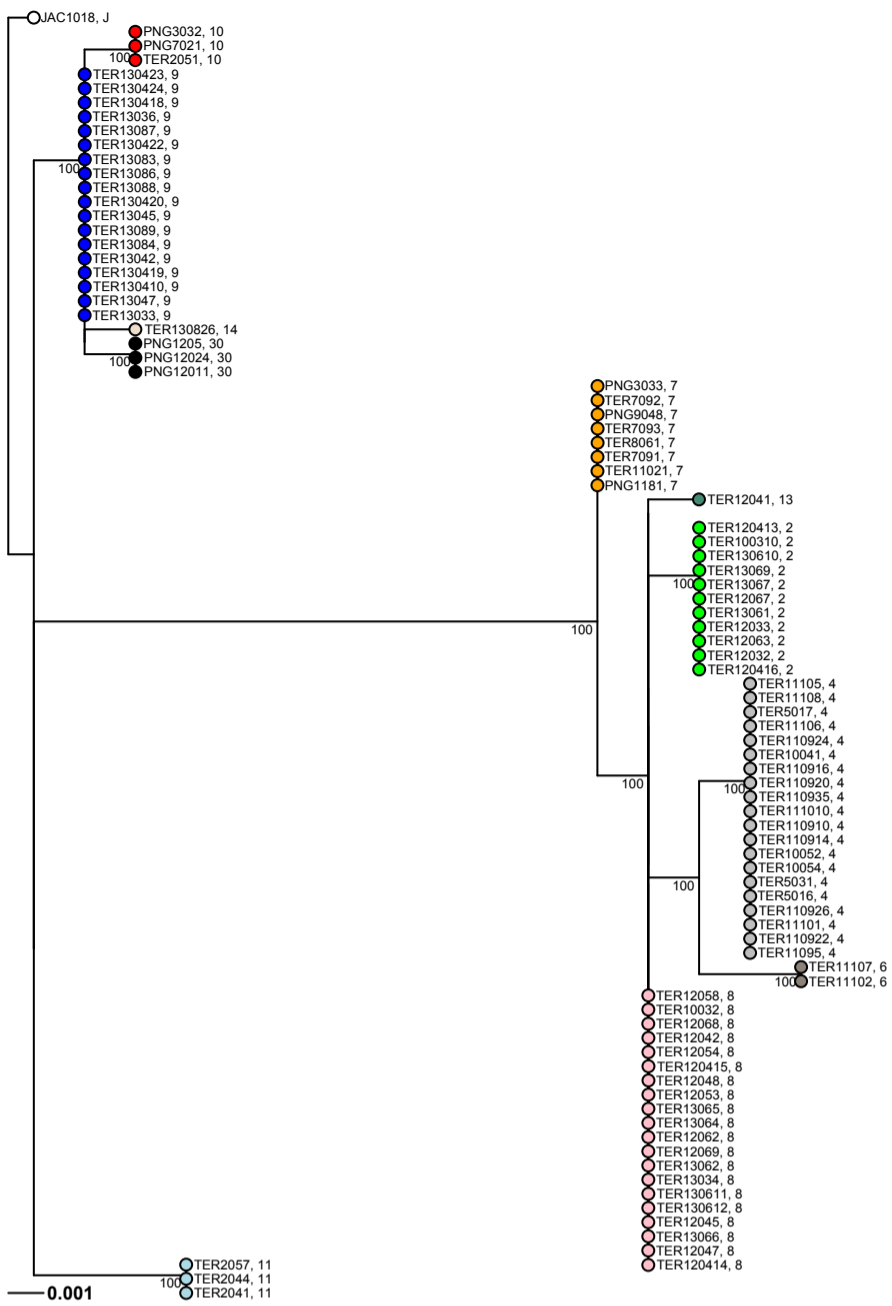
**Density of variable sites  
across loci**



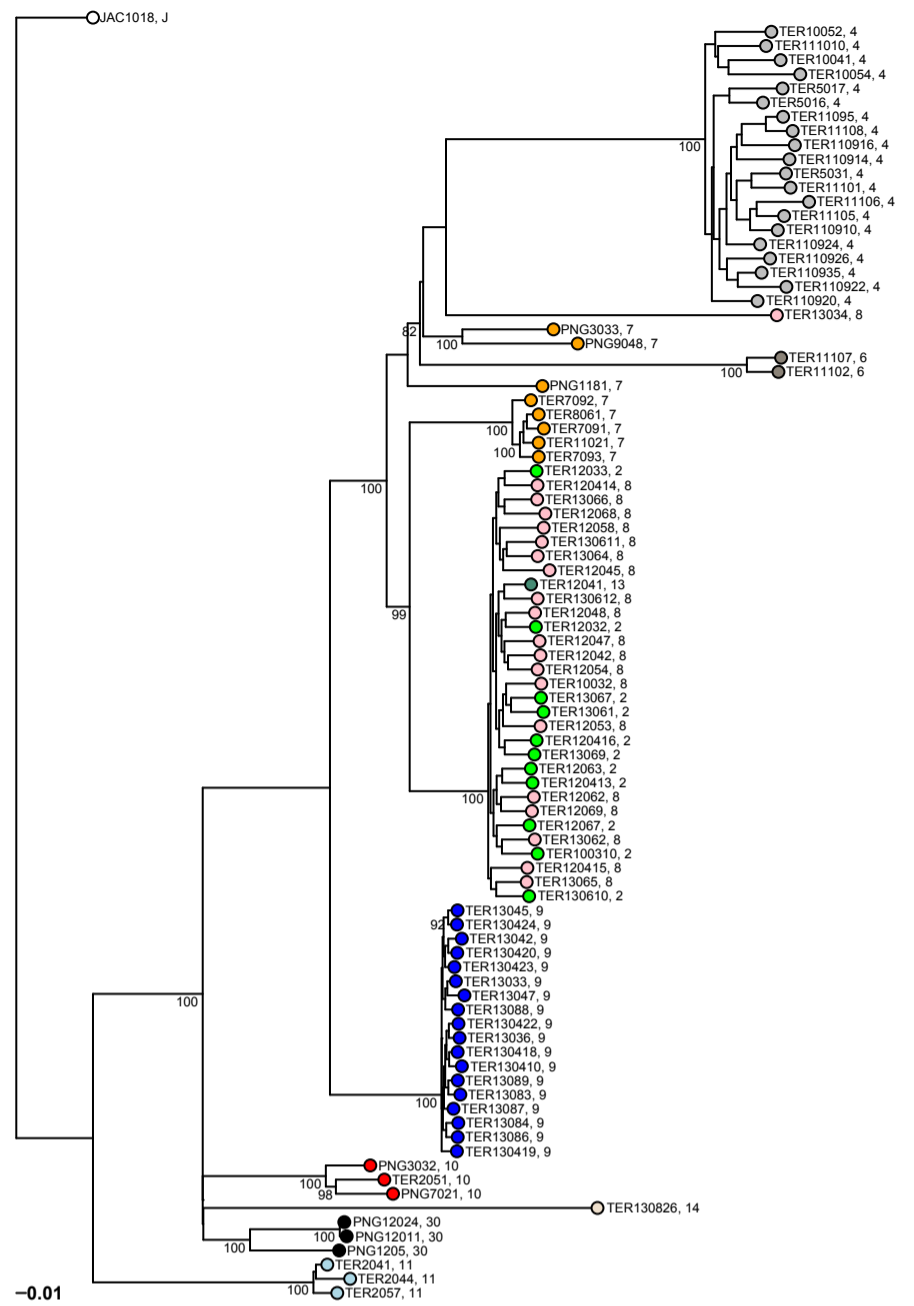
**Density of parsimony-informative sites  
across loci**



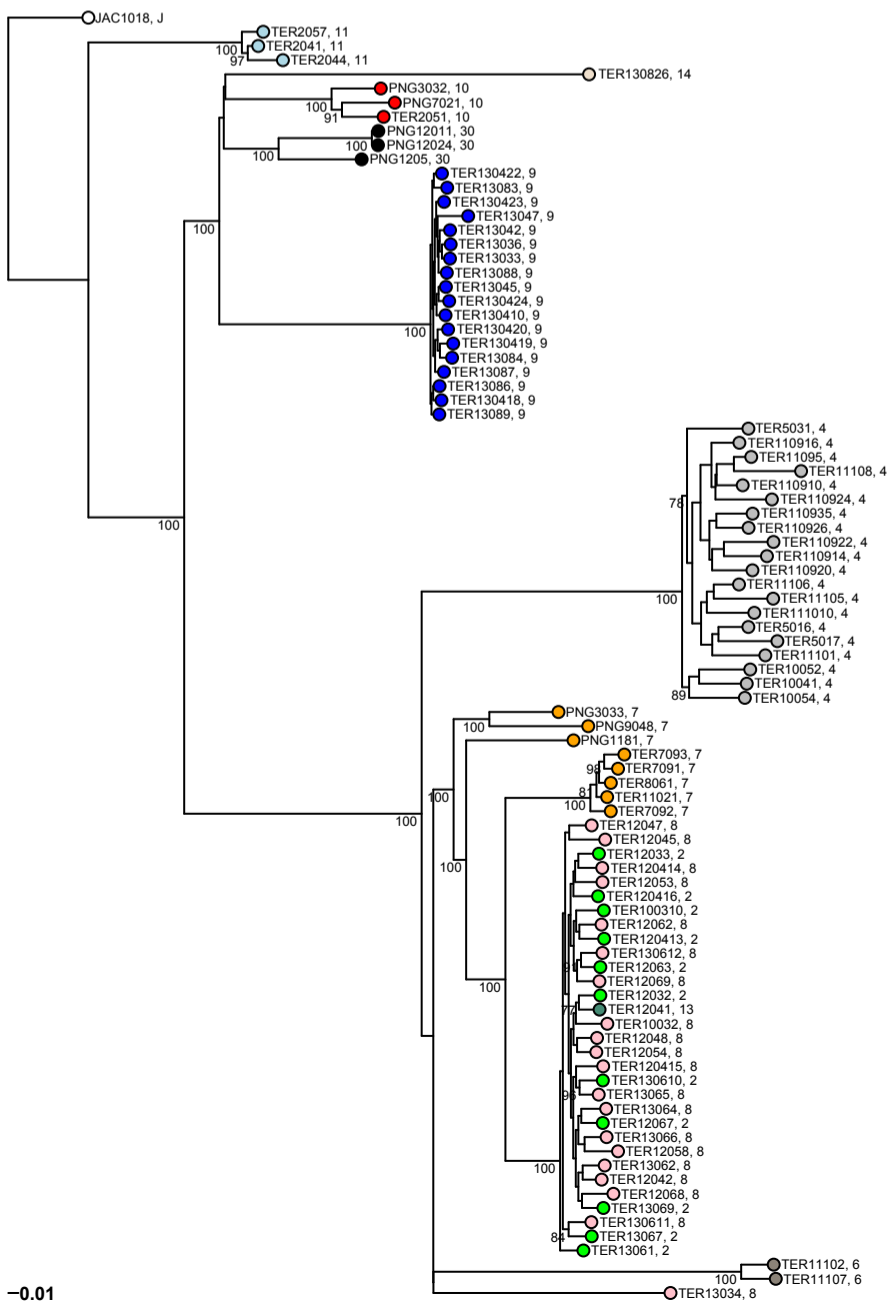
a. Mitochondrial, *mtMutS*



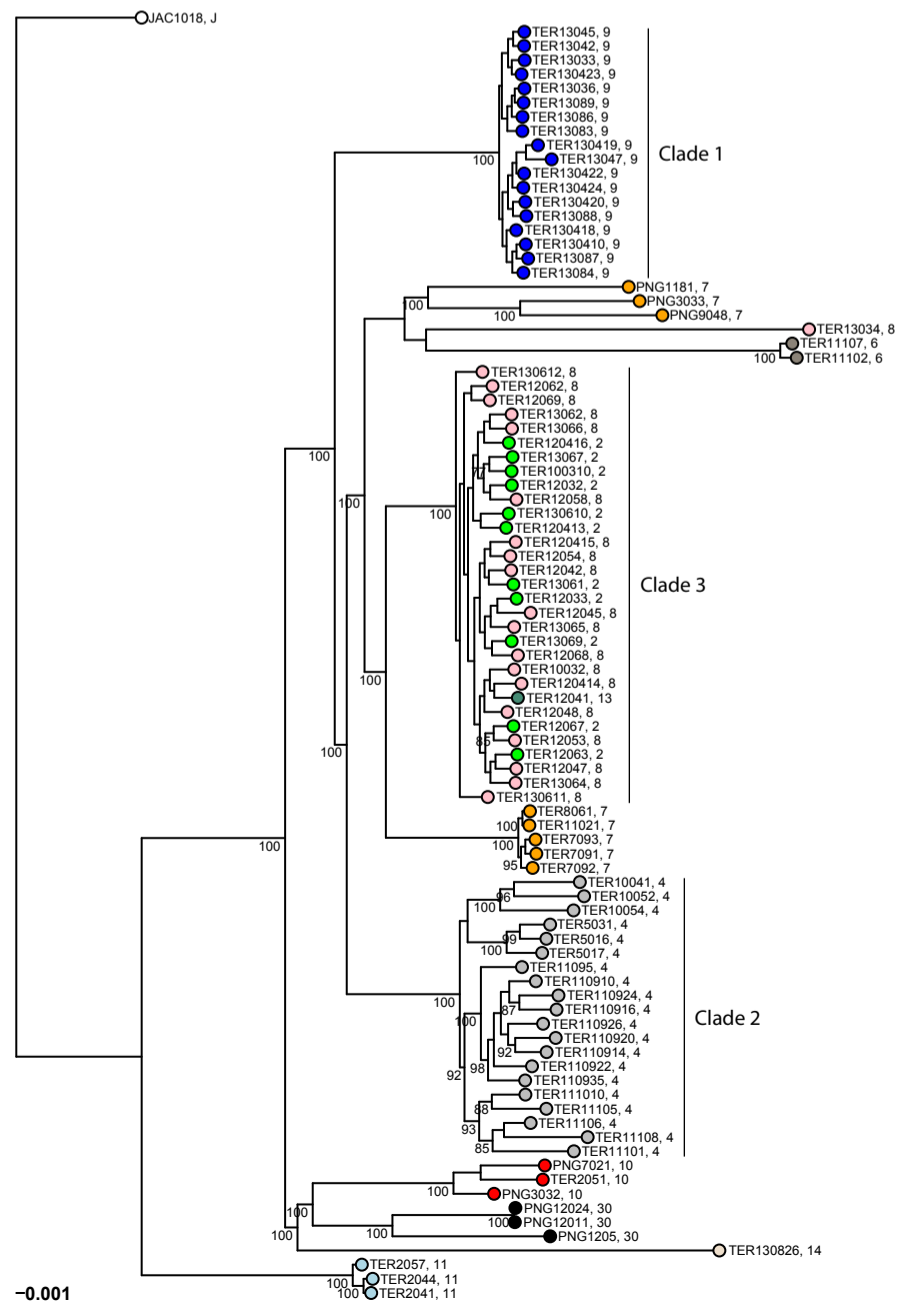
b. RAD-tags, Stacks m3M4n4 dataset



c. RAD-tags, Stacks m3M10n12 dataset



d. RAD-tags, PyRAD m6s93 dataset



Haplotype	N. colonies	Geography	Habitat	Depth range (m)
J	1	Atlantic	slope	627 - 627
2	11	NC	slope	390 - 500
4	20	NC	slope & seamoun	150 - 330
6	2	NC	seamount	270 - 310
7	8	NC-PNG	slope & seamoun	300 - 880
8	20	NC	slope	390 - 500
9	18	NC	slope	390 - 450
10	3	NC	slope & seamoun	458 - 880
11	3	NC	seamount	750 - 840
13	1	NC	slope	460 - 490
14	1	NC	slope	400 - 420
30	3	PNG	slope	220 - 1020

---

**Delimitation**

mtMutS / RAD congruence  
mtMutS / RAD incongruence  
mtMutS / RAD congruence  
mtMutS / RAD congruence  
mtMutS / RAD incongruence  
mtMutS / RAD incongruence  
mtMutS / RAD congruence  
mtMutS / RAD congruence  
mtMutS / RAD congruence  
mtMutS / RAD incongruence  
mtMutS / RAD congruence  
mtMutS / RAD congruence

Specimen	Haplotype	read.category	N. reads (M)	89% divergence	
				N. loci (pyRAD)	N. loci (Stacks)
TER2044	11	high	5.82	6580	866
JAC1018	J	high	5.49	3305	1851
TER7092	7	high	4.04	6867	1363
TER130424	9	median	1.61	6151	1198
TER13064	8	median	1.61	6876	4183
TER13087	9	median	1.60	5959	1131
TER11101	4	low	0.09	1046	228
TER13047	9	low	0.08	1145	396
TER11108	4	low	0.04	441	50

<b>93% divergence</b>			
<b>Intersect (%)</b>	<b>N. loci (pyRAD)</b>	<b>N. loci (Stacks)</b>	<b>Intersect (%)</b>
7.84	6720	607	5.54
24.57	2717	1202	21.46
13.03	6862	1246	11.40
12.73	6323	850	8.86
39.89	6584	4607	42.72
13.81	6189	821	9.26
1.15	944	138	0.64
9.96	1107	297	8.67
2.49	384	32	1.04