MICROSATELLITE DEVELOPMENT IN RHODOPHYTA USING HIGH-THROUGHPUT SEQUENCE DATA

Lucía Couceiro, Isabel Maneiro, Stéphane Mauger, Myriam Valero, José Miguel Ruiz, Rodolfo Barreiro

MICROSATELLITE DEVELOPMENT IN RHODOPHYTA USING HIGH-THROUGHPUT SEQUENCE DATA

Lucia Couceiro[1,2], Isabel Maneiro[1], Stéphane Mauger[2], Myriam Valero[2], Jose Miguel Ruiz[1] and Rodolfo Barreiro[1]

[1] Area de Ecoloxia, Facultade de Ciencias, Universidade da Coruna, Campus de A Zapateira s/n, 1571-A Coruna, Spain

[2] CNRS-UPMC, UMR 7144, Equipe BEDIM, Station Biologique de Roscoff, Place Georges Teissier, F-29682 Roscoff Cedex, France

Shotgun genome sequencing is rapidly emerging as the method of choice for the identification of microsatellite loci in nonmodel organisms. However, to the best of our knowledge, this approach has not been applied to marine algae so far. Herein, we report the results of using the 454 next-generation sequencing (NGS) platform to randomly sample 36.0 and 40.9 Mbp (139,786 and 139,795 reads, respectively) of the genome of two red algae from the northwest Iberian Peninsula [Grateloupia lanceola (J. Agardh) J. Agardh and a still undescribed new member of the family Cruoriaceae]. Using data mining tools, we identified 4,766 and 5,174 perfect microsatellite loci in 4,344 and 4,504 sequences/contigs from G. lanceola and the Cruoriaceae, respectively. After conservative removal of potentially problematic loci (redundant sequences, mobile elements), primer design was possible for 1,371 and 1,366 loci, respectively. A survey of the literature indicates that microsatellite density in our Rhodophyta is at the low end of the values reported for other organisms investigated with the same technology (land plants and animals). A limited number of loci were successfully tested for PCR amplification and polymorphism finding that they may be suitable for population genetic

studies. This study demonstrates that random genome sequencing is a rapid, effective alternative to develop useful microsatellite loci in previously unstudied red algae.

Key index words: 454 sequencing; conservation genetics; expressed sequence tags; microsatellite development; primer design; Rhodophyta; seaweed

Abbreviations: bp, base pairs; ESTs, expressed sequence tags; GS-FLX, Genome Sequencer FLX; Mbp, mega base pairs; MIDs, multiplex identifiers; NGS, next-generation sequencing; PAL, potentially amplifiable loci; TE, Tris-EDTA

Population genetics studies are an important tool for developing effective management strategies for endangered species (Frankham et al. 2004). How- ever, obtaining appropriate molecular markers— a prerequisite for such studies—often becomes a limiting step. This is the case for microsatellites. Although they are often considered the marker of choice for population genetics analysis due to their high levels of polymorphism and codominant nature, their de novo development is a costly and time-consuming endeavor (Zane et al. 2002). These difficulties have acted in many cases as a deterrent for the utilization of these markers, particularly in poorly studied taxonomic groups (Squirrell et al. 2003).

The recent accessibility to NGS technologies is revolutionizing life sciences in several research fields (reviewed in Hudson 2008). Even though NGS is typically cited as a fast and cheap route to assemble genomes or transcriptomes, it also provides avenues to many broader purposes, including marker devel- opment (Thomson et al. 2010). In particular, high- throughput sequence data from the Roche 454 platform offer the potential for a rapid identification of microsatellite loci on nonmodel organisms, and several examples of this have been published already (Abdelkrim et al. 2009, Allentoft et al. 2009, Santana et al. 2009, Castoe et al. 2010, Csencsics et al. 2010). Briefly, the Genome Sequencer FLX System (GS-FLX) powered by 454 sequencing (454 Life Sciences, Branford, CT, USA) enables a random and extensive survey of

the genome provid- ing thousands of short reads (average length of 400 bp using GS-FLX Titanium Series Reagents). This huge data set of reads is then screened with simple bioinformatic tools to detect microsatellite motifs and to design PCR primers. Thus, this approach eliminates the prolonged, expensive, and laborious steps of constructing genomic libraries, including cloning, hybridization to detect positive clones, plasmid isolation, and Sanger sequencing. Herein, we report the results obtained by applying this approach to two Rhodophyta of conservation concern in Galicia (northwest Iberian Peninsula): G. lanceola and a still undescribed new member of the Cruoriaceae. Compared to other taxa, seaweeds have not been traditionally included in the debate about endangered species. However, there is a growing recognition that many seaweeds are restricted in their

distribution and may thus be vulnerable to stochastic environmental and anthropogenic events (Brodie et al. 2009). The two species studied herein are candidates for inclusion in the regional catalog of endangered species (Xunta de Galicia 2007). In Galicia, G. lanceola occurs in only a few (<30) enclaves mostly concentrated in 30 km of coastline (Peez- Cirera et al. 1989, Ba´rbara and Cremades 2004). Besides their regional rarity, these populations possess biogeographic value derived from their marginality and isolation (Eckert et al. 2008) (Galician populations are located > 1,000 km north of the species' main range in the Atlantic coast of Africa). The conservation status of the second species also deserves attention. Its Galician populations were traditionally identified as Furcellaria lumbricalis (Huds.) J. V. Lamour., a common inhabitant of northern latitudes with a southern limit of distribution in the Iberian Peninsula (Pe´ rez-Cirera 1975, Granja et al. 1993). However, some anatomical peculiarities led us to produce DNA sequence data to determine its taxonomic identity. Surprisingly, both plastid (rbcL) and nuclear (rDNA LSU) gene sequences have revealed that the plant found in Galicia is a still undescribed erect member of the Cruoriaceae family (unpublished results). Its limited presence in Galicia (it is known only from 12 enclaves, Barbara et al. 2006) renders this new species an interesting case to investigate its population structure and, eventually, to

78 determine its suitability for inclusion in the regional catalog of seaweeds of conservation
79 concern. Previous experience with conventional protocols for microsatellite isolation suggests
80 that seaweeds may display comparably low genomic frequencies of microsatellites (Wattier and
81 Maggs 2001, Andreakis et al. 2007). Since we were particularly interested in discovering
82 whether or not the use of NGS technology will still support this conventional wisdom, our
83 results with the 454 technology are compared to those for other taxa obtained using comparable
84 high-throughput approaches.
85
86 **MATERIALS AND METHODS**
87
88 DNA extraction and pyrosequencing. For each species, a single specimen was used as the sole
89 source of tissue ⁄ DNA. Prior to DNA extraction, samples were frozen in liquid nitrogen and
90 homogenized using a mechanical tissue disrupter (Mini-Bead- beater 1; Bio Spec Products Inc.,
91 Bartlesville, OK, USA) and steel beads. Total DNA was then extracted with the Wizard
92 Magnetic 96 DNA Plant System (Promega Corp., Madison, WI, USA) following
93 manufacturer's recommendations. To achieve the required amount of DNA, multiple DNA
94 isolations from each specimen were pooled, lyophilized, resuspended in Tris- EDTA (TE)
95 buffer, and purified with QIAquick PCR Purification kit (Qiagen Sciences Inc., Germantown,
96 MD, USA).
97 A total of 2–5 lg of DNA from each species was used to prepare their respective GS-FLX
98 Titanium general libraries (454 Life Sciences) following manufacturer's protocols and quality
99 control steps. Briefly, genomic DNA was fragmented by nebulization, small DNA fragments
100 (<350 bp) were removed, short adaptors were ligated to the ends of each fragment, and, finally,
101 only those fragments with adaptors in both 5' and 3' ends were retained. Each DNA sample
102 was appropriately tagged using two different multiplex identifiers (MIDs) to enable pooling for
103 simultaneous sequencing. DNA library fragments were then captured onto beads and clonally

104 ampli- fied within individual emulsion droplets (emPCR). Finally, amplified fragments from
105 both species were evenly mixed and sequenced on ¼ PicoTiterPlate regions using the Titanium
106 GS-FLX chemistry (454 Life Sciences). Libraries preparation, amplification, and sequencing
107 were carried out at the external facilities Centre de Recerca en Agrigenomica (Barcelona,
108 Spain).
109 Data preprocessing and microsatellite discovery. Each read was automatically assigned to its
110 correct library by GS-FLX software (454 Life Sciences) based on sample-specific MID
111 sequence. In an initial attempt to remove redundancy, sequence reads from each species were
112 analyzed using 454 GS De Novo Assembler v2.3 (454 Life Sciences) with default parameters.
113 The assembled contigs and remaining singletones resulting from this analysis were clustered
114 into a single fasta format file and used as input for the software QDD, a collection of freely
115 available modules (ActivePerl, BLAST, CLUSTALw2 and Primer3) that proceeds in three
116 successive stages (Meglecz et al. 2010). Firstly, it selects sequences that are longer than a user-
117 defined limit and have perfect microsatellites with a minimum number of repeats. Secondly,
118 QDD automatically detects and removes redundancy as well as groups of sequences that
119 potentially are part of a repetitive region of the genome and, consequently, might be problematic
120 for microsatellite amplification (although QDD is not designed to describe mobile elements,
121 this is a conservative method that eliminates many of these sequences). Finally, the software
122 selects sequences that contain a target microsatellite and a flanking region free from
123 nanosatellites and designs PCR primers to amplify these repeats when flanking regions enable
124 it. All QDD parameters, including the parameters of the module Primer3, were set to default
125 values.
126 Data analysis. Tab-delimited files generated from QDD steps 1 (identified loci) and 3
127 (potentially amplifiable loci, PAL) were converted to spread-sheet files for use in Microsoft
128 Office Excel 2003 (Microsoft Corporation, Redmond, WA, USA). Microsatellite frequency for
129 each species was expressed as total number of identified microsatellites per mega base pair

(Mbp), and subsequent data analysis included sorting according to microsatellite category (i.e., di-, tri-, tetra-, penta-, and hexanucleotides), specific motifs and number of tandemly repeated units. Reverse-complement repeat motifs and translated or shifted motifs were grouped together so that there were four unique dinucleotides repeats and 10, 33, 102, and 350 unique tri-, tetra-, penta, and hexanucleotides repeats, respectively (Jin et al. 1994). To provide some comparison with other Rhodophyta, microsatellite frequency ⁄ attributes were also investigated using exactly the same procedure in seven species with >1,000 expressed sequence tag (EST) sequences available in the public database EST-GenBank: Porphyra yezoen- sis, Furcellaria lumbricalis, Eucheuma denticulatum, Gracilaria changii, Porphyra haitanensis, Chondrus crispus, and Griffithsia okiensis. Portions of this data set were screened for micro- satellite motifs by other authors (e.g., EST data for the two Porphyra; Liu et al. 2005, Sun et al. 2006, Wang et al. 2007, Xie et al. 2009), but they were reanalyzed here under common search criteria to facilitate comparison.

Microsatellite testing. A subset of the PAL for each species was tested for amplification (18 and 23 PAL for the Cruoria- ceae and G. lanceola, respectively). To increase the chance of targeting polymorphic loci, we selected PAL with the highest number of repetitions. Markers producing clear electrophero- gram patterns were assessed for polymorphism using a number of individuals from various populations in the Iberian Penin- sula (32 specimens from four populations in the case of the Cruoriaceae; 26–30 specimens from three populations in G. lanceola). DNA extractions were performed as described above, and PCR reactions (10 lL) contained 10 ng of DNA, 200 nM of each primer (Life Technologies Corp., Carlsbad, CA, USA), 150 lM of dNTPs (Fermentas Inc., Geln Burnie, MA, USA), 1.5 mM of MgCl2 (Promega Corp.), 1· PCR Buffer (Promega Corp.), and 0.35 U Taq DNA polymerase (Promega Corp.). Amplifications were carried out separately for each locus on PTC-200 Thermo Cycler (GMI Inc., Ramsey, MI, USA) with initial denaturing at 94 C for 5 min, followed by 30 cycles of 94 C for 30 s, 57 C–60 C for 30 s, and 72 C for 30 s with a final extension of 72 C for 10 min.

DNA fragments were separated on an ABI 3730 Genetic Analyzer (Life Technologies Corp.) with GeneScan 600 LIZ as size standard (Life Technologies Corp.). Finally, electropherograms were analyzed with the software GeneMapper v4.0 (Life Technologies Corp.), and diversity estimates (number of alleles per locus, NA; observed heterozygosity, Ho; expected heterozygosity, He) were calcu- lated with GenALEx 6 software (Peakall and Smouse 2006).

RESULTS

Undescribed Cruoriaceae species. A total of 139,795 reads and 40,896,217 bp were obtained from the 454 shotgun library for the Cruoriaceae (mean read length 293 bp). When sequence reads were assembled using 454 GS De Novo Assembler v 2.3 with default parameters, the new data set consisted in 4,329 contigs plus 51,622 singletones. A total of 5,174 microsatellite loci (di-, tri-, tetra-, penta-, and hexanucleotides) with ‡4 repeats were identified in 4,504 of these sequences amounting to a yield of 0.037 microsatellites per initial read and 126.5 microsatellites per mega base pairs (Mbp). Most loci contained di- and trinucleotide repeats (4,309 and 699, respectively), followed far behind by hexa- (65), tetra- (64), and pentanucleotides (37) (Table 1). After a conservative removal of redun- dancy and potentially problematic sequences (e.g., those located in repetitive regions of the genome), PCR primers design was possible in 1,366 loci (26% of the loci identified). Again, these potentially PAL were dominated by di- and trinucleotide repeats (1,178 and 168 loci, respectively), while larger motifs lagged well behind (£8 loci in each repeat category). A detailed description of these results including a unique accession for each PAL, Gen- Bank accession number, repeat monomer sequence, number of perfect tandemly repeated units, and forward ⁄ reverse primer sequences is provided in Appendix S1 (see the supplementary material).

181 The relative numbers of perfect tandemly repeated units were fairly similar across the various
182 microsatellite categories (Table 1). In all five categories, there were many short repeats and
183 fewer long repeats although the relative abundance of loci with the lowest number of repeated
184 units (four) decreased with microsatellite complexity (from >70% for di- and tri- nucleotides to
185 <55% for larger motifs). Differences were even more pronounced in PAL where most loci
186 (>95%) only contained four to five repeat units. Since di- and trinucleotides were 97% of all
187 iden- tified microsatellite loci and 99% of all PAL, differences in the relative abundance of
188 specific motifs were only investigated in these two categories (Fig. 1). For the dinucleotide
189 repeats, three of the four possible motifs were evenly represented and accounted 88% of the
190 identified loci, while AT was less frequent (12%). Among the trinucleo- tides, one of the 10
191 possible motifs (CCG) was clearly overrepresented and nearly doubled the fre- quency expected
192 simply by chance; the other nine motifs showed comparable frequencies except for two motifs
193 (ACT and AGT) with frequencies <5%. In general, the same pattern was applicable to PAL
194 data. Twelve of the 18 PAL tested for amplification (Table 1) yielded a product of the expected
195 size. Reliable scoring of allele sizes (one to two clear peaks, no multiband pattern) was possible
196 for seven of them, and they were used for subsequent poly- morphism assessment. Among these
197 seven loci, five were polymorphic and heterozygous in at least one individual (Table S1 in the
198 supplementary material) with three to four (mean 3.4) alleles per locus. The observed and
199 expected heterozygosities ranged from 0.031 to 0.969 (mean 0.569) and from 0.254 to 0.656
200 (mean 0.515), respectively. The five polymor- phic loci produced 28 distinct multilocus
201 genotypes in 32 individuals (88%). Grateloupia lanceola. The 454 shotgun library for G.
202 lanceola provided a number of reads very close to those obtained for the Cruoriaceae (139,786
203 reads) but a slightly lower number of base pairs (36,013,235 bp, mean length read 258 bp). The
204 pre- liminary assembling with the 454 GS De Novo Assembler v 2.3 resulted in 2,526 contigs
205 plus 67,005 singletones. A total of 4,766 microsatellite loci (di-, tri-, tetra-, penta-, and
206 hexanucleotides) with ‡4 repeats were identified in 4,344 of these sequences amounting to a

yield of 0.034 microsatellites per initial read or 132.3 microsatellites per Mbp (Table 1). As in the other species, most loci contained di- and trinucleotide repeats (3,959 and 686, respectively) followed far behind by tetra- (59), hexa- (33), and pentanucleotides (29). After removing redundancy in its widest sense, the pro- portion of loci for which PCR primers design was possible was again very similar to that obtained for the other red alga (1,371 loci, 29% of the identified loci). Likewise, PAL were dominated by di- and tri- nucleotide repeats (1,176 and 165, respectively), while more complex motifs lagged well behind (£14 loci). A detailed description of these results includ- ing a unique accession of each PAL, GenBank acces- sion number, repeat monomer sequence, number of perfect tandemly repeated units, and for- ward/ reverse primers sequences is provided in Appendix S2 (see the supplementary material).

As already observed in the Cruoriaceae, the rela- tive abundance of perfect tandemly repeated units was fairly similar across microsatellite categories (Table 1) with abundant short repeats and fewer long repeats. However, the relative abundance of loci with the lowest number of repeated units (four) decreased with microsatellite complexity (86, 75, 69, 62, and 60% for di- , tri-, tetra-, penta-, and hexanu- cleotides, respectively). The higher relative abundance of loci with a small number of repeated units was even more marked in PAL as >95% loci only contained four to five repeats.

Both di- and trinucleotide loci revealed notable changes in the relative abundance of specific motifs and, more interestingly, substantial differences with the other species (Fig. 1). For the dinucleotide repeats, one of the four possible motifs (AG) accounted for 50% of the identified loci, while AC showed a frequency similar to that expected by chance (31%) and the other two motifs (AT and CG) were <15%. Among the trinucleotides, one of the 10 possible motifs (AAG) doubled the frequency expected by chance; the other nine motifs exhibited comparable frequencies (10%) except for AGC and CCG with frequencies <7%. PAL showed the same general pattern of motif frequencies.

A large proportion of the PAL tested for this spe- cies showed good, size-expected amplification prod- ucts (20 of 23 tested PAL, 87%) and easily scorable profiles (19 of 23 tested PAL, 83%) (Table 1). Twelve of these loci were monomorphic, and the seven remaining ones were polymorphic; however, six were heterozygous in at least one individual, while the last one was homozygous across all sam- ples. The mean number of different alleles per locus was lower than in the Cruoriaceae (2.4 vs. 3.4, two to four alleles per locus), and the observed and expected heterozygosities ranged from 0 to 0.933 (mean 0.196) and from 0.034 to 0.531 (mean 0.204), respectively (Table S1). Likewise, the seven loci produced 13 distinct multilocus genotypes in 30 individuals (43%). The low levels of polymorphism detected in G. lanceola are consistent with the low genetic diversity found in this species using AFLP markers (unpublished results).

Microsatellite frequency / attributes in other Rhodophyta. A total of 60,444 public EST sequences (27 Mbp) from the other seven red algae were screened for perfect tandemly repeats units with QDD using the same criteria applied to our shotgun sequencing data (Table S2 in the supplementary material). Micro- satellite frequency oscillated from 261.8 counts Æ Mbp)1 in E. denticulatum to 860.89 counts Æ Mbp)1 in G. changii. As in our Cruoriaceae and G. lanceola, most of the microsatellite loci (>95%) contained di- and trinucleotide repeats. Relative abundances of perfect tandemly repeated units were also similar to our results with genome sequencing data. More than 65% of the detected loci contained just four repeats, while the percentage of microsatellites with >10 repeats was <1% in three species (E. denticula- tum, F. lumbricalis, and G. changii), ranged 1%–5% in C. crispus and G. okiensis, and was >5% in the two Porphyra. The relative abundance of specific motifs also showed considerable variation. AG was the most abundant dinucleotide motif in three species (E. denticulatum, P. haitanensis, and G. changii; 34.4%–44.9%), AC dominated in F. lumbricalis (49.2%) as well as in G. okiensis (39.4%), and CG was the most common motif in C. crispus (44.4%) and P. yezoensis (48.2%). Among the trinucleotides, CCG prevailed in all species except F. lumbricalis, but its relative abundance was variable.

DISCUSSION

Population genetics studies are still comparatively scarce in seaweeds (Valero et al. 2001) and have mostly focused on a limited number of genera ⁄ species (Andreakis et al. 2007). This shortage of intraspecific studies has been attributed, at least in part, to a lack of appropriate molecular markers. In particular, the detection of useful microsatellite loci has been challenging, and it has been suggested that their genomic frequency in this group might be low (Olsen et al. 2002, Andreakis et al. 2007). Our relatively extensive and random survey of the genome of two Rhodophyta should allow some test- ing of this conventional wisdom.

Microsatellite frequency is known to vary to a con- siderable extent among eukaryotes. Even when complete genomes are screened for microsatellite occurrence with consistent procedures, among-taxa differences can be as high as three orders of magni- tude (Sharma et al. 2007). A review of the literature reveals that microsatellite frequency is likewise variable in taxa investigated with a genome sequencing approach comparable to ours (Table S3 in the sup- plementary material). Estimates for a number of plants and animals range from 11.7 microsatellite counts per Mbp to 544.0 counts Æ Mbp)1 (mean 148.4 counts Æ Mbp)1). In this context, our results may seem comparable to those obtained for various animals and higher than the estimates reported for some plants. However, our review also shows that search parameters vary between studies, and, consequently, caution must be exercised in doing this comparison. In particular, the minimum number of repeats for microsatellite detection can be especially influential. Most studies used cutoff values slightly larger than ours, especially for the smaller motifs that often constitute a major fraction of the detected loci. When our estimates were recalculated for a higher cutoff (five for every category), the new microsatellite frequencies (21.9 and 28.8 counts Æ Mbp)1) placed our two red algae in the lower end of the values reported in the literature (Table S3).

As our study is the first case where 454 shotgun genome sequencing was used to detect microsatel- lite motifs in algae, it is difficult to assess to what extent our estimates may be relevant for other taxa. However, it is worth mentioning that our two taxa yielded verysimilar microsatellite frequencies despite the fact that they belong to different orders and show obvious differences in other microsatellite attributes. Thus, the genome of the Cruoriaceae was particularly enriched in CG motifs, while G. lanceola was dominated by AG motifs. This variation is con- sistent with the great differences in di- and trinu- cleotide motifs reported in other photosynthetic organisms (Von Stackelberg et al. 2006). It also sug- gests that microsatellite attributes are species speci- fic rather than group specific, an observation with obvious implications for the enrichment protocols often used for isolating microsatellite loci (Zane et al. 2002) as inappropriate a priori choices about repeat size and motif sequence may lower the effi- cacy of obtaining useful microsatellite loci.

In contrast with the absence of random genomic sequencing data for Rhodophyta, the growing body of publicly available DNA sequences—including vast collections of ESTs—can be used to explore the density/attributes of repeat motifs in the genome of other red algae. Our analyses of EST data for seven species of Rhodophyta (covering four orders and six families) reveal that microsatellite density is 2- to 7-fold higher in EST sequences than in our 454 sequencing data, suggesting that EST collections may be a more effective source for microsatellite mining. As in our random sequencing, EST-derived loci were dominated (>95%) by di- and trinucleo- tide repeats (Table S2); also, dinucleotides were three to five times more abundant than trinucleo- tides in all taxa except the two Porphyra. This preva- lence of dimer repeats is in agreement with the higher density of dinucleotides found in nine of the 25 photosynthetic organisms investigated by Von Stackelberg et al. (2006), and contrasts with the notion that EST-derived microsatellites may be prone to trinucleotides (Varshney et al. 2002, Rota et al. 2005, Wang et al. 2006). However, the particu- lars of our search criteria may have had some influence on our

309 estimates, and alternative estimates for the two Porphyra can be found in the literature (Liu et
310 al. 2005, Sun et al. 2006, Wang et al. 2007, Xie et al. 2009).
311 A survey of service providers indicates that many companies guarantee the development of 10
312 poly- morphic microsatellite loci in 1–3 months with a cost of $5,000–10,000 (Abdelkrim et al.
313 2009). Using shotgun sequencing, we were able to develop 12 polymorphic microsatellite loci
314 in a similar per- iod of time with a final cost of €7,000. Polymorphism (as number of alleles) in
315 our loci was comparable to values reported for other red algae (Guillemin et al. 2005, Andreakis
316 et al. 2007). In this regard, previous analysis with dominant AFLPs markers has revealed very
317 low levels of genetic diver- sity in one of the species included in this study (G. lanceola; I.
318 Maneiro, L. Couceiro, I. Ba´rbara, J. Cremades, J. M. Ruiz, and R. Barreiro, unpub- lished),
319 and the slightly lower number of alleles found for this species, as well as the high propor- tion
320 of monomorphic loci (12 ⁄ 19), seems consistent with these previous results. Therefore, our
321 approach shows a time- and cost-effectiveness comparable to those of traditional methods of
322 microsatellite development. However, the genome sequencing approach still offers a number
323 of alternatives that may enhance its effectiveness. Using microsatellite- enriched DNA before
324 pyrosequencing, Santana et al. (2009) reported that 25%–97% of their initial reads contained
325 microsatellite motifs (estimates vary depending on the species and enrichment protocol used),
326 values that clearly surpass the <10% found in our two red algae. Likewise, our microsatellite
327 search was restricted to perfect motifs, but there is some evidence that complex repeated core
328 motifs (i.e., multiple or interrupted repeats) dominate in other red algae (Wattier et al. 1997,
329 Luo et al. 1999, Guillemin et al. 2005, Andreakis et al. 2007). The possibility that complex
330 patterns may be more abundant in red algae surely warrants further investiga- tion and suggests
331 that the development of polymorphic loci from NGS data could benefit from less stringent
332 screening criteria. Although inter- rupted repeats are expected to have lower mutation rates
333 compared to pure ones (Buschiazzo and Gemmell 2006), empirical data from red algae reveal

that many useful microsatellite loci have complex patterns (Wattier et al. 1997, Luo et al. 1999, Guillemin et al. 2005, Andreakis et al. 2007).

Microsatellite detection with genome sequencing data has other additional benefits. The access to a large volume of sequences allows the detection of putative mobile elements by spotting sequence simi- larity groups (Meglecz et al. 2010). Using the package QDD (Meglecz et al. 2010), we removed 25.8%–26.2% of the sequences with potential loci, possibly saving time and money in primer design and laboratory testing for useless loci since PCR amplification can be seriously affected by microsatel- lite and mobile element associations. On the other hand, genome sequencing produces a large number of novel genomic resources that may be useful for both related and unrelated research such as gene content, gene functions, or even the association of PAL with coding regions (Sharma et al. 2007, Tang- phatsornruang et al. 2009).

In summary, we have reported the cost- and labor-effective development of microsatellite markers in two red algae using shotgun sequencing data (454 technology). To our knowledge, this is the first time that random genome sequencing is used for microsatellite identification in seaweeds. Our results suggest that microsatellite density in our two red algae is below the levels reported for other taxa studied with shotgun technology. However, and despite the use of a stringent filtering procedure to avoid redundancy, we still obtained a substantial number of PAL. A limited number of these PAL were successfully tested for PCR amplification and polymorphism finding that they may be suitable for population genetic studies. Shotgun sequencing appears as an efficient alternative for the fast devel- opment of microsatellite loci in difficult and/or poorly studied red algae species.

The authors are deeply indebted to Ignacio Barbara, Javier Cremades, Viviana Pena, and Pilar Dıaz for their assistance with sampling and species identification. This research was supported by Spain's Secretaria de Estado de Investigacion, Ministerio de Ciencia e Innovacion (project CTM2007- 61011).

360

361 **References**

362 Abdelkrim, J., Robertson, B., Stanton, J. A. & Gemmell, N. 2009. Fast, cost-effective
363 development of species-specific microsatellite markers by genomic sequencing. BioTechniques
364 46:185–92.

365 Allentoft, M., Schuster, S. C., Holdaway, R., Hale, M., McLay, E., Oskam, C., Gilbert, M. T.,
366 Spencer, P., Willerslev, E. & Bunce, M. 2009. Identification of microsatellites from an extinct
367 moa species using high-throughput (454) sequence data. BioTechniques 46:195–200.

368 Andreakis, N., Kooistra, W. & Procaccini, G. 2007. Microsatellite markers in an invasive strain
369 of Asparagopsis taxiformis (Bon- nemaisoniales, Rhodophyta): insights in ploidy level and sex-
370 ual reproduction. Gene 406:144–51.

371 Barbara, I. & Cremades, J. 2004. Grateloupia lanceola versus Grate- loupia turuturu
372 (Gigartinales, Rhodophyta): en la Penınsula Iberica. An. Jard. Bot. Madr. 61:103–18.

373 Barbara, I., Dıaz, P., Cremades, J., Pen˜ a, V., Lopez-Rodrıguez, C., Berecibar, E. & Santos, R.
374 2006. Catalogo gallego de especies amenazadas y lista roja de las algas bentonicas marinas de
375 Galicia. Algas 35:9–19.

376 Brodie, J., Andersen, R. A., Kawachi, M. & Millar, A. J. K. 2009. Endangered algal species
377 and how to protect them. Phycologia 48:423–38.

378 Buschiazzo, E. & Gemmell, N. J. 2006. The rise, fall and renaissance of microsatellites in
379 eukaryotic genomes. Bioessays 28:1040–50.

380 Castoe, T. A., Poole, A. W., Gu, W., de Koning, A. P. J., Daza, J. M., Smith, E. N. & Pollock,
381 D. D. 2010. Rapid identification of thousands of copperhead snake (Agkistrodon contortrix)
382 micro- satellite loci from modest amounts of 454 shotgun genome sequence. Mol. Ecol. Resour.
383 10:341–7.

384 Csencsics, D., Brodbeck, S. & Holderegger, R. 2010. Cost-effective, species-specific
385 microsatellite development for the endangered dwarf bulrush (Typha minima) using next-
386 generation sequencing technology. J. Hered. 101:789–93.
387 Eckert, C. G., Samis, E. & Lougheed, C. 2008. Genetic variation across species' geographical
388 ranges: the central-marginal hypothesis and beyond. Mol. Ecol. 17:1170–88.
389 Frankham, R., Ballou, J. D. & Briscoe, D. A. 2004. A Primer of Con- servation Genetics.
390 Cambridge University Press, Cambridge, UK, 220 pp.
391 Granja, A., Cremades, J. & Ba´rbara, I. 1993. Contribucio´ n al co- nocimiento de la flora bento´
392 nica marina del noroeste de la Pen´ınsula Ibe´ rica I. Primeros datos sobre el cara´cter flor´ıstico
393 del litoral de Lugo (Galicia). Nova Acta Cient. Compostelana (Biol.) 4:15–23.
394 Guillemin, M. L., Destombe, C., Faugeron, S., Correa, A. & Valero, M. 2005. Development of
395 microsatellites DNA markers in the cultivated seaweed, Gracilaria chilensis (Gracilariales,
396 Rhodo- phyta). Mol. Ecol. Notes 5:155–7.
397 Hudson, M. E. 2008. Sequencing breakthroughs for genomic ecology and evolutionary biology.
398 Mol. Ecol. Resour. 8:3–17. Jin, L., Zhong, Y. & Chakraborty, R. 1994. The exact numbers of
399 possible microsatellite motifs. Am. J. Hum. Genet. 55:582.
400 Liu, B. Q., Zeng, Q. G., Luo, Q. J., Wang, Y. J. & Li, S. H. 2005. Isolation of microsatellite
401 loci from dbEST of algae Porphyra yezoensis and primer amplification of interspecies transfer.
402 Oceanol. Limnol. Sin. 36:248–54.
403 Luo, H., Moerchen, M., Engel, C. R., Destombe, C., Epplen, J. T., Epplen, C., Saumitou-
404 Laprade, P. & Valero, M. 1999. Charac- terization of microsatellite markers in the red alga
405 Gracilaria gracilis. Mol. Ecol. 8:685–702.
406 Meglecz, E., Costedoat, C., Dubut, V., Gilles, A., Malausa, T., Pech, N. & Martin, J. F. 2010.
407 QDD: a user-friendly program to select microsatellite markers and design primers from large
408 sequencing projects. Bioinformatics 26:403–4.

409 Olsen, J. L., Sadowski, G., Stam, W. T., Veldsink, J. H. & Jones, K. 2002. Characterization of
410 microsatellite loci in the marine seaweed Ascophyllum nodosum (Phaeophyceae; Fucales).
411 Mol. Ecol. Notes 2:33–4.
412 Peakall, R. & Smouse, P. E. 2006. GENALEX 6: genetic analysis in Excel. Population genetics
413 software for teaching and research. Mol. Ecol. Notes 6:288–95.
414 Pe´ rez-Cirera, J. L. 1975. Notas sobre la vegetacio´ n ficolo´ gica bento´ nica de la R´ıa de
415 Cedeira (NO. de Espan˜ a). An. Inst. Bot. A. J. Cavanilles 32:161–71. Pe´ rez-Cirera, J. L.,
416 Cremades, J. & Ba´rbara, I. 1989. Grateloupia lanceola (Cryptonemiales, Rhodophyta) en las
417 costas de la Pen´ınsula Ibe´ rica: estudio morfolo´ gico y anato´ mico. Lazaroa 11:123–34.
418 Rota, L. M., Kantety, R. V., Yu, J. K. & Sorrells, M. E. 2005. Non- random distribution and
419 frequencies of genomic and EST- derived microsatellite markers in rice, wheat, and barley.
420 BMC Genomics 6:23–34.
421 Santana, Q., Coetzee, M., Steenkamp, E., Mlonyeni, O., Hammond, G., Wingfield, M. &
422 Wingfield, B. 2009. Microsatellite discovery by deep sequencing of enriched genomic libraries.
423 BioTech- niques 46:217–23.
424 Sharma, P. C., Grover, A. & Kahl, G. 2007. Mining microsatellites in eukaryotic genomes.
425 Trends Biotechnol. 25:490–8.
426 Squirrell, J., Hollingsworth, P. M., Woodhead, M., Russell, J., Lowe, A. J., Gibby, M. &
427 Powell, W. 2003. How much effort is re- quired to isolate nuclear microsatellites from plants?
428 Mol. Ecol. 12:1339–48.
429 Sun, J., Liu, T., Guo, B., Jin, D., Weng, M., Feng, Y., Xu, P., Duan, D. & Wang, B. 2006.
430 Development of SSR primers from EST sequences and their application in germplasm
431 identification of Porphyra lines (Rhodophyta). Eur. J. Phycol. 41:329–36.
432 Tangphatsornruang, S., Somta, P., Uthaipaisanwong, P., Chanpra- sert, J., Sangsrakru, D.,
433 Seehalak, W., Sommanas, W., Tra- goonrung, S. & Srinives, P. 2009. Characterization of

434 microsatellites and gene contents from genome shotgun sequences of mungbean (Vigna radiata
435 (L.) Wilczek). BMC Plant Biol. 9:137.
436 Thomson, R. C., Wang, I. J. & Johnson, J. R. 2010. Genome-enabled development of DNA
437 markers for ecology, evolution and conservation. Mol. Ecol. 19:2184–95.
438 Valero, M., Engel, C. R., Billot, C., Kloareg, B. & Destombe, C. 2001. Concepts and issues of
439 population genetics in seaweeds. Cah. Biol. Mar. 42:53–62.
440 Varshney, R. K., Thiel, T., Stein, N., Langridge, P. & Graner, A. 2002. In silico analysis on
441 frequency and distribution of mi- crosatellites in ESTs of some cereal species. Cell Mol. Biol.
442 Lett. 7:537–46.
443 Von Stackelberg, M., Rensing, S. A. & Reski, R. 2006. Identification of genic moss SSR
444 markers and a comparative analysis of twenty-four algal and plant gene indices reveal
445 species-specific rather than group-specific characteristics of microsatellites. BMC Plant Biol.
446 6:9.
447 Wang, C. B., Guo, W. Z., Cai, C. P. & Zhang, T. Z. 2006. Characterization, development and
448 exploitation of EST-derived microsatellites in Gossypium raimondii Ulbrich. Chin. Sci. Bull.
449 51:557–61.
450 Wang, M. Q., Hu, J. J., Zhuang, Y. Y., Zhang, L., Liu, W. & Mao, Y. X. 2007. In silico
451 screening for microsatellite markers from expressed sequence tags of Porphyra yezoensis
452 (Bangiales, Rho- dophyta). J. Ocean Univ. China (Oceanic Coast. Sea Res.) 6:161–6.
453 Wattier, R., Dallas, J. F., Destombe, C., Saumitou-Laprade, P. & Valero, M. 1997. Single
454 locus microsatellites in Gracilariales (Rhodophyta): high levels of genetic variability within
455 Gracilaria gracilis and conservation in related species. J. Phycol. 33:868–80.
456 Wattier, R. & Maggs, C. A. 2001. Intraspecific variation in seaweeds: the application of new
457 tools and approaches. Adv. Bot. Res. 35:171–212.

458 Xie, C. T., Chen, C. S., Ji, D. & Xu, Y. 2009. Characterization, development and exploitation
459 of EST-derived microsatellites in Porphyra haitanensis Chang et Zheng (Bangiales,
460 Rhodophyta). J. Appl. Phycol. 21:367–74.
461 Xunta de Galicia 2007. Decreto 88/2007 do 19 de abril polo que se regula o Cata´logo
462 galego de especies ameazadas. Available at:
463 http://www.xunta.es/Doc/Dog2007.nsf/FichaContenido/12742? OpenDocument (last accessed
464 29 August 2011).
465 Zane, L., Bargelloni, L. & Patarnello, T. 2002. Strategies for microsatellite isolation: a
466 review. Mol. Ecol. 11:1–16.
467

TABLE 1. Counts of loci for each combination of microsatellite category (di-, tri-, tetra-, penta-, and hexanucleotides) and number of perfect tandemly repeat units in each species. Within each cell, values are given for identified/potentially amplifiable microsatellite loci.

	Cruoriaceae species					*Grateloupia lanceola*				
	di-	tri-	tetra-	penta-	hexa-	di-	tri-	tetra-	penta-	hexa-
4	3,409/1,019	515/131	27/5	20/4*	25/4	3,385/1,052	513/124	41/8	18/3	20/7
5	536/118	85/18	11/1*	5/0	9/3*	385/92	107/31	11/5*	8/1	6/3*
6	134/21	38/8	3/1*	4/0	4/0	77/16	35/5	3/0	0/1	4/0
7	69/5	12/6*	7/0	0/0	0/0	30/7	4/0	2/1*	3/0	0/0
8	37/7	14/2*	4/1*	1/0	2/0	11/4*	6/2*	1/0	0/0	2/1*
9	23/5*	9/0	2/0	2/0	4/0	15/2*	5/2*	0/0	0/0	0/0
≥10	73/3*	21/3*	8/0	3/0	17/0	54/3*	16/1*	1/0	0/0	1/0
≥25	17/0	3/0	1/0	2/0	4/1*	2/0	0/0	0/0	0/0	0/0
≥50	11/0	2/0	1/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
Total	4,309/1,178	699/168	64/8	37/4	65/8	3,959/1,176	686/165	59/14	29/5	33/11

*, Denotes categories of potentially amplifiable loci that were tested for amplification and polymorphism.

468
469

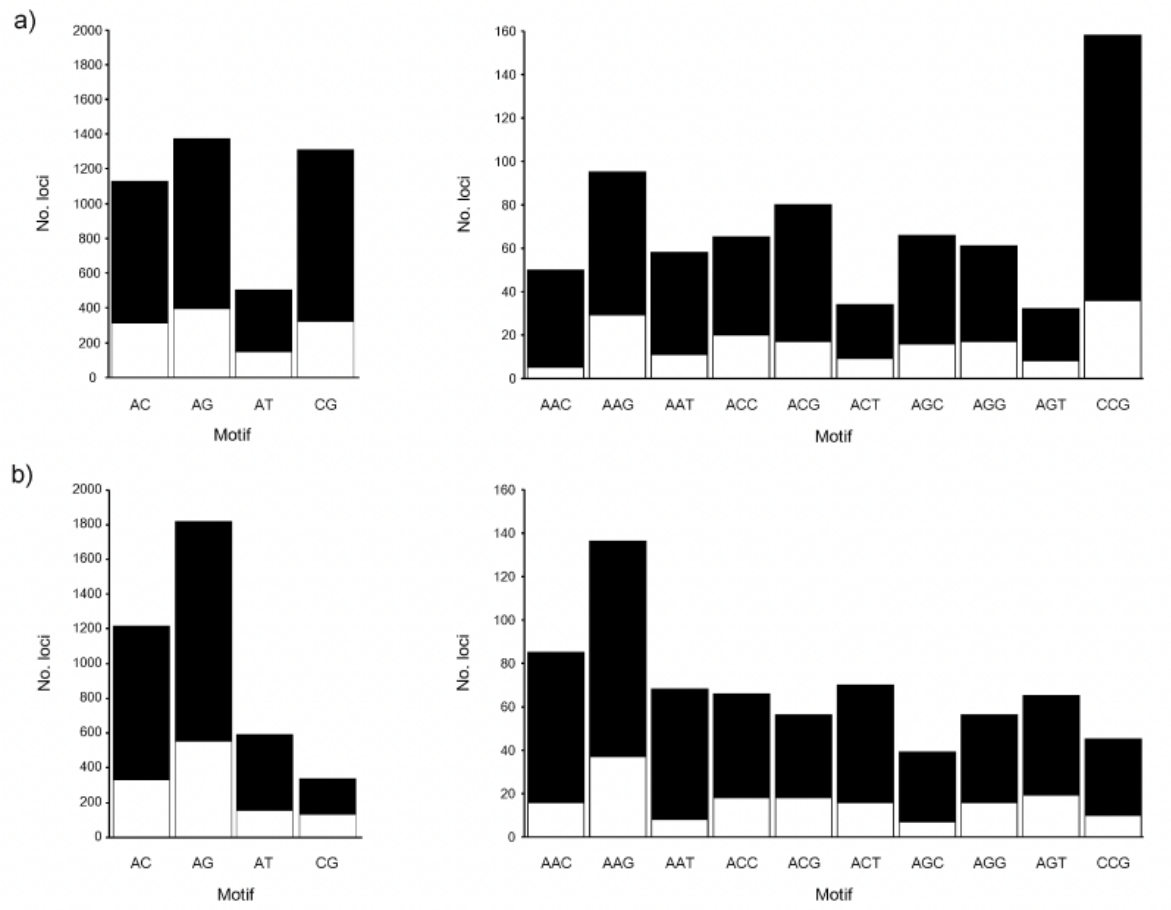

FIG. 1. Motif frequency in di- and trinucleotide repeats. Number of identified microsatellite loci (black) and subset for which PCR primers could be designed (potentially amplifiable loci, PAL; white) for different repeat sequence motifs in (a) undescribed Cruoriaceae and (b) *Grateloupia lanceola*.

470

Supplementary Material

The following supplementary material is available for this article:

Appendix S1. Details of Cruoriaceae PAL (GenBank accession, number of perfect tandemly repeated units, repeat monomer sequence) and designed primers (forward and reverse sequences, melting temperatures).

Appendix S2. Details of *Grateloupia lanceola* PAL (GenBank accession, number of perfect tandemly repeated units, repeat monomer sequence) and designed primers (forward and reverse sequences, melting temperatures).

Table S1. Characteristics of five and seven polymorphic microsatellite loci in the undescribed Cruoriaceae and *Grateloupia lanceola*, respectively (attributes/primer sequences for monomorphic loci are available from authors upon request). *N*: sample size; N_A: number of alleles; H_o: observed heterozygosity; H_e: expected heterozygosity.

Table S2. Frequency/attributes of EST-derived microsatellites for seven Rhodophyta species.

Table S3. Microsatellite frequencies (microsatellite counts per Mbp) reported for genome sequencing approaches in this (as shaded data) and other studies.

This material is available as part of the online article.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

471