Journal of FISH BIOLOGY
Published for the Fisheries Society of the British Isles

Journal of Fish Biology (2013) **82**, 944–958
doi:10.1111/jfb.12034, available online at wileyonlinelibrary.com

Rapid, economical single-nucleotide polymorphism and microsatellite discovery based on *de novo* assembly of a reduced representation genome in a non-model organism: a case study of Atlantic cod *Gadus morhua*

J. Carlsson*†‡, D. T. Gauthier§, J. E. L. Carlsson*, J. P. Coughlan*, E. Dillane*, R. D. Fitzgerald‖, U. Keating*, P. McGinnity*, L. Mirimin‖ and T. F. Cross*

**Beaufort Genetics Research Programme, School of Biological, Earth and Environmental Sciences/Aquauculture and Fisheries Development Centre, University College Cork, Distillery Fields, North Mall, Cork, Ireland, †School of Biology and Environmental Science, UCD Science Education and Research Centre – West, University College Dublin, Dublin 4, Ireland, §Department of Biological Sciences, 202E Mills Godwin Bldg., Old Dominion University, Norfolk, VA, 23529, U.S.A. and ‖Carna Research Station, Ryan Institute, National University of Ireland, Galway, Carna, Connemara, Co., Galway, Ireland*

(Received 10 March 2011, Accepted 21 November 2012)

By combining next-generation sequencing technology (454) and reduced representation library (RRL) construction, the rapid and economical isolation of over 25 000 potential single-nucleotide polymorphisms (SNP) and >6000 putative microsatellite loci from *c*. 2% of the genome of the non-model teleost, Atlantic cod *Gadus morhua* from the Celtic Sea, south of Ireland, was demonstrated. A small-scale validation of markers indicated that 80% (11 of 14) of SNP loci and 40% (6 of 15) of the microsatellite loci could be amplified and showed variability. The results clearly show that small-scale next-generation sequencing of RRL genomes is an economical and rapid approach for simultaneous SNP and microsatellite discovery that is applicable to any species. The low cost and relatively small investment in time allows for positive exploitation of ascertainment bias to design markers applicable to specific populations and study questions.

Key words: 454; Celtic Sea; next-generation sequencing; pyrosequencing; teleost.

INTRODUCTION

Genotyping is a fundamental technique with numerous applications in genomics, phylogenetics, population genetics, pedigree construction, forensics, genetic mapping and investigation of association of loci with traits affected by selection (Beebee & Rowe, 2008). While microsatellites are currently the workhorse for many genetic studies, there has been a shift towards using single-nucleotide polymorphisms (SNP) as the primary marker for many applications. SNPs consist of predominately bi-allelic single base-pair substitutions and share many of the positive characteristics of

‡Author to whom correspondence should be addressed. Tel.: +353 1 716 2197; email: jens.carlsson@ucd.ie

microsatellites, including Mendelian inheritance and co-dominance. SNPs are more densely distributed throughout the genome than microsatellites. SNPs are not prone to common PCR errors that affect microsatellite genotyping, such as non-specific amplification, large allele drop-out (van Oosterhout *et al*., 2004) and homoplasy (Estoup *et al*., 2002). Because the amplicons of SNP alternate alleles are of identical size, subjectivity in interpreting amplicon size is eliminated, and SNP-based genotyping assays are transferrable between laboratories. SNPs also avoid uncertainties regarding appropriate mutation models. Most mutations in microsatellites are stepwise with addition or deletion of complete motifs (stepwise mutation model; Kimura & Ohta, 1978). The mutation rate, however, varies greatly among loci and the proportion of stepwise *v*. other types of mutations (*e.g*. infinite allele model; Kimura & Crow, 1964) is rarely known (Fan & Chu, 2007). The bi-allelic nature of SNPs also enables the use of robotics and high or massive-throughput screening platforms.

Microsatellites have traditionally been discovered by searching existing databases for repetitive sequences, or from construction and sequencing of microsatellite-enriched libraries (Zane *et al*., 2002). Expressed sequence tag (EST) libraries have proven to be a rich source of microsatellites (Vasemägi & Primmer, 2005; Carlsson & Reece, 2007; Higgins *et al*., 2009), however, EST resources are only available for relatively few, well-studied model organisms. Generation of microsatellite-containing libraries can also be outsourced to commercial laboratories (www.genetic-id-services.com/index.htm; www.srel.edu/microsat/Microsat_DNA_Development.html). This is expensive and typically results in relatively low numbers of microsatellite containing sequences (*i.e. c*. 40–100 sequences containing microsatellites with enough flanking sequence to design primers). Thus, while microsatellite discovery is technically straightforward, it can be both time consuming and expensive. Next-generation sequencing techniques such as 454 (Roche; www.roche-applied-science.com) and Illumina platforms (www.illumina.com) have recently been used for microsatellite discovery (Allentoft *et al*., 2009; Santana *et al*., 2009; Castoe *et al*., 2010), and owing to the large amount of sequence data that may be obtained at relatively low cost, it is likely that such techniques will be used to greater extent in the future.

Similar to microsatellite discovery, SNP discovery has traditionally relied on alignment of genomic or EST sequences available in public databases and identifying sequences that show nucleotide substitutions. The major drawback for this technique is that it is not feasible in species for which little or no sequence information is available (*i.e*. most non-model organisms). EST libraries may also be constructed *de novo* and searched for SNPs. This incurs costs, however, similar to those for microsatellite marker development. While EST-linked SNPs are useful for genome-wide expression and mapping studies, their usefulness in estimating population genetic parameters can sometimes be problematic, because accurate estimates of many population genetic parameters assume that the loci used are not affected by selection. Using EST-linked SNPs (or microsatellites), which are more likely to be under selection than loci spread randomly throughout the genome, can bias these types of estimates and lead to erroneous conclusions (Beaumont & Nichols, 1996; Laval *et al*., 2010; Heylar *et al*., 2011). For instance, Allendorf *et al*. (2010) suggested that a small percentage of loci (1–5%) affected by selection could bias F_{ST} estimates up to 50%. SNP discovery based on shotgun sequencing partially solves this problem, as it yields both

coding and non-coding markers. Sequencing depth, however, must be great enough to differentiate SNPs from sequencing errors.

The level of resources necessary to perform Sanger sequencing with enough coverage to discover SNPs at a genome-wide level is only available for a very limited number of species (*e.g. Homo sapiens*). Indeed, for non-model organisms, even pyrosequencing may be cost-prohibitive at the average level of coverage (*c.* ×4) necessary for large-scale SNP discovery in moderately sized genomes. One solution to this problem is to reduce the effective size of the sequenced genome by generation of reduced representation libraries (RRL) (Altshuler *et al*., 2000; Van Tassell *et al*., 2008; Wiedmann *et al*., 2008). In this approach, whole genomic DNA is restriction-digested and separated by gel electrophoresis, and a fraction of DNA within a certain size range is then used for sequencing. In this method, the genomic DNA pool is reduced *c.* 10 to 100 fold (Altshuler *et al*., 2000). RRLs have been used successfully in concert with Sanger sequencing to identify hundreds of thousands of SNP loci from the human genome (Altshuler *et al*., 2000). More recently, pyrosequencing and Illumina sequencing have been used in concert with RRL to identify SNPs in major livestock species such as swine *Sus scrofa* (Wiedmann *et al*., 2008) and cattle, *Bos taurus* (Van Tassell *et al*., 2008) as well as a model finfish species, rainbow trout *Oncorhynchus mykiss* (Walbaum 1792) (Castaño Sánchez *et al*., 2009). Another approach to reduce the genome, which also allows for SNP discovery and simultaneous SNP typing is sequencing of restriction site-associated DNA (RAD) tags using Illumina sequencing techniques (Baird *et al*., 2008).

One disadvantage of SNPs relative to microsatellites is that they are more likely to suffer from ascertainment bias, which stems from selection and development of markers in one population and deployment of those markers in other populations. Some alleles will show high frequency in the population for which they were developed, but will be rare or absent in other populations (Nielsen, 2000). Therefore, utilizing SNPs developed for a certain population (or groups of populations) on other populations may be sub-optimal owing to lack of variability. Developing SNPs for individual populations has been problematic in that sufficient resources for this undertaking were rarely available for non-model species. The RRL–pyrosequencing approach, however, appears to provide the means for relatively low-cost, population-specific marker development.

In this work, RRL and small-scale pyrosequencing were used to discover SNPs and microsatellites in Atlantic cod *Gadus morhua* L. 1758 from the Celtic Sea south of Ireland. The Celtic Sea *G. morhua* is at the southern limit of this species' east Atlantic Ocean distribution and efforts are underway to introduce Celtic Sea *G. morhua* to aquaculture under the National Development Programme for Ireland. The genome of *G. morhua* has recently been sequenced (Star *et al*., 2011) and while there are both microsatellites (Bentzen *et al*., 1996) and EST-linked SNPs (Moen *et al*., 2008; Hemmer-Hansen *et al*., 2011) available for *G. morhua*, markers designed specifically for Celtic Sea *G. morhua* are unavailable. It is anticipated that markers developed in this work will be useful for fine-scale population genetic studies, pedigree reconstruction and genetic mapping and association studies in Celtic Sea *G. morhua*. In addition, the aim is to demonstrate the feasibility of small-scale RRL or pyrosequencing to generate large numbers of potential SNP and microsatellite loci at low cost.

MATERIALS AND METHODS

SAMPLES

Wild *G. morhua* were collected by trawling off Dunmore East, County Waterford, Ireland (52° 09′ N; 6° 59′ W), in the Celtic Sea, during a Ryan Institute–Marine Institute Fisheries Science Services survey on 24 March 2009. Muscle samples from five individuals were preserved in 96% ethanol for genetic analyses. Total body length (L_T) and total body mass (M_T) of individual fish ranged between 22·1–26·5 cm and 0·091–0·190 kg.

TOTAL DNA EXTRACTION

A piece of muscle tissue (*c*. 2 mm^3) was placed in 300 µl cell lysis solution (Qiagen; www.qiagen.com) and 4 µl of proteinase K (20 mg ml^{-1}). The mixture was incubated overnight at 37° C. A volume of 100 µl of protein precipitation solution (Qiagen catalogue number 1045701) was added. The tubes were then vortexed for 20 s, and placed on ice for 10 min, followed by centrifuging at 12470 *g* for 5 min. DNA was precipitated with isopropanol, washed with ethanol and resuspended in 0·1× Tris–EDTA (TE buffer). DNA was quantified using a GeneQuant II RNA–DNA spectrophotometer (Pharmacia Biotech; www.gelifesciences.com).

RESTRICTION DIGEST

Identical amounts of DNA from five individuals were pooled and digested with HaeIII (New England Biolabs; www.neb.com) for 18 h (nucleotide cut site CC—GG). Restriction enzyme was inactivated at 80° C for 20 min, and digested DNA was electrophoresed on a 1·5% agarose gel, using GelPilot 1 kb + size standards (Qiagen). The region of the gel containing fragments of *c*. 600–800 bp in length was excised, and DNA was extracted from the gel fragment (QIAquick; Qiagen). DNA was eluted in 1× TE buffer and purity was assessed by spectrophotometry (A_{260}/A_{280} 2·04). Digested DNA from these five individuals was sent to the Virginia Commonwealth University, Center for the Study of Biological Complexity, U.S.A., for pyrosequencing of a single half-plate using a Roche-454 GS FLX instrument and titanium reagents (Roche) according to standard protocols (mechanical shearing of fragments was omitted from the normal library-construction protocol).

ASSEMBLY, SNP AND MICROSATELLITE DISCOVERY

The acquired sequence data were analysed as follows. After removal of key-tag sequences as well as the two first base pairs (CC, the cut site for the restriction digests), reads were assembled *de novo* using MIRA ver. 3.0.5 (Chevreux *et al*., 2004) with the highly repetitive and no-end clipping switches turned on (−highlyrepetitive -CL:pec). Unedited original sequence reads were then aligned to the reference *de novo* sequence using the ssaha2_pileup pipeline (Ning *et al*., 2001). Potential SNPs were accepted only if two alleles were present and the alternative allele appeared at least twice (*i.e*. a sequence depth of at least ×4 was used for detection) with a minor allele frequency (MAF) of at least 10%. Raw sequence reads were edited (key-tags were removed) and screened for putative di-, tri-, tetra-, penta- and hexanucleotide microsatellites (with at least six repeats for dinucleotide and five repeats for all other motifs) using MISA 1.0 (http://pgrc.ipk-gatersleben.de/misa/). The programme Primer3 (Rozen & Skaletsky, 2000) was used to screen identified microsatellites for potential primers using the default settings. All microsatellite containing sequences ($n = 287$) found on Genbank (www.ncbi.nlm.nih.gov/genbank) using the search term '*Gadus morhua* microsatellites' were imported into the software GENEIOUS PRO ver. 5.4.2 and assembled (not allowing for mismatches or ambiguities, in seven instances when the Primer3-generated primer largely assembled with the microsatellite motif, the assembly was disregarded) with primers suggested by the Primer3 pipeline (using the first suggested primer pair).

TABLE I. Summary statistics for 11 validated single-nucleotide polymorphism loci among *Gadus morhua* from the Celtic Sea

Locus	Variant	Primer sequence (5′–3′)	*n*	Hom	Het
GmoUCC-S04067*	C/A	F-TCTGGATGAGGAACGTAGGG R-GGTGAACTCCAAGGGCTAGA	6	0	6
GmoUCC-S04571*	A/T	F-AAAAAGCTTCATCAACAGGATCA R-ACTCTGGTAGGAGGCGTTCA	5	1	4
GmoUCC-S05529*	A/G	F-GCAAGGGTGTAGTGGTGACA R-AGTGGACGCTAAACGCAAGT	6	2	4
GmoUCC-S05726*	C/T	F-ACGAGGGTGCAGTATGGAAG R-TCGCTGTGGATGTCACTCTC	6	0	6
GmoUCC-S07193*	C/G	F-TTACGGGTGTTTCTGTGTGC R-GACGCTTCAGAGGAGAGGAA	4	3	1
GmoUCC-S32846*	A/G	F-CAAAAGCACAGCCGTAAACA R-GCGTACTGCGAGCAATAACA	5	3	2
GmoUCC-S33793*	C/T	F-CGGAGTGAAAAGTACCACATCA R-CATTCATTCATTGACCTGTTCC	6	3	3
GmoUCC-S37473*	C/T	F-TCAGAATTCACTCTGGATGAAAGT R-GGACGCCTTTCCAAACAGTA	5	4	1
GmoUCC-S37762†	A/T	F-TGAACACCTGGACTCGTTCC R-GCGCCGTGAAAGAGAAAC	5	0	5
GmoUCC-S41809*	A/G	F-TGTCGTGGTGGAGTAATGAGA R-GAAAGCCCCTGGGACTAAAC	6	5	1
GmoUCC-S44855*	A/C	F-GGAGAGGCTACATCATTTGG R-GTGTCTTGCTCCTTTCTTT	6	0	6

F, forward; R, reverse; n, number of individuals; Hom, number of homozygotes; Het, number of heterozygotes.

*Genbank accession numbers: JX878896-JX878905.

†SNP containing sequence: TACAGAGTGAGAATGAACACCTGGACTCGTTCCTCCGGACCTCCT-GAGCCTGACTGAAGTCCAGTTCTCCGTGGTTTCCCCGATAAAGAGACTTAAAATGTACGAAC CCAACAAATCGGAGACTGCTTATTGTAGAAGACA(A/T)TTGGGATCGCGTTTCTCTTTCACGG CGCAGAAGAGAA.

VALIDATION

Six individuals (including five from the original SNP and microsatellite discovery panel) comprised the validation panel and a total of 14 putative SNP loci (Table I) were selected for validation by direct nuclear Sanger sequencing (both strands). SNP loci were selected based on their putative variability with a minimum MAF of 40%. Primers were designed with Primer3 software and subsequently amplified and submitted for sequencing (Macrogen; www.macrogen.com). The finding of two peaks with base composition identical to the original 454 sequence at the SNP locus was considered as a positive validation of the locus.

In all, 48 *G. morhua* from the Celtic Sea, sampled in 2007, were chosen as the microsatellite validation panel. Primers were designed with Primer3 for 15 putative tetranucleotide microsatellite loci (Table II). An M13 modified tail (5′-GGATAACAATTTCACACAGG-3′) or Hill tail (5′-TGACCGGCAGCAAAATTG-3′) was added to the 5′ end of the forward primers following methods described by Schuelke (2000). These loci were amplified and amplicons were visualized and analysed on a Li-Cor 4300 Global IR^2 automated sequencer (Li-Cor Biosciences; www.licor.com) with a size standard (50–350 bp). The software Genepop 4.0.10 (Rousset, 2008) was used to estimate observed and expected heterozygosity and for testing

deviations from Hardy–Weinberg proportions (exact G-test using 1000 dememorization steps and 100 000 Markov chain steps). Alpha levels for significance were corrected for multiple tests using the sequential Bonferroni technique (Rice, 1989).

RESULTS

The half-plate 454 run of the RRL from Celtic Sea *G. morhua* resulted in a total of *c*. 88·2 Mbp spread over 414 237 individual sequence reads with an average read length of 213 bp. MIRA successfully assembled 294 458 sequences (71%), resulting in 74 119 contigs with a median length of 253 bp [mean ± S.D. = 267 ± 116, range 40–1201 bp; Fig. 1(a)] and an average read depth of 4·0 [range 1–112; Fig. 1(b)]. Sequence depth of $\geq \times 4$ was present for 26 160 contigs. The combined contigs had a length of 19·8 Mbp and represent *c*. 2·2% of the *G. morhua* genome, which has been estimated to be *c*. 900 Mbp (Hardie & Hebert, 2004).

All available sequence reads were used against the assembled reference sequence in the pileup assembly for SNP discovery. Ssaha2_pileup analysis identified 25 233 potential SNPs where only two alleles were present and each allele was observed at least twice with an MAF of 10% or more. MAF mean ± S.D. = 0·343 ± 0·122 and ranged from 0·10 to 0·50. The most common SNP variant consisted of A/G or C/T (Fig. 2). The density of SNPs in the Celtic Sea *G. morhua* RRL genome was estimated to be one SNP per 785 bp (25 233 SNPs in 19 808 583 bp).

MISA identified 84 795 microsatellite repeats in the 414 237 sequence reads. Of these, 19 748 were compound microsatellites. As not all sequences used for microsatellite discovery were unique, it is not straightforward to estimate the density of microsatellite repeats. Considering that the average read depth in contigs was 4× and assuming a similar read depth across all sequence reads, the total sequence length of unique sequence would be *c*. 22·1 Mbp [by dividing the total number of sequences by four and multiplying that with the average sequence length (213 bp)]. Hence, the density of microsatellite loci in the RRL genome of Celtic Sea *G. morhua* can be estimated to one microsatellite repeat per 260 bp. Numbers of the repeat classes were 72 466 di-, 7507 tri-, 4160 tetra-, 346 penta- and 316 hexanucleotide repeats (compound microsatellites may be represented in several repeat classes and multiple times). The most common dinucleotide motif was AC, while TCC and ACAG were the most common motifs in tri- and tetranucleotides, respectively (as stated above, compound loci are represented more than once as they can have multiple motifs; Fig. 3). The motif CCGG was not encountered, consistent with the recognition site of the restriction digest enzyme HaeIII. The number of microsatellite loci per motif ranged from 162 to 41 946 in hexa- and dinucleotide loci, respectively (Fig. 4). Primer3 was able to design primers for 11 341 (Fig. 4) microsatellite loci discovered here with a design (unvalidated) success rate ranging between 7% (compound loci) and 54% (pentanucleotide loci).

As the average read depth of the contigs was 4×, it is likely that a number of microsatellite loci were sequenced more than once. In order to estimate the proportion of such loci, 100 forward primers resulting from the Primer3 analysis were selected at random. The number of exact matches for a given primer sequence occurring within original microsatellite containing sequences varied greatly (median = 2, range 1–44) with a mean ± S.D. of 3·9 ± 6·3. Of these, 38% of primers were only

TABLE II. Summary statistics for nine microsatellite loci among *Gadus morhua* from the Celtic Sea

Locus	Repeat structure	Primer sequence (5′–3′)	Tail	Fluorescent label	*n*	*a*	*as*	H_E	H_O	HW
GmoUCC-01*	$(ACAT)_n$	F-ATGCGTCTCGAAATGGACA R-GAGTGTGCGTGCGTGAGT	M13	IRD-700	34	2	90–94	0·25	0·24	>0·05
GmoUCC-04*	$(CAGA)_n$	F-GAGGAACACTGTCACCACCA R-GTTTCCCCTCGGTTGAGACT	M13	IRD-700	47	4	78–90	0·21	0·17	>0·05
GmoUCC-05*	$(CTTT)_n$	F-TGAATAAAACCCGCCATCTC R-CAGGTCAACGGACCACATAA	M13	IRD-700	34	14	192–264	0·85	0·59	<0·001
GmoUCC-10*	$(AAAT)_n$	F-CCGAAGATTTGCAGTTCAGA R-ATTGGCTGCCCAGGTCTTAT	M13	IRD-700	34	2	160–168	0·42	0·41	>0·05
GmoUCC-11*	$(AAAT)_n$	F-CGGAGGTTATTTGGGATTT R-CAGGCTGCATCAGAACAACT	HILL	IRD-800	41	8	158–198	0·69	0·32	<0·001
GmoUCC-15*	$(AACT)_n$	F-AATGACCGTGTGTCACTGC R-ACATCCCTTCAGGGTTAGGG	HILL	IRD-800	45	2	105–109	0·12	0·13	>0·05

F, forward; R, reverse; n, number of individuals; *a*, number of alleles; *as*, allele size range in base pairs; H_E, expected heterozygosity; H_O, observed heterozygosity; HW, probability values of concordance with Hardy–Weinberg expectations [estimates after correction for multiple tests (initial $\alpha = 0{\cdot}05/6 = 0{\cdot}0083$)].
*Genbank accession numbers: JX887870–JX887875.

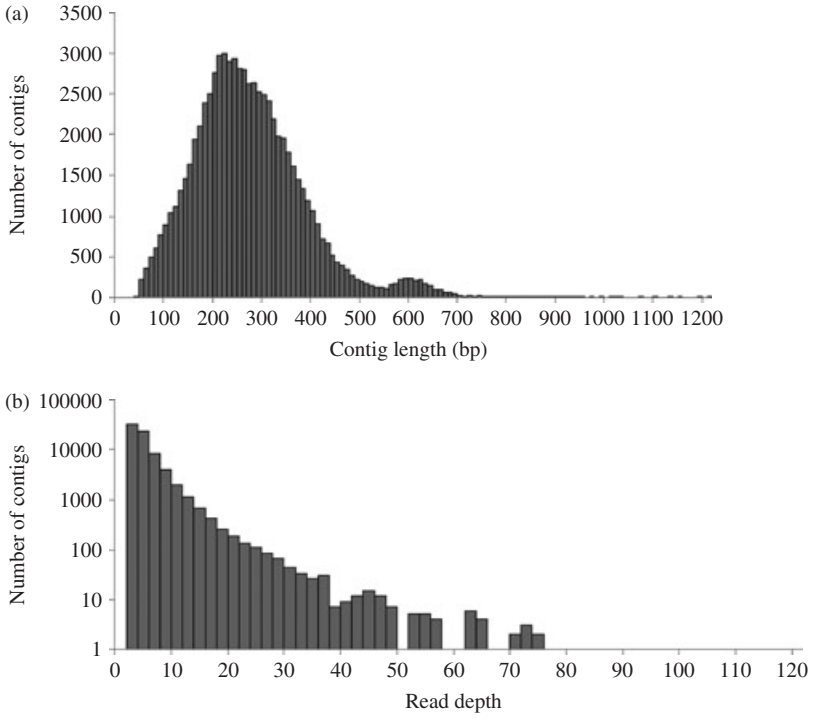

FIG. 1. Distribution of (a) contig lengths and (b) read depth from single half-plate 454 run of the *Gadus morhua* reduced representation library genome. Logarithmic scale is given in (b) for number of contigs.

found once and 19, 10 and 13% were encountered two, three and four times, respectively. This result indicates that of the 11 341 microsatellite loci for which primers were designed, *c*. 4309 were likely to have been encountered only once. Similarly, as 19% (2154) were observed twice, half of this number (1077) can be considered unique. Using the same approach for all loci tri-, tetra-, penta- and hexanucleotide loci, the total number of unique microsatellite loci expected, from the original 11 341 for which primers could be designed, was 6424. The assembly of available *G. morhua* microsatellites on GenBank with the 11 341 primer pairs found by Primer3 resulted in a recovery of 11 microsatellites from GenBank. Hence, 11 microsatellites detected in this study were already in GenBank (GenBank accession numbers: EU860239, FJ007680, FJ007690, FJ007694, FJ007702, FJ007707, FJ007708, FJ007726, FJ007728, FJ007774 and FJ007784).

A total of 14 putative SNP loci were sequenced (Table I) and 11 of these showed variability. Three of the putative SNP loci showed no sequence variability and could hence not be validated. Of the 11 loci that were validated, seven showed both homozygotes and heterozygotes while three loci showed only heterozygotes. It is possible that the four loci showing 100% heterozygosity could be due to alignment of pseudogenes and hence false positives as has been observed in many salmonids (Hohenlohe *et al*., 2011). *Gadus morhua*, however, is not known to have undergone genome duplication, and because putative SNPs with MAFs of 40% or more were

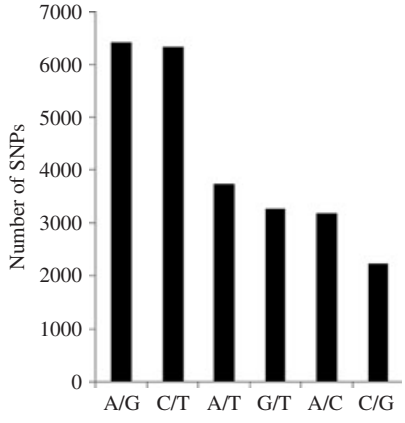

FIG. 2. Frequency distribution of single-nucleotide polymorphism types discovered in Celtic Sea *Gadus morhua* reduced representation library.

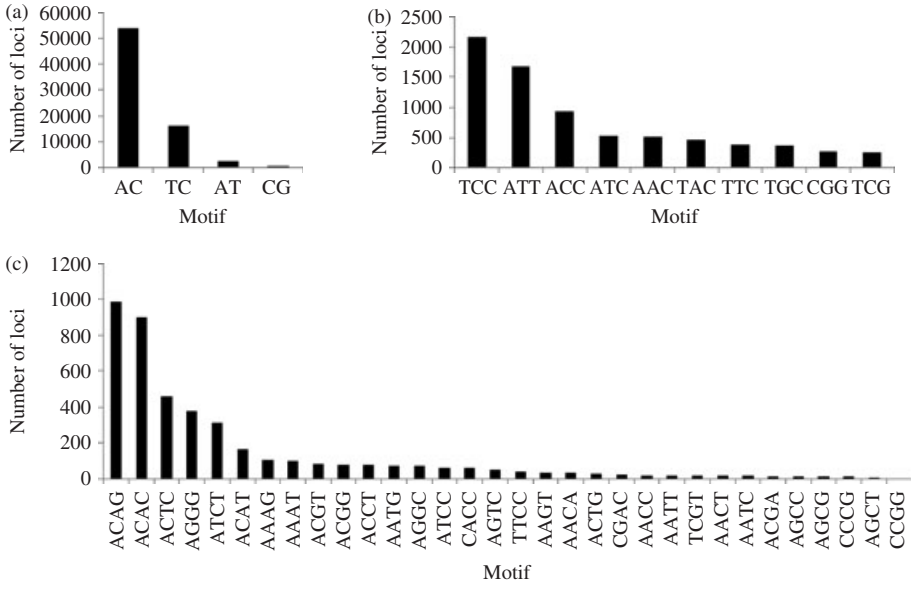

FIG. 3. Frequency distribution of motifs in (a) di-, (b) tri- and (c) tetranucleotide microsatellites discovered in the Celtic Sea *Gadus morhua*. Compound microsatellites can be represented more than once as they might have more than one motif class.

selected, it is not surprising that some loci would show 100% heterozygosity over the six individuals that were used for validation.

Twelve of the 15 chosen tetranucleotide microsatellite loci for which primers were designed amplified successfully. A total of six loci were not further pursued as three were monomorphic in the test samples of *G. morhua* from the Celtic Sea and three showed inconstant ambiguous amplifications. Hence, the six loci that were deployed

showed allelic variability that will be useful for genetic investigations of wild and reared *G. morhua* (Table II). The number of alleles in the validation panel varied from two to 14, while observed and expected heterozygosity ranged from 0·13 to 0·59 and 0·12 to 0·85, respectively (Table II). Two loci deviated significantly from Hardy–Weinberg expectations (Table II). The successful validation rate was therefore 80% (11 of 14) for SNPs and 40–60% (9 of 15 or 6 of 12) for microsatellite loci.

DISCUSSION

The SNP and microsatellite discovery pipeline described here identified over 25 000 putative SNPs and primers for *c*. 6400 potential microsatellite loci in *G. morhua* from the Celtic Sea area of the species distribution. The technique is based on a RRL and a single half-plate run on a 454 GS FLX (Roche) sequencing instrument using titanium chemistry. This approach can be used for rapid (<2 weeks of laboratory work and bioinformatic analyses), comparatively low-cost SNP and microsatellite discovery and validation in any organism, without the need for prior genetic information. While the cost of the method described here is comparable to traditional or commercial alternatives for microsatellite discovery, the approach used has the added benefit of detecting far more microsatellites and simultaneously huge numbers of SNPs. As these loci are not derived from EST sequences it is likely that the majority are selectively neutral (Vigouroux *et al*., 2002), a prerequisite for estimating demographic and population genetic parameters. The ascertainment bias purposely introduced into the marker discovery process by only using Celtic Sea *G. morhua* ensures that SNP loci will be variable in *G. morhua* at least from this area.

If individuals from only one or a few populations are used for SNP and microsatellite discovery, it will result in only a sub-set of the variable markers in the species being detected. This will lead to preferential development of SNPs showing high

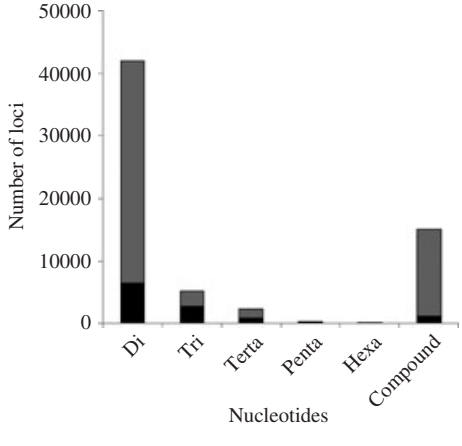

FIG. 4. Number of perfect di-, tri- tetra-, penta-, hexanucleotide and compound (both imperfect and perfect) microsatellite loci discovered (▒) and the proportion of these for which Primer3 could design primers (■) in the Celtic Sea *Gadus morhua*.

MAFs in populations for which the markers were developed and, while these markers will be variable in the discovery panel, they are not necessarily equally variable in other populations of the species (Brandström & Ellegren, 2008). Because of the potentially biased selection of markers, this approach can lead to ascertainment bias, and may reduce the ability to estimate demographic and population genetic parameters (Nielsen, 2000, 2004). In trying to minimize ascertainment bias, many SNP and microsatellite discovery studies have included individuals from many different populations representative of a broad geographical range, in order to more accurately capture the variability of these markers over an entire species (Morin *et al*., 2004; Brandström & Ellegren, 2008). This approach, however, selects markers that are variable among all populations included in the discovery panel and might generate markers that have limited variability and hence usefulness in an individual population. Although ascertainment bias is problematic for developing markers representing genetic variation on a species-wide level, it is possible to take advantage of ascertainment bias by developing a set of markers (SNPs and microsatellites) for the target populations involved in a specific study, which ensures that the markers are variable in the focal populations. The high cost of sequencing has previously limited this approach. The advent of next-generation sequencing technologies, however, has greatly reduced the cost of generating large amounts of sequencing information *de novo*. Here, the feasibility of low-cost, mid-scale marker discovery in a specific population is demonstrated in the Celtic Sea *G. morhua*. This approach should yield useful markers for genotyping this stock, and the technique is broadly applicable to other species for which existing sequences are limited. For a relatively low cost, markers can be developed for specific populations and questions. For instance, if it is desirable to have a marker set for species range-wide variability, individuals from throughout the species range can be included in the discovery panel. Similarly, if variable markers for a small sub-set of populations are needed, individuals from these populations can be included in the discovery panel.

The frequency of SNPs found in the Celtic Sea *G. morhua* by this study (one per 785 bp) fits well within the reported range for a wide range of other organisms (one per 200–1000 bp; Brumfield *et al*., 2003). The estimate of microsatellite density (one per 260 bp) is similarly consistent with previous reports, which in teleosts can range from one per 6·56 kbp in three-spined stickleback, *Gasterosteus aculeatus* L. 1758, to one per 32 bp (3·21% of the genome) of the spotted green pufferfish *Tetraodon nigroviridis* Marion de Procé 1822 (Chistiakova *et al*., 2005) (the wide range being indicative of different methods of discovery, with or without subsequent validation and also differences in genome sizes). It should, however, be noted that the number of SNPs present in *G. morhua* is probably higher than stated in this work, as some loci that are variable on a species-wide geographic scale might be monomorphic in Celtic Sea *G. morhua*.

The objective of this study was primarily to demonstrate a pipeline for fast, economical SNP and microsatellite discovery and not to validate a large number of these markers. A small number of putative SNP (11 of 14) and microsatellite (6 of 15) loci identified by the pipeline, however, were validated, demonstrating that this technique produces usable markers. In this limited sampling, a high validation success was achieved for both SNPs (80% based on SNPs with MAFs of 40% or more and the validation success could change if SNPs with lower MAFs were chosen for validation) and microsatellites (40%). Additional microsatellite markers were

validated by sequencing but were monomorphic in Celtic Sea *G. morhua*. It is possible, however, that these loci would show variability in a larger test sample or in samples from other geographical regions. In similar SNP discovery approaches, validation success has varied. For instance, Wiedmann *et al*. (2008) validated (Sequenom platform; www.sequenom.com) 168 of 176 (95·5%) putative SNPs found in the *S. scrofa* genome *via* RRL–pyrosequencing, while Maughan *et al*. (2009) successfully resequenced (Sanger sequencing) and validated 34 of 35 (97%) putative SNPs in the amaranth *Amaranthus caudatus* genome. In contrast, Castaño Sánchez *et al*. (2009) indicated substantial problems in validating SNPs from *O. mykiss* after RRL or pyrosequencing, with only 183 of 381 (48%) putative SNPs validated by Golden Gate array technology (Illumina). Castaño Sánchez *et al*. (2009) suggested that the major cause for the false positives is the evolutionary recent genome duplication that has occurred in salmonids. *Gadus morhua* have not undergone recent genome duplication as seen in salmonids (Leggatt & Iwama, 2003) and consequently are less likely to suffer from this type of false-positive result.

The validation of polymorphic microsatellites for which primers can be designed varies among species. Santana *et al*. (2009) identified microsatellite loci *via* pyrosequencing of three unrelated organisms (the fungus *Fusarium circinatum*, the insect *Sirex noctilio* and the nematode *Deladenus siricidicola*) and generated 29 primer sets, of which 13 (33%) gave polymorphic amplifications of the expected size. Hence, the validation success of 40% is similar or above to what has previously been reported. Validation of additional putative SNPs and microsatellite markers generated by this study is currently in process. These markers will be deployed in both population genetic and marker-assisted breeding programmes of Celtic Sea *G. morhua*.

The technique described here for SNP and microsatellite discovery is somewhat similar to the RAD-tag sequencing described by Baird *et al*. (2008). Both techniques take advantage of restriction enzymes and next-generation sequencing. They differ, however, in a number of aspects; RAD-tags can produce larger numbers of reads owing to use of Illumina sequencing, but the read lengths are shorter than with 454. While RAD-tags can be used for simultaneous SNP discovery and typing, the short reads from an Illumina sequencer cannot easily be used for microsatellite discovery and typing (it is unlikely that both flanking regions can be sequenced to allow for microsatellite primer design except for very short microsatellites). The technique described here allows for the detection of microsatellites, markers that will remain extremely important for applications, where they are better suited than SNPs or other markers (*e.g*. parentage and relatedness analysis). Therefore, while there are similarities between the techniques, the goals of the two approaches are different.

In order to further improve effectiveness of the described approach, it is suggested that future modifications of the protocol could take advantage of individual barcodes whereby each individual used in the discovery panel is individually tagged with a short sequence. Thus, this approach could allow for simultaneous SNP and microsatellite discovery and genotyping. The read lengths from a 454 sequencer are, in contrast to Illumina sequences, long enough to allow for detection of microsatellites and their flanking regions. Hence, microsatellite loci could be identified by their flanking regions across individuals and the size variation of the microsatellite could be detected in the sequence.

In summary, a pipeline for SNP and microsatellite discovery that utilizes reduced representation genomic libraries and next-generation sequencing was developed, and a number of SNPs and microsatellites were validated. This approach is rapid, economical and will generate thousands of SNP and microsatellite markers in any type of organism with relatively little effort. This method allows for designing markers for specific populations and questions by taking advantage of the inherent ascertainment bias that can affect marker development. The markers developed with this technique will be useful for a wide range of genetic disciplines including population genetics, phylogenetics and genetic mapping, and for identifying footprints of selection and for marker-assisted breeding programmes. In addition to the very high number of markers generated by this approach, large amounts of sequence data are generated that can be annotated and used for other genomic research.

This work was funded by the EIRCOD cod breeding and broodstock project (Grant-Aid Agreement No. PBA/AF/07/004) carried out under the Sea Change Strategy with the support of the Marine Institute and the Marine Research Sub-programme of the National Development Plan 2007–2013, co-financed by the European Regional Development Fund. J.C., J.P.C. and P.M. were supported by the Beaufort Marine Research Award in Fish Population Genetics funded by the Irish Government under the Sea Change Programme. M. Cross provided laboratory support.

References

Allendorf, F. W., Hohenlohe, P. A. & Luikart, G. (2010). Genomics and the future of conservation genetics. *Nature Reviews Genetics* **11,** 697–709.

Allentoft, M., Schuster, S. C., Hodaway, R. N., Hale, M. L., McLay, E., Oskam, C., Gilbert, M. T. P., Spencer, P., Willerslev, E. & Bunce, M. (2009). Identification of microsatellites from an extinct moa species using high-throughput (454) sequence data. *Biotechniques* **46,** 195–200.

Altshuler, D., Pollara, V. J., Cowles, C. R., Van Etten, W. J., Baldwin, J., Linton, L. & Lander, E. S. (2000). An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407,** 513–516.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A. & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3,** 3376. doi: 10.1371/journal.pone.0003376

Beaumont, M. A. & Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society B* **263,** 1619–1626.

Beebee, T. & Rowe, G. (2008). *Introduction to Molecular Ecology*, 2nd edn. Oxford: Oxford University Press.

Bentzen, P., Taggart, C. T., Ruzzante, D. E. & Cock, D. (1996). Microsatellite polymorphism and the population structure of Atlantic cod (*Gadus morhua*) in the northwest Atlantic. *Canadian Journal of Fisheries and Aquatic Sciences* **53,** 2706–2721.

Brandström, M. & Ellegren, H. (2008). Genome-wide analysis of microsatellite polymorphism in chicken circumventing the ascertainment bias. *Genome Research* **18,** 881–887.

Brumfield, R. T., Beerli, P., Nickerson, D. A. & Edwards, S. V. (2003). The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology and Evolution* **18,** 249–256.

Carlsson, J. & Reece, K. S. (2007). Eight PCR primers to amplify EST-linked microsatellites in the Eastern oyster, *Crassostrea virginica* genome. *Molecular Ecology Notes* **7,** 257–259.

Castaño Sánchez, C. C., Smith, T. P. L., Wiedmann, R. T., Vallejo, R. L., Salem, M., Yao, J. & Rexroad, C. E. III (2009). Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics* **10,** 559.

Castoe, T. A., Poole, A. W., Gu, W., DEKoning, A. P. J., Daza, J. M., Smith, E. N. & Pollock, D. D. (2010). Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequencing. *Molecular Ecology Resources* **10,** 341–347.

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A. J., Miller, W. E., Wetter, T. & Suhai, S. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research* **14,** 1147–1159.

Chistiakova, D. A., Hellemans, B. & Volckaert, F. A. M. (2005). Microsatellites and their genomic distribution, evolution, function and applications: a review with special reference to fish genetics. *Aquaculture* **225,** 1–29.

Estoup, A., Janre, P. & Cournet, J.-M. (2002). Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology* **11,** 1591–1604.

Fan, H. & Chu, J.-Y. (2007). A brief review of short tandem repeat mutation. *Genomics, Proteomics and Bioinformatics* **5,** 7–14.

Hardie, D. C. & Hebert, P. D. N. (2004). Genome-size evolution in fishes. *Canadian Journal of Fisheries and Aquatic Sciences* **61,** 1636–1646.

Hemmer-Hansen, J., Nielsen, E. G., Meldrup, D. & Mittelholzer, C. (2011). Identification of single nucleotide polymorphisms in candidate genes for growth and reproduction in a nonmodel organism; the Atlantic cod, *Gadus morhua*. *Molecular Ecology Resources* **11**(Suppl. 1), 71–80.

Heylar, S. J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M. I., Ogden, R., Limborg, M. T., Cariani, A., Males, G. E., Diopere, W., Carvalho, G. R. & Nielsen, E. E. (2011). Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources* **11,** 123–136.

Higgins, B., Hubert, S., Simpson, G., Stone, C. & Bowman, S. (2009). Characterization of 155 EST-derived microsatellites and validation for linkage mapping. *Molecular Ecology Resources* **9,** 733–737.

Hohenlohe, P. A., Amish, S. J., Catchen, J. M., Allendorf, F. W. & Luikart, G. (2011). Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources* **11**(Suppl. 1), 117–122.

Kimura, M. & Crow, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49,** 725–738.

Kimura, M. & Ohta, T. (1978). Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proceedings of the National Academy of Sciences of the United States of America* **75,** 2868–2872.

Laval, G., Patin, E., Barreiro, L. B. & Quintana-Murci, L. (2010). Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLoS One* **5,** e10284.

Leggatt, R. A. & Iwama, G. K. (2003). Occurrence of polyploidy in the fishes. *Reviews in Fish Biology and Fisheries* **13,** 237–246.

Maughan, P. J., Yourstone, S. M., Jellen, E. J. & Udall, J. A. (2009). SNP discovery via genomic reduction, barcoding and 454-pyrosequencing in amaranth. *Plant Genome* **2,** 260–270.

Moen, T., Hayes, B., Nilsen, F., Delghandi, M., Fjalstad, K. T., Fevolden, S.-E., Berg, P. R. & Sigbørn, L. (2008). Identification and characterisation of novel SNP markers in Atlantic cod: evidence for directional selection. *BMC Genetics* **9,** 18.

Morin, P. A., Luikart, G. & Wayne, R. K. (2004). SNPs in ecology, evolution and conservation. *Trends in Ecology and Evolution* **19,** 208–216.

Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154,** 931–942.

Nielsen, R. (2004). Population genetic analysis of ascertained SNP data. *Human Genomics* **3,** 218–224.

Ning, Z., Cox, A. J. & Mullikin, J. C. (2001). SSAHA: a fast search method for large DNA databases. *Genome Research* **11,** 1725–1729.

van Oosterhout, C., Hutchinson, W. F., Wills, D. P. M. & Shipley, P. (2004). Micro-checker: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes* **4,** 135.

Rice, W. R. (1989). Analysing tables of statistical tests. *Evolution* **43,** 223–225.

Rousset, F. (2008). Genepop'007: a complete re-implementation of the Genepop software for Windows and Linux. *Molecular Ecology Resources* **8,** 103–106.

Rozen, S. & Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology* **132,** 365–386.

Santana, Q. C., Coetzee, M. P. A., Steenkamp, E. T., Mlonyeni, O. X., Hammond, G. N. A., Wingfield, M. J. & Wingfield, B. D. (2009). Microsatellite discovery by deep sequencing of enriched genomic libraries. *Biotechniques* **46,** 217–233.

Schuelke, M. (2000). An economic method for the fluorescent labelling of PCR fragments. *Nature Biotechnology* **18,** 233–234.

Star, B., Nederbragt, A. J., Jentoft, S., Grimholt, U., Malmstrøm, M., Gregers, T. F., Rounge, T. B., Paulsen, J., Solbakken, M. H., Sharma, A., Wetten, O. F., Lansén, A., Winer, R., Knight, J., Vogel, J.-H., Aken, B., Andersen, Ø., Lagesen, K., Tooming-Klunderud, A., Edvardsen, R. B., Tina, K. G., Espelund, M., Nepal, C., Previt, C., Karlsen, B. O., Moum, T., Skage, M., Berg, P. R., Gjøen, T., Kuhl, H., Thorsen, J., Malde, K., Reinhardt, R., Du, E., Johansen, S. D., Searle, S., Lien, S., Nilsen, F., Jonassen, I., Omholt, S. W., Stenseth, N. C. & Jakobsen, K. S. (2011). The genome sequence of Atlantic cod reveals a unique immune system. *Nature* **477,** 207–210.

Van Tassell, C. P., Smith, T. P. L., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., Haudenschild, C. D., Moore, S. S., Warren, W. C. & Sonstegard, T. S. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods* **5,** 247–252.

Vasemägi, A. & Primmer, C. R. (2005). Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Molecular Ecology* **14,** 3623–3642.

Vigouroux, Y., McMullen, M., Hittinger, C. T., Houchins, K., Schulz, L., Kresovich, S., Matsuoka, Y. & Doebley, J. (2002). Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 9650–9655.

Wiedmann, R. T., Smith, T. P. L. & Nonneman, D. J. (2008). SNP discovery in swine by reduced representation and high throughput pyrosequencing. *BMC Genetics* **9,** 81.

Zane, L., Bargelloni, L. & Patarnello, T. (2002). Strategies for microsatellite isolation: a review. *Molecular Ecology* **11,** 1–16.