A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding


Published in:
Royal Society Open Science

Document Version:
Publisher's PDF, also known as Version of record

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
© 2016 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Open Access
This research has been made openly available by Queen's academics and its Open Research team. We would love to hear how access to this research benefits you. – Share your feedback with us: http://go.qub.ac.uk/oa-feedback

Download date: 18. Oct. 2023
This study examines the potential of next-generation sequencing based ‘genotyping-by-sequencing’ (GBS) of microsatellite loci for rapid and cost-effective genotyping in large-scale population genetic studies. The recovery of individual genotypes from large sequence pools was achieved by PCR-incorporated combinatorial barcoding.
using universal primers. Three experimental conditions were employed to explore the possibility of using this approach with existing and novel multiplex marker panels and weighted amplicon mixture. The GBS approach was validated against microsatellite data generated by capillary electrophoresis. GBS allows access to the underlying nucleotide sequences that can reveal homoplasy, even in large datasets and facilitates cross laboratory transfer. GBS of microsatellites, using individual combinatorial barcoding, is potentially faster and cheaper than current microsatellite approaches and offers better and more data.

1. Introduction

The advent of next-generation sequencing (NGS) technologies has fundamentally changed how genetic sequence data are generated [1]. While NGS was primarily introduced to substantially increase sequence yield for genome projects [2–4], it has in addition enabled high-throughput genotyping that can be used for genetic studies, including population genetics, by using a range of protocols (e.g. RAD [5], ddRAD [6], 2bRAD [7]). This is collectively known as ‘genotyping-by-sequencing’ (GBS) [8,9].

The primary advantage of GBS for population genetic studies is the generation of increased quantities of data that allows for improved statistical power and high genome representation [9]. The concurrent development of single-nucleotide polymorphism (SNP) genotyping platforms (e.g. SNP-chips [10,11], microfluidic TaqMAN assays [12]) and the persistent problems associated with microsatellite genotyping has led to a shift from using microsatellites to SNPs as the preferred marker for genetic studies [13]. The main problems with capillary and gel-based microsatellite studies include fragment size homoplasy, poor levels of inter-laboratory calibration, the genotype not including the underlying sequence information and inherently laborious genotyping [14,15]. However, there are also unresolved problems regarding SNPs, such as ascertainment bias, transferability among SNP genotyping platforms and the requirement for high template DNA quality [16–18]. While SNPs do not suffer many of the issues associated with microsatellite genotyping, the major advantage of microsatellites over SNP-based approaches for population analyses is the higher statistical power per locus [19–21]. Additionally, microsatellites are preferred to SNPs in forensic, parentage and kinship studies owing to their higher mutation rates and polyallelic nature [22–25].

Many of the issues associated with microsatellite-based population studies could be mitigated using a GBS approach. The effects of size-homoplasy [26] can be effectively eliminated because the genotype incorporates the underlying sequence information. The difficulties of inter-laboratory calibration would be significantly reduced, as GBS considers actual base-pair lengths of the alleles, and not the estimated allele size sequence length [15]. Additionally, the elimination of time consuming capillary or gel electrophoresis runs [27] can reduce genotyping time, thereby increasing efficiency. Further, GBS has the potential to use established microsatellite multiplex panels, enabling calibration with existing datasets and facilitating inter-laboratory collaboration. However, few GBS studies based on microsatellites have been published [28–30].

The current study aims to assess the potential of microsatellite GBS using Atlantic cod Gadus morhua (L.) as the study organism. The primary objective was to examine the potential of microsatellite GBS using de novo and existing capillary/gel electrophoresis-based multiplex marker panels. The secondary objective was to develop a rapid and cost-effective method for microsatellite GBS, that can take advantage of modern NGS platforms, using combinatorial barcoding for implementation in large-scale population genetics studies.

2. Material and methods

A total of 64 Atlantic cod were collected from the Celtic Sea (n = 32) and the West of Ireland (n = 32) in 2011. Tissue samples were preserved in molecular grade ethanol. DNA was extracted using a Chelex® protocol [31]. DNA was quantified using a NanoDrop 1000 Spectrophotometer (Thermo Scientific, Wilmington, DE, USA) and diluted to a concentration of 50 ng µl⁻¹. These samples had previously been genotyped in six original PCR multiplexes at 53 microsatellite loci using a conventional ABI capillary-based approach [32] thus providing a platform for direct comparison with the current study.

Six multiplex PCR panels comprising 53 microsatellite loci [32] were analysed to test the capability of microsatellite GBS on all individuals. Each of the six original multiplex panels included loci in
three different size classes (class I: 115–213 bp, class II: 203–320 bp and class III: 265–416 bp, see the electronic supplementary material, table S1). To assess the preferential amplification of shorter fragments during PCR and NGS sequencing [30,33], weighted ratio mixtures of fluorescently labelled amplicons were visualized on an ABI 3130xl Genetic Analyzer (Applied Biosystems; conditions according to [32] suggesting an optimal size-class amplicon-ratio of 1(I) : 1(II) : 8(III).

Three PCR size-class-combinations for GBS were evaluated (electronic supplementary material, table S1). The first de novo multiplex panels were weighted for amplicon size by combining all markers of the same size classes (3-PCR) from the six original multiplexes using a subset of 16 individuals from the Celtic Sea. The second size-class-combination consisted of the existing multiplexes (6-PCR) using the same subset of 16 individuals from the Celtic Sea (sensu [32]). The third PCR size-class-combination comprised the six existing multiplex panels, subdivided into three smaller multiplex panels according to the three size classes using all 64 individuals (18-PCR).

The three-primer PCR approach [32,34] was modified into a four-primer PCR to produce unique individual combinatorial barcoded amplicons suitable for pooled amplicon NGS in population genetics studies. Each barcode comprised 10 bp of synthesized DNA sequence [35]. The modified protocol consisted of locus-specific primers (forward and reverse) that were adapted to include universal primer sequence (figure 1; electronic supplementary material, table S2). Two barcoded universal primers were included to incorporate two barcodes into each of the resulting amplicons. In total, 12 forward and eight reverse DNA barcodes allowed for the recovery of 96 unique individual combinatorial barcodes (electronic supplementary material, table S3).

Multiplex PCRs (cycling condition as in [32]) were performed in 10 µl reactions (size classes I and II) and 20 µl reactions (size class III, in 3-PCR and 18-PCR size-class-combinations) with 1 µl template DNA, 1 x Multiplex PCR Master Mix (Qiagen, Hilden, Germany), 2 pmol of each locus-specific primer and 8 pmol of each barcoded universal primer per locus-specific primer with the corresponding universal primer sequence. For each multiplex and PCR size-class-combination equal amounts of PCR product from each individual were pooled and separated by gel electrophoresis on a 2% agarose gel at 5 V cm⁻¹. PCR products were visualized using Safe Imager® (Invitrogen, Life Technologies, Carlsbad, CA, USA). Products corresponding to expected amplicon sizes were cut out from the gel and placed into 2 ml centrifuge tubes. DNA was extracted using the QIAquick® Gel Extraction Kit (Qiagen, Hilden, Germany). Amplicon DNA was quantified using Qubit® 2.0 fluorometer with a dsDNA HS Assay kit (Life Technologies). Amplicons from the 3-PCR and 18-PCR size-class-combinations were normalized to the optimal size-class amplicons-ratio (1 : 1 : 8). The amplicons were concentrated using the Amicon® Ultra 30 K procedure (Merck Millipore Ltd., Merck KGaA, Darmstadt, Germany). An aliquot of the final library was diluted to a concentration of 500 ng in 16 µl for sequencing on a 454 Roche® platform ([36]; 454 Life Sciences Corp., a Roche Company, Branford, CT, USA).

Roche adaptors were added to the amplicon library using the Rapid Library Preparation Kit [37]. Amplicon concentrations were estimated using quantitative PCR to assess differential amplification success depending on amplicon size. Amplicons were sequenced on two 1/4 regions of a Pico TiterPlate using a 454 GS-FLX Sequencer with titanium reagents at Teagasc Food Research Centre, Moorepark.
Ireland. The Shotgun Data Processing Pipeline was used for signal processing to increase sequence yield [38]. A python script was developed to process raw sequence data by identifying sequence reads containing the forward and reverse (combinatorial) barcodes and the locus-specific primers and grouping them accordingly. The script was based on the Levenshtein distance metric which measures the distance between two sequences of characters. The error rate of Roche 454 GS-FLX amplicon sequences when including both sequencing and PCR errors has been estimated in *G. morhua* at 6% [39]. It was therefore necessary to allow for sequencing and PCR errors in the reads in order to avoid a significant loss of reads during the identification and grouping process. The python script allowed for up to two and three sequencing errors (both substitutions and indels) in the combinatorial barcodes and primers, respectively, to reflect a potential sequencing and PCR error rate of approximately 6%. All the scripts used are available on github (https://github.com/egenomics/micomba).

Grouped and classified sequences were imported into GENEIOUS® 7 as fasta files and organized into folders per locus per individual. Loci were manually genotyped by viewing all of the reads of a particular individual at a specific locus, as a read length histogram and verified by read alignment (GENEIOUS ALIGNMENT—default settings) and manually edited [40]. Only individuals with five or more reads for a given locus were genotyped.

Read alignments of a subset of 16 individuals, with 10 or more sequence reads, from the Celtic Sea dataset (six PCR size-class-combinations) were screened for homoplasy. The aligned sequences were scrutinized for the presence of SNPs and indels that were not part of the microsatellite repeat structure and would not change the amplicon size. The SNP or indel had to occur in at least 20% of the reads to be considered as homoplasy.

Correspondence between GBS and ABI capillary-based genotyping data [32] was fitted to a binomial model [41] (electronic supplementary material, equation S1). The impact of several factors on GBS–ABI correspondence was evaluated. These factors included microsatellite type (e.g. mono-, di-), PCR size-class-combination (3-, 6- or 18-PCR), and read depth used for genotyping (broken into 2 and 3 read depth categories). Model fit was compared using the finite-sample Akaike information criterion (AIC) [42–46] (electronic supplementary material, equations S3 and S4).

Binomial proportions of the correspondence data were modelled for a restricted subset of the data, 16 individuals from the Celtic Sea (‘Celtic Sea’ dataset), for which data existed across all PCR size-class-combinations. Differences in correspondence proportions among PCR size-class-combinations were evaluated using Bonferroni-corrected z-tests [47]. The effectiveness of PCR size-class-combinations on the total number of reads produced for the Celtic Sea data was then evaluated. Total reads were modelled as a function of PCR size-class-combination, locus-specific primer and universal primer type, in addition to potential intrinsic confounding variables, such as inter-locus or inter-individual variability. A multinomial probability likelihood model [41] (electronic supplementary material, equation S2) was employed to evaluate model support using AIC. Bonferroni-corrected Mann–Whitney tests [47] were performed on total, median and maximum number of reads per individual between the PCR size-class-combinations to assess significant differences in the performance of PCR size-class-combinations.

### 3. Results

Sequencing of the two 1/4 regions of a PicoTiterPlate resulted in 228,246 reads for region 1 and 226,848 for region 2. Read length ranged from 54 to 1200 bp with an average read length of 275 bp (s.d. 84 bp). A total of 95.5% of reads had a quality score higher than Q20 and the average quality score was Q35.9. The proportions of A, C, G and T nucleotides were 27.0%, 22.7%, 26.7% and 23.6%, respectively, with a GC content of 49.4%.

A total of 180,054 reads were successfully assigned to a specific individual and locus combinations (electronic supplementary material, table S4). The recovered reads per individual (combined across all loci) ranged from 57 to 9192 (median: 999) (electronic supplementary material, table S3), and the number of reads per locus (combined over all individuals) ranged from 33 to 11,197 (median: 2294). The numbers of reads per genotype ranged from 0 to 1097 (median: 12). Analyses of PCR size-class-combinations resulted in 3325 successful genotypes from 5088 possible calls, with 3196 genotypes used for correspondence checks with the ABI-capillary genotyping data. Of the 53 loci analysed, 10 loci had low numbers of sequence reads preventing genotyping in more than 50% of the individuals (electronic supplementary material, table S5).
In total, 529 genotypes were screened for homoplasy, which was detected in 32% of the genotypes and was present in 38 loci. SNPs represented 80% of the homoplasy with 20% represented by indels.

AICc model selection for the Celtic Sea data demonstrated that PCR size-class-combination was the only important variable for predicting read yield. There was no support for other modelled variables (e.g., forward or reverse primers) or confounding variables (such as inter-individual and among-locus variation; table 1). Correspondence between the full GBS and ABI datasets was best explained by a model incorporating read depth and microsatellite motif type (table 2). PCR size-class-combination had an effect on the GBS–ABI correspondence in the Celtic Sea data (electronic supplementary material, table S6), with significantly higher correspondence for both 3-PCR and 6-PCR when either is compared with 18-PCR. However, there was no significant difference between 3- and 6-PCR (electronic supplementary material, table S6). In addition, significantly more reads per individual were observed for 3- and 6-PCR, although again, there was no significant difference between the two (electronic supplementary material, table S7). No significant differences were noted for median number of reads among size-class-combinations (electronic supplementary material, table S8). Maximum number of reads per individual showed a significant difference between 6-PCR and 18-PCR only (electronic supplementary material, table S9).

4. Discussion

This study demonstrates the potential for NGS-based GBS as a method for microsatellite genotyping using de novo and existing capillary/ gel electrophoresis-based multiplex marker panels. It also illustrates the potential for a rapid and cost-effective method for microsatellite GBS, that can take advantage of modern NGS platforms, using combinatorial barcoding for implementation in large-scale population genetics studies. This method provides access to the underlying sequence data, providing an additional advantage over traditional fragment length genotyping by resolving issues of size homoplasy and revealing potentially hidden genetic variation in the amplicons.

Analysis of the Celtic Sea data indicates that size-class combination is the controlling factor for both sequence yield and GBS–ABI correspondence. Both the 3- and 6-PCR size class conditions outperformed 18-PCR, with no significant differences between the 3- and the 6-PCR conditions observed. It is possible that 6-PCR would not have performed as well as 3-PCR in the absence of pre-optimization [32]. However, the increased number of PCR reactions and DNA quantification steps in the 18-PCR method may have introduced quantification errors that were amplified at the pooling stage, lowering read yield, and hence correspondence. As such, it is proposed that future studies minimize the number of steps to reduce variability among PCRs.

Binladen et al. [48] reported biases in sequence recovery with 454 NGS as a result of different base pair composition of barcodes. In contrast to this, the current study found no bias in sequence
polymorphism is time consuming, and therefore carries a significant financial cost to genetic studies. For deployment of the approach described here, in a population genetics setting, it may be more expedient to use a larger number of initial loci to increase the likelihood of generating a sufficient number of informative genotypes.

When employing other NGS platforms this requirement may be alleviated by the increased read yield of these platforms. The current study was not optimized for population genetics scale genotyping as a larger number of loci and a lower number of individuals were analyzed than is often used in population genetic studies (cf. [32]). However, in an effort to address multiple questions about the feasibility of GBS of microsatellites, it was judged advantageous to include a larger number of loci than needed for many population genetic studies. In addition, if this method was used on de novo developed loci, the success rate from microsatellite containing sequences to genotyped loci is only known post hoc and hence a larger number of initial loci would increase the likelihood of generating a sufficient number of informative loci. For deployment of the approach described here, in a population genetics setting, it may be more beneficial to interrogate fewer loci, but more individuals. This could be facilitated by increasing the number of forward and reverse combinatorial barcodes used to tag an increased number of individuals.

Manual scoring of conventional capillary and gel-based electrophoresis fragment length polymorphism is time consuming, and therefore carries a significant financial cost to genetic studies.

### Table 2. Model selection using AICc on the correspondence between 454 microsatellites and ABI microsatellites, using the full dataset. (K is the number of parameters estimated for a given model structure, LogL is the log-likelihood of the model, AICc is the finite-sample AIC score for the model, dAICc is the difference in AICc score between the given model and the optimal model score, and Post. Prob is the model posterior probability.)

<table>
<thead>
<tr>
<th>model comments</th>
<th>K</th>
<th>LogL</th>
<th>AICc</th>
<th>dAICc</th>
<th>Post. Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>no effects average over all data</td>
<td>1</td>
<td>−1724.64</td>
<td>3451.28</td>
<td>131.96</td>
<td>2.22 × 10−29</td>
</tr>
<tr>
<td>no. reads (2) 0–5 versus 5 + reads</td>
<td>2</td>
<td>−1693.93</td>
<td>3391.87</td>
<td>72.55</td>
<td>1.77 × 10−16</td>
</tr>
<tr>
<td>no. reads (3) 0–5 versus 5–10 versus 10 + reads</td>
<td>3</td>
<td>−1676.04</td>
<td>3358.10</td>
<td>38.78</td>
<td>3.80 × 10−9</td>
</tr>
<tr>
<td>MST type reads (2) by MST type 0–5 versus 5 + reads</td>
<td>5</td>
<td>−1693.54</td>
<td>3397.11</td>
<td>77.79</td>
<td>1.29 × 10−12</td>
</tr>
<tr>
<td>reads (3) by MST type 0–5 versus 5–10 versus 10 + reads</td>
<td>10</td>
<td>−1663.76</td>
<td>3347.59</td>
<td>28.27</td>
<td>7.26 × 10−7</td>
</tr>
<tr>
<td>PCR (3,6) v. 18 reads (3) by MST type 0–5 versus 5 + reads</td>
<td>15</td>
<td>−1644.58</td>
<td>3319.32</td>
<td>0.00</td>
<td>0.999</td>
</tr>
<tr>
<td>PCR 3,6,18 reads (2) by PCR 3,6,18 0–5 versus 5 + reads</td>
<td>2</td>
<td>−1706.09</td>
<td>3416.18</td>
<td>96.86</td>
<td>9.27 × 10−22</td>
</tr>
<tr>
<td>reads (3) by PCR 3,6,18 0–5 versus 5–10 versus 10 + reads</td>
<td>3</td>
<td>−1675.28</td>
<td>3362.58</td>
<td>43.26</td>
<td>4.03 × 10−10</td>
</tr>
<tr>
<td>reads (3) by PCR 3,6,18 0–5 versus 5–10 versus 10 + reads</td>
<td>6</td>
<td>−1658.37</td>
<td>3334.79</td>
<td>15.47</td>
<td>0.0004</td>
</tr>
</tbody>
</table>
This study describes an approach that does not require manual correction of internal size standards or genotype calls owing to spectral bleed-through, thus reducing the genotyping time. In addition, the availability of the underlying sequence data lends itself to the development of automated genotyping [55]. The numbers of loci that can be multiplexed in capillary or gel-based electrophoresis in conventional studies are limited by the availability of fluorescent labels/detection channels and fragment size overlaps (rarely more than 12 loci per multiplex) [56]. The GBS approach described in this study using PCR-incorporated combinatorial barcoding has no limitations on the number of markers (other than number of sequences produced by the chosen NGS platform), as size overlap does not affect sequence yield.

The high proportion of homoplasy (32%) observed is in concordance with other studies addressing the prevalence of homoplasy [57–60]. This homoplasy would have been undetectable using traditional ABI-based microsatellite genotyping as the amplicon size would be unaffected. Previous detection methods include single-strand conformation polymorphism analysis and direct sequencing, with or without cloning, however, these detection methods can be laborious [57–60]. It should be noted that low read depth may prevent distinguishing homoplasy from true mutations and sequencing/PCR error. However, the purpose of the analysis in the current study was to explore the capacity to detect and quantify prevalence of homoplasy using GBS-based methods. The prevalence of homoplasy implies that allelic diversity in fragment size-based studies is probably substantially underestimated. Consequently, inferred population structures in these studies may also underestimate true levels of genetic variability. Aside from homoplasy, determining microsatellite repeat numbers could improve genetic diversity comparisons [61]. Essentially, accessing the molecular structure of the microsatellite markers will increase our understanding of the mutation model of the studied loci and thus improve the quality of the information retrieved from the data. In addition, access to actual sequence lengths and sequence information will greatly facilitate inter-laboratory calibration and data storage in repositories as sequence files offering a significant advantage.

This study presents a novel method for microsatellite GBS using individual combinatorial barcoding that can be faster and cheaper than current approaches while offering better and more data.
and F.C. provided expertise in the sequencing library design and F.C. carried out the sequencing. J.C. conceived the study, designed the study and helped draft the manuscript. All authors participated in writing of the manuscript and gave final approval for publication.

Competing interests. We declare we have no competing interests.

Funding. S.V. and R.F. acknowledge funding from 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 (Grant-Aid Agreement no. PBA/AF/07/004, ERICROD). J.L.V.-C. acknowledges funding from the Spanish Ministerio de Economía y Competitividad (FPI BES-2010-038494 and EEBB-I-13–06270). G.H. acknowledges funding by the Irish Research Council (IRC) Graduate Research Education Programme (GREP). P.M. and J.C. acknowledge funding from the Beaufort Marine Research Award in Fish Population Genetics funded by the Irish Government under the Sea Change Programme. P.C. acknowledges funding from Science Foundation Ireland (SFI 12/IP/1308). E.D.F. acknowledges funding from Irish Research Council (Funding Agency Ref no. GOIPD/2013/320).

Acknowledgements. The authors would like to acknowledge the staff at Carna Research Station who acquired the Celtic Sea G. morhua samples and John Brittain who acquired the West of Ireland samples. The authors would also like to thank the referees for their constructive input.

References


3. Oh JD et al. 2006 The complete genome sequence of a chronic atrophic gastritis Helicobacter pylori strain: evolution during disease progression. Proc Natl Acad Sci USA 103, 9999–10 004. (doi:10.1073/pnas.0602784103)


35. 2009 Using Multiplex Identifier (MID) adaptors for the GS FLX titanium chemistry - extended MID set. 454 Sequencing Technical Bulletin No. 005 – 2009, 454 Life Sciences Corp., a Roche Company, Branford, CT, USA.


38. 2012 454 Sequencing system guidelines for amplicon experimental design: Branford, CT, 454 Life Sciences Corp., a Roche Company.


51. 2013 The Ion PGM™ System, with 400-base read length chemistry, enables routine high-quality de novo assembly of small genomes. Life Technologies Corporation.


