Less is more: extreme genome complexity reduction with ddRAD using Ion Torrent semiconductor technology

LILIAN PUKK,* FREED AHMAD,† SHIHAB HASAN,** VELJO KISAND,‡ RIHO GROSS* and ANTI VASEMÄGI**

*Department of Aquaculture, Estonian University of Life Sciences, Tartu 51006, Estonia, †Department of Information Technology, University of Turku, Turku 20014, Finland, ‡Bioinformatics Lab, QIMR Berghofer Medical Research Institute, Brisbane, Qld, Australia, §§chool of Medicine, The University of Queensland (UQ), Brisbane, Qld 4006, Australia, ¶Institute of Technology, University of Tartu, Tartu 50411, Estonia, **Department of Biology, University of Turku, Turku 20520, Finland

Abstract

Massively parallel sequencing a small proportion of the whole genome at high coverage enables answering a wide range of questions from molecular evolution and evolutionary biology to animal and plant breeding and forensics. In this study, we describe the development of restriction-site associated DNA (RAD) sequencing approach for Ion Torrent PGM platform. Our protocol results in extreme genome complexity reduction using two rare-cutting restriction enzymes and strict size selection of the library allowing sequencing of a relatively small number of genomic fragments with high sequencing depth. We applied this approach to a common freshwater fish species, the Eurasian perch (Perca fluviatilis L.), and generated over 2.2 MB of novel sequence data consisting of ~17 000 contigs, identified 1259 single nucleotide polymorphisms (SNPs). We also estimated genetic differentiation between the DNA pools from freshwater (Lake Peipus) and brackish water (the Baltic Sea) populations and identified SNPs with the strongest signal of differentiation that could be used for robust individual assignment in the future. This work represents an important step towards developing genomic resources and genetic tools for the Eurasian perch. We expect that our ddRAD sequencing protocol for semiconductor sequencing technology will be useful alternative for currently available RAD protocols.

Keywords: de novo assembly, DNA pooling, Eurasian perch (Perca fluviatilis L.), next-generation sequencing, population genomics, SNP development and validation

Received 24 January 2014; revision received 9 February 2015; accepted 13 February 2015

Introduction

As the introduction of next-generation sequencing (NGS) technologies, whole genome sequencing has been carried out on an increasing number of species, triggering a major breakthrough in genetics. For many studies, however, obtaining the complete genome sequence from a large number of individuals still remains prohibitively expensive and is not even necessary. Therefore, several genome complexity reduction methods which allow only a subset of the genome to be sequenced have been recently described (Davey et al. 2011). These methods rely on the generation of reduced representation libraries (RRL), generated by restriction enzymes (Alshuler et al. 2000), selective amplification (van Orsouw et al. 2007), targeted amplicon sequencing (Hyten et al. 2010), capture probes (Shen et al. 2013), methylation filtering (Palmer et al. 2003), high-Ct selection (Yuan et al. 2003) or by sequencing the transcribed proportion of the genome, that is transcriptome (Barbazuk et al. 2007).

One of the most widely used genome complexity reduction methods that have been applied in both model (e.g. humans, laboratory mice and Drosophila) and non-model species is restriction-site associated DNA (RAD) sequencing. This method targets DNA sequences flanking specific restriction enzyme cutting sites throughout the genome (Baird et al. 2008; Hohenlohe et al. 2010). To date, RAD sequencing has been applied in various research fields ranging from molecular evolution and evolutionary biology (Emerson et al. 2010) to animal and plant breeding (Yang et al. 2012), conservation (Sharma et al. 2012) and forensics (Ogden et al. 2013). Importantly, RAD sequencing is highly suitable for generating genome-wide genotype data in situations where not much is known about the target genome. As a result, RAD sequencing has been increasingly used in population genetic, biogeographic and phylogenetic studies of
nonmodel organisms providing a genome-wide view on the role of different evolutionary forces during population differentiation, adaptation and speciation (Baxter et al. 2011; Bruneaux et al. 2013; Keller et al. 2013).

Earlier RAD library preparation protocols typically consisted of the following steps: restriction enzyme digestion, ligation of the first adapter, physical shearing, end-repair, ligation of the second adapter and size selection. However, more recent studies have simplified, and further developed, the library preparation procedure by eliminating physical shearing and introducing two restriction enzyme digestion with strict size selection (termed double RAD (Bruneaux et al. 2013) or double-digest RAD (Peterson et al. 2012)). These modifications enable greater flexibility in targeting the optimal number of genomic regions ranging from thousands of consensus sequences required for QTL mapping to hundreds of thousands of contigs necessary for association studies and hitchhiking mapping.

Most of the RAD studies to date have been carried out using Illumina GAII and HiSeq NGS platforms, but to fully exploit the potential of RAD sequencing, other NGS technologies can be used. For example, the new semiconductor sequencing technology (Ion Torrent PGM) represents a simple, fast (hours instead of days) and flexible solution for small laboratories and individual research groups, but currently, only a few studies have reported the utility of Ion Torrent PGM technology for RAD sequencing (Mascher et al. 2013; Kai et al. 2014).

This can be partially explained by the relatively small sequencing throughput of PGM technology, which produces tens of millions of short reads compared to, for example, hundreds of millions of sequences generated by the Illumina HiSeq platform. However, for RAD projects which require sequencing of a smaller number of genomic regions, such as QTL mapping and basic population genetic analysis, Ion Torrent PGM technology is more accessible to smaller research groups making it a faster and simpler alternative compared to other NGS platforms. Therefore, it is important to further develop and evaluate RAD sequencing protocols adopted for Ion Torrent PGM that enable high genome complexity reduction by targeting a relatively small number of genomic regions while retaining the high coverage necessary for robust SNP discovery and reliable estimation of various population genetic parameters.

In this study, we describe the development of a two-enzyme RAD sequencing approach for Ion Torrent PGM platform by carrying out extreme genome complexity reduction using rare-cutting restriction enzymes and strict size selection of the library while using DNA pooling as a cost-effective approach to estimate allele frequencies on a genome-wide scale. We applied this approach to the common freshwater fish species, the Eurasian perch (Perca fluviatilis L.). Eurasian perch contributes significantly to both commercial and recreational fisheries in the coastal Estonian waters of the Baltic Sea and in Lake Peipus, the fifth largest lake in Europe (Moran 2003; Pukk et al. 2013). Because of different fishing regulations in these water bodies, the ability to use genetic markers for assigning individuals back to their source population would be of considerable importance for the identification of illegal fishing and fish trade. To develop genomic resources and genetic tools for the Eurasian perch, which could be used in a court of law, we: i) developed a restriction-site associated DNA (RAD) sequencing approach for Ion Torrent PGM platform which results in extreme genome complexity reduction allowing sequencing of a relatively small number of genomic fragments with high sequencing depth; ii) characterized a small proportion of the genome and identified more than a thousand single nucleotide polymorphisms (SNPs) in perch; iii) estimated genetic differentiation between the DNA pools constructed based on individuals collected from freshwater (Lake Peipus) and brackish water (the Baltic Sea) environments; and iv) identified SNPs with the strongest signal of differentiation which could be used for individual assignment in the future.

Materials and methods

DNA isolation and ddRAD library preparation

Seventy-six wild Eurasian perch (Perca fluviatilis L.) were used for ddRAD library preparation. Genomic DNA was isolated from fin clips and dried scale samples as described by Pukk et al. (2013). These samples were divided into two pools: the Baltic Sea pool (Turku Bay, \( n = 33 \); Matsalu Bay, \( n = 6 \) and Pärnu Bay, \( n = 11 \)) and Lake Peipus pool (\( n = 26 \)). DNA quality and concentration were assessed by agarose gel electrophoresis and with a NanoDrop spectrophotometer (Thermo Fisher, Inc.). RAD library preparation protocol broadly followed the methods described by Bruneaux et al. (2013), with some modifications as outlined by Pukk et al. (2014). In short, pooled DNA (800 ng) was digested for 2 h at 37°C in 20 μL reaction volume, simultaneously with two restriction enzymes, 20 U of PstI (restriction site 5’ CTGCAG 3’) and BamHI (restriction site 5’ GGATCC 3’) (New England Biolabs) and heat-inactivated for 15 min at 75°C (Fig. 1). The ligation consisted of forward (0.016 pmol) and reverse (0.004 pmol) adapters which were added to 20 μL of restriction reaction together with 5 μL 10 x T4 DNA Ligase buffer and 1200 U of T4 DNA ligase (M0202S; NEB) (forward adapter, top: 5’ CCATCTCATCCCTGCTGTCTCCGACTCAGXXXXXTCGAC 3’, forward adapter, bottom: 5’ XXXXCTGAGTCCGGAGA CACGCAGGATGAGTGG 3’, where XXXX is a bar-
code sequence; reverse adapter, top: 5’ GATCATCA CCGACTGCCCATAGAGAGG 3’, reverse adapter, bottom: 5’ CCTCTCTATGGGCAGTCGGTGAT 3’. To differentiate between the pools, two following 5 bp barcodes were used (Sea pool: 5’ AGAAC 3’ and Lake pool: 5’ TCGTT 3’). To avoid integration of an extra nucleotide at the end of the restriction enzyme (BamHI) cutting site. Second, after filtering was carried out to exclude contigs that most likely contain multiple loci. First, using TABLET version 3.9.15 (Chevreux et al. 1999) with job specified as genome, de novo and accurate in manifest file using Ion Torrent specific settings. After de novo assembly, the following filtering was carried out to exclude contigs that most likely contain multiple loci. First, using TABLET version 1.13.07.31 (Milne et al. 2013), all the contigs with more than a 2.5% mismatch were discarded. This threshold was set based on the overall distribution of mismatch percentage from the whole data set. Similarly to mismatch percentage, we evaluated the distribution of sequencing depth of all contigs and excluded contigs with more than 500× coverage from the subsequent analysis to avoid repetitive sequences. To avoid mis-assembled contigs due to chimeric reads, the remaining contigs were split into two or three shorter contigs, depending upon the number of uncut restriction sites found. The contigs consisting of 40 or more nucleotides were selected as reference for mapping in MIRA. For mapping, two files containing population-specific reads were used. The longer reads containing uncut motifs were split from the point of restriction site into shorter (> 20 bp) reads. The job mentioned in the manifest file for MIRA was genome, mapping and accurate, and the chimeric read clipping option was also used. The BAM file produced by MIRA was then used as an input for
SAMTOOLS v.0.1.18 (Li et al. 2009) for SNP discovery. Potential SNPs were accepted only if they met the following stringent selection criteria: read alignment depth of \( \leq 500 \times; \geq 5 \) reads per SNP allele; \( \leq 2\% \) of polymorphic sites per contig; quality threshold \( \geq 20 \); and \( \leq 2\% \) of mismatch. To avoid incorrect SNP calling caused by frequent homopolymer sequencing errors of the Ion Torrent PGM platform, SNPs associated with three or more consecutive homopolymer were discarded.

**Homopolymer sequencing**

To avoid incorrect SNP calling caused by frequent homopolymer sequencing errors of the Ion Torrent PGM platform, SNPs associated with three or more consecutive homopolymer were discarded.

**Contig annotation and categorization**

RAD contigs were annotated using a BLAST v2.2.28 (Altschul et al. 1990) (BLASTX – against a Uniprot database and BLASTN – against a nucleotide database) with the E-value threshold set at \( < 1 \times e^{-04} \). Gene names and Gene Ontology (GO) terms were retrieved from the Uniprot database according to the BLASTX top hit results. Repetitive genomic elements were detected using REPEATMASKER v4.0.3 software (Smit et al. 1996–2010).

**Results**

**RAD tag sequencing, de novo assembly and mapping**

The sequencing of the RAD libraries with 314 and 318 chips generated 126 327 and 4 045 509 raw reads, respectively. After filtering, 16 886 contigs covering 2 134 779 bp were retained (N50 contig size 82.56 bp; mean coverage 45.19×) and used as a reference for mapping. From the combined reads of both pools (1 377 041 for Sea and 1 158 807 for Lake pool), MIRA used 1 259 177 sequences (2 086 684 bp; 711 194 and 547 983 reads from the Sea and Lake pool, respectively) for mapping, with an N50 contig size of 148 bp and a mean coverage of 74.6× (Table S1, Supporting information).

**Contig annotation and categorization**

A total of 4369 (25.9%) consensus sequences gave significant BLASTN (3242 contigs; 19.2%) and/or BLASTX (2518 contigs; 14.9%) hits. The largest proportion of significant BLASTX hits came from **Orochromis niloticus** (542 contigs; 21.8%), **Gasterosteus aculeatus** (424 contigs; 16.8%), **Takifugu rubripes** (233 contigs; 9.3%) and **Tetraodon nigroviridis** (196 contigs; 7.8%). From all of the BLASTX hits, GO terms were assigned to 1821 contigs (72.3%).

**SNP validation**

For SNP validation, primers for PCR amplification were developed for 24 putative SNP markers, which were selected from the pre-reanalysed data set according to their chi-square value (12 markers showing high differentiation between Lake and Sea pools) or the presence of a PstI restriction site to test the potential presence of chimeras (\( n = 12 \)). For primer development, an online PRIMER3 v.0.4.0 (Koressaar & Remm 2007; Unterghasser et al. 2012) program was used and primers were selected only if they met the following selection criteria: any’ and 3’ close to 0; melting temperature 57–60 °C and SNP position at least 40 bp from left or right primer to ensure successful sequencing results (primer sequences available on Dryad). Sanger sequencing of seven individuals (four from Lake Peipus and three from Pärnu Bay) was carried out at Estonian Biocentre (http://vvv.ebc.ee) using 3130xl Genetic Analyzer (Applied Biosystems). For sequence alignment and analysis, CLUSTALW multiple alignment option in BIOED-IT v7.2.5 (Hall 1999) was used when possible, otherwise sequences were aligned manually.

**SNP identification and classification**

When all available sequence reads (1 259 177) were used against the assembled reference sequence for SNP
discovery, 7325 potential SNPs were identified. After stringent filtering, a total of 1259 putative SNPs remained (Table S2, Supporting information), corresponding to a SNP density of one per 1754 bp. The most common SNP variant consisted of G/A or C/T. The distribution of the SNPs along the contigs showed that the frequency of identified polymorphisms decreased after the first 100–150 bp (Fig. 2), indicating that an elevated sequencing error rate towards the end of the reads most likely does not affect the identified 1259 SNPs. The mean overall coverage so-called SNPs was 81.7× (mean coverage of Sea and Lake pools: 46.1× and 35.5×, respectively), and 842 of SNPs (67%) had ten or more reads per DNA pool. In total, 32.9% of the filtered SNP-containing contigs had significant BLASTN (23.3%; 293) and/or BLASTX (18.8%; 237) hits. Based on the BLASTX results, 168 and 69 SNPs were located in noncoding and coding regions, respectively, the latter corresponding to 40 synonymous and 29 non-synonymous substitutions.

Genetic differentiation between Sea and Lake pools

Genetic differentiation, measured as $G_{ST}$, varied considerably (0–0.8933) among 1 228 SNP loci, which contained alleles counted for both pooled samples (Fig. 3). $G_{ST}$ values were strongly affected by coverage – most of the SNPs with very high level of differentiation showed rather low coverage (<50×), while loci with higher coverage showed similar or slightly higher levels of differentiation than individually genotyped microsatellite loci (for 17 microsatellites average $G_{ST}$ is 0.016; 95% CI: 0.007–0.025; 24 individuals analysed per population; (Pukk et al. 2014)). Chi-square test identified loci with high levels of differentiation and low sampling error (e.g. high coverage in both pools) between Sea and Lake pools.

SNP validation

From 24 putative SNP markers selected for validation from pre-reanalysed data set (12 highly differentiated loci and 12 loci containing seemingly uncut PstI restriction site), twelve yielded positive PCR amplification products. Eleven of them represented highly differentiated markers while only a single fragment containing seemingly undigested PstI restriction site was successfully amplified. After cleaning, trimming and aligning the sequences, the presence of both homozygote and heterozygote genotypes could be confirmed in 6 of 12 markers while only homozygotes were observed at four SNP loci and the sequence quality was too low for reliable SNP calling at two loci (Table S2).

Discussion

In this study, we described an efficient genome complexity reduction protocol that uses two rare-cutting restriction enzymes and stringent size selection to make a RAD library compatible for sequencing with Ion Torrent semiconductor technology. We obtained a total of 4 M reads, which were de novo assembled into ~7 000 contigs predicted to cover 0.18% of the P. fluviatilis genome (assuming a genome size of 1.19 GB, based on Animal Genome Size Database; http://www.genomesize.com).
Compared to earlier RAD sequencing studies carried out on a wide range of species, our RAD protocol enables very efficient genome complexity reduction (Fig. 4; Table S3, Supporting information). Therefore, our RAD protocol for the Ion Torrent is expected to be useful for studies which require sequencing of a smaller number of genomic regions, such as mapping QTL loci influencing variation in traits of interest, phylogenetic and population genomic analysis. Using two 6-cutter restriction enzymes (PstI and BamHI) in combination with stringent size selection, we were able to target a small proportion of the genome while retaining high sequencing depth allowing reliable SNP calling despite the higher sequencing errors of the Ion Torrent platform compared to currently available light-based sequencing approaches (Loman et al. 2012; Merriman et al. 2012; Bragg et al. 2013). Unlike earlier RAD sequencing studies, that typically suffer the risk of high DNA loss due to physical shearing and purification steps during agarose gel size selection, our protocol included a fast and simple, yet reproducible, narrow size selection strategy using an E-Gel® SizeSelect™ Agarose Gel. This enabled quick separation and collection of size fractions suitable for a particular sequencing platform, significantly reducing the overall library preparation time.

Using our RAD protocol, the mean distance between two contigs in the Eurasian perch is more than 75 Kb, and only a small number of published RAD studies have resulted in a similar level of genome complexity reduction (Fig. 4; Gagnaire et al. 2013; Hale et al. 2013; Ogden et al. 2013; Richards et al. 2013). All these studies have used a single restriction enzyme (SbfI) with an 8-bp recognition site (5’ CCTGCAGG 3’) and size selection of DNA fragments from 300 bp to 600–700 bp. Importantly, 8-cutter SbfI creates the identical four base pair overhangs (5’ TGCA 3’) compared to 6-cutter PstI used in this study and therefore, the same forward adapters can be used in relation to both restriction enzymes to target the optimal number of RAD loci to address a particular research question. As such, our protocol enables fast and flexible RAD sequencing and efficient genome complexity reduction with high coverage using semiconductor sequencing technology.

Eurasian perch, as the most common and widely distributed member of the Percidae family, contributes significantly to commercial and recreational fisheries in Estonia. Because of different fishing regulations, the ability to use genetic markers for assigning individuals back to their source population would be of considerable importance. Also, perch has a great potential for freshwater aquaculture (Rossi et al. 2007), where the use of molecular markers could be beneficial. However, not much is known about the P. fluviatilis genome as the genomic resources of the perch currently include a small number of mitochondrial and nuclear sequences. In this study, we obtained over 2.2 MB of novel sequence data for perch and fifteen per cent of all the consensus sequences gave significant BLASTX hits. As expected, the largest proportion of those hits came from the Nile tilapia O. niloticus (Perciformes). In total, over 1200 putative SNPs were identified and loci with large allele frequency differences between freshwater (Lake Peipus) and brackish water DNA pools (the Baltic Sea) were found, which could be used for the development of genotyping assays to correctly assign fish to populations of origin in the future.

Compared to the microsatellite markers (Pukk et al. 2014), DNA pooling resulted in similar or higher levels of genetic differentiation estimates for the high coverage SNPs (Fig. 3). On the other hand, the SNP markers with relatively low coverage (< 20–50×) showed inflated G\textsubscript{ST} values, demonstrating the importance of high coverage for accurate estimation of genetic differentiation from DNA pools. However, SNP validation of a small subset of loci using Sanger sequencing indicated that not all putative SNP markers represent genuine polymorphisms and therefore, great care should be taken during development of SNP panels for identifying illegal fishing and fish trade. The other critical issue related to allelotyping is the importance of adding equal amount of DNA from each individual to the DNA pool. However, when using large number of individuals for generation of DNA.

---

**Fig. 4** Varying levels of genome complexity reduction in published RAD studies – genome size (MB) versus number of contigs (log10 scale). Empty circles correspond to earlier studies from 2010 to 2013 that refer to (Baird et al. 2008) and contained all necessary information (genome size, number of contigs, platform used and fragment size selection). Filled circle corresponds to this study.
pools, accurate allele frequency estimation can be achieved even with relatively unequal contributions of each individual forming the pool (Gautier et al. 2013a,b). The results of this study also revealed that majority of the loci with apparently uncut PstI and/or BamHI sites represent most likely chimeric sequences. Therefore, we recommend that future studies either exclude putative chimeric sequences before assembly or alternatively carry out in silico cutting of the fragments before subsequent analysis steps. Further optimization of the library protocols and developing alternative strategies to reduce the occurrence of chimeras would be useful.

In summary, this study represents an important step towards developing genomic resources and genetic tools for the Eurasian perch, which could be used in a court of law. Our RAD protocol enables very high genome complexity reduction by targeting a small number of genomic regions while retaining the high coverage necessary for robust SNP discovery and reliable estimation of various population genetic parameters. We expect that the described RAD sequencing protocol for semiconductor sequencing technology will be particularly useful for quick SNP discovery, testing different size selection procedures, restriction enzymes and other critical steps during library preparation. However, Ion Torrent PGM is currently not the cheapest available NGS technology in the market, which combined with rather low throughput and relatively high sequencing error rate makes it less suitable for larger genotyping-by-sequencing projects. Therefore, it is important that both the strengths and limitations of the approach are taken into account when choosing appropriate genome complexity reduction method and sequencing platform (Arnold et al. 2013; Andrews & Luikart 2014; Puritz et al. 2014).

Acknowledgements

We thank M. Lindqvist and O. Thalmann for technical assistance of the Ion Torrent sequencing, two reviewers and the Subject Editor Travis Glenn for the helpful feedback on an earlier version of the manuscript. We also thank the Finnish Center for Scientific Computing for providing computational resources. The study was supported by the Academy of Finland, Estonian Science Foundation (Grant No. 8215), the Estonian Ministry of Education and Research (institutional research funding project IUT8-2), the European Social Fund’s Doctoral Studies and Internationalization Programme DoRa, the Performance Computing Centre of University of Tartu, Estonia, and the European Regional Development Fund through the Center of Excellence in Chemical Biology.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology, 215, 403–410.

© 2015 The Authors. Molecular Ecology Resources published by John Wiley & Sons Ltd
Sharma R, Goosens B, Kun-Rodrigues C et al. (2012) Two different high throughput sequencing approaches identify thousands of de novo genomic markers for the genetically depleted Bornean elephant. *PLoS ONE*, 7, e49533.

Shen PD, Wang WY, Chi AK et al. (2013) Multiplex target capture with double-stranded DNA probes. *Genome Medicine*, 5, 8.

Smit AFA, Hubley R, Green P (1996–2010) RepeatMasker Open-3.0., [http://www.repeatmasker.org](http://www.repeatmasker.org).

Untergasser A, Cutcutache I, Koressaar T et al. (2012) Primer3 - new capabilities and interfaces. *Nucleic Acids Research*, 40, 12.

Yang HA, Tao Y, Zheng ZQ et al. (2012) Application of next-generation sequencing for rapid marker development in molecular plant breeding: a case study on anthracnose disease resistance in Lupinus angustifolius L. * BMC Genomics*, 13, 11.

Yuan YN, SanMiguel PJ, Bennetzen JL (2003) High-Cot sequence analysis of the maize genome. *Plant Journal*, 34, 249–253.

A.V. designed the RAD protocol, L.P. and A.V. designed the study. L.P. conducted the laboratory work. A.V., L.P., F.A., S.H., R.G. and V.K. carried out bioinformatics analysis. L.P. wrote first draft of the manuscript. All authors contributed to the writing of the manuscript, read and approved the final manuscript.

**Data Accessibility**

Raw sequence data, Sanger sequence alignments, Python scripts and primer sequences are available on Dryad, doi:10.5061/dryad.s2405.

Table 1 SuppInfo.xlsb contains the annotated assembled contigs.

Table 2 SuppInfo.xlsb contains the filtered SNP data set.

**Supporting Information**

Additional Supporting Information may be found in the online version of this article:

Table S1 Mapped consensus sequences with BLASTN, BLASTX and RepeatMasker results.

Table S2 Filtered putative SNPs with allele frequencies, Gst, BLASTN, BLASTX and RepeatMasker results.

Table S3 Table of next-generation sequencing studies using restriction-site associated DNA approach that cite Baird et al. (2008).