MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2015) **15**, 1316–1329 doi: 10.1111/1755-0998.12406

Double-digest RAD sequencing using Ion Proton semiconductor platform (ddRADseq-ion) with nonmodel organisms

HANS RECKNAGEL,* ARNE JACOBS,* PAWEL HERZYK†‡ and KATHRYN R. ELMER*
*Institute of Biodiversity, Animal Health & Comparative Medicine, College of Medical, Veterinary & Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK, †Glasgow Polyomics, Wolfson Wohl Cancer Research Centre, University of Glasgow, Garscube Campus, Bearsden G61 1QH, UK, ‡Institute of Molecular, Cell & Systems Biology, College of Medical, Veterinary & Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK

Abstract

Research in evolutionary biology involving nonmodel organisms is rapidly shifting from using traditional molecular markers such as mtDNA and microsatellites to higher throughput SNP genotyping methodologies to address questions in population genetics, phylogenetics and genetic mapping. Restriction site associated DNA sequencing (RAD sequencing or RADseq) has become an established method for SNP genotyping on Illumina sequencing platforms. Here, we developed a protocol and adapters for double-digest RAD sequencing for Ion Torrent (Life Technologies; Ion Proton, Ion PGM) semiconductor sequencing. We sequenced thirteen genomic libraries of three different nonmodel vertebrate species on Ion Proton with PI chips: Arctic charr *Salvelinus alpinus*, European whitefish *Coregonus lavaretus* and common lizard *Zootoca vivipara*. This resulted in ~962 million single-end reads overall and a mean of ~74 million reads per library. We filtered the genomic data using *Stacks*, a bioinformatic tool to process RAD sequencing data. On average, we obtained ~11 000 polymorphic loci per library of 6–30 individuals. We validate our new method by technical and biological replication, by reconstructing phylogenetic relationships, and using a hybrid genetic cross to track genomic variants. Finally, we discuss the differences between using the different sequencing platforms in the context of RAD sequencing, assessing possible advantages and disadvantages. We show that our protocol can be used for Ion semiconductor sequencing platforms for the rapid and cost-effective generation of variable and reproducible genetic markers.

Keywords: double-digest, genotyping by sequencing, Ion Proton, Ion Torrent, RAD sequencing, semiconductor sequencing, single nucleotide polymorphism (SNP) genotyping

Received 12 November 2014; revision received 10 March 2015; accepted 16 March 2015

Introduction

Recent technical advances in genomics have propelled research in ecology and evolution and promoted the integration of these two fields. In particular, the development of next-generation sequencing technologies, which have massively parallelized DNA sequencing, has had a major impact (Stapley *et al.* 2010). High-throughput genotyping of wild populations of nonmodel organisms opens new possibilities to unravel the genetic material leading to phenotypic change and adaptation (Barrett & Hoekstra 2011). Only by the accumulation of such research, we will be able to understand the genetics of adaptation and gain an integrative view of the environment, the phenotype and the genotype.

Correspondence: Kathryn R. Elmer, Fax: +44 141 330 5971; E-mail: kathryn.elmer@glasgow.ac.uk

While next-generation sequencing generates a vast amount of genomic data, the interpretation of such data constitutes a major challenge to scientists. The limit is usually no longer technical, but rather a combination of time, effort and money. The analysis of complex whole genomes is costly and time-consuming and often unnecessary for understanding evolution and genetics. Hence, several methods for reducing the genome to a representative, but more manageable, fraction have been developed recently (Baird *et al.* 2008; Andolfatto *et al.* 2011; Elshire *et al.* 2011; Peterson *et al.* 2012; Narum *et al.* 2013). These reduced genome representation methods make use of restriction enzymes to digest and fragment the genome, followed by targeted sequencing of those fragments. From mutations identified in the sequences of these fragments, hundreds to tens of thousands of single nucleotide polymorphisms (SNPs) can be detected and

serve as genetic markers used to identify genetic structure and adaptive variation in populations. Which approach is most useful depends on several aspects, including the availability of a reference genome/genetic map relevant to the organisms being studied, extant genomic diversity, the level of coverage per marker and individual that can be attained, etc. (details have been reviewed recently quite extensively, in for example Davey *et al.* 2011; Poland & Rife 2012; Narum *et al.* 2013 and will not be covered here).

With no reference genome or other genomic information available, as it is the case for most nonmodel organisms, restriction site associated DNA sequencing (RADseq) has been shown to be a valuable method for the generation of SNP data (Baird *et al.* 2008; Davey *et al.* 2011, 2013; Rowe *et al.* 2011). A modification to the original protocol, double-digest RADseq (ddRADseq) digests the genome with two restriction enzymes rather than one. This reduces library preparation biases induced by DNA shearing and increases the time and cost efficiency by maximizing flexibility in marker quantity across individuals and libraries (Peterson *et al.* 2012). Other RAD sequencing approaches have been developed and all modifications have strengths and weaknesses, as recently summarized in Puritz *et al.* (2014). In general, the complexity and organization of the genome (size, ploidy level, number and type of repetitive elements, GC content, etc.) has considerable implications for the calling of SNPs and defining homologous markers (Rowe *et al.* 2011; Mastretta-Yanes *et al.* 2014). Incorporating size selection to double-digest library preparation allows the greatest flexibility for the trade-off of marker number vs. sequencing effort in reduced representation libraries (Peterson *et al.* 2012; Poland & Rife 2012).

At present, Illumina sequencing by synthesis and Ion Torrent semiconductor sequencers are the most suitable platforms for the high-throughput generation of DNA sequence (Loman *et al.* 2012; Quail *et al.* 2012). The reduced genome representation sequencing methods have almost exclusively been adapted to the Illumina platform [e.g. genotyping by sequencing (Elshire *et al.* 2011), RADseq (Baird *et al.* 2008); ddRADseq (Peterson *et al.* 2012)]. To date, only the genotyping-by-sequencing (GBS) approach has been adapted to Ion Torrent semiconductor sequencing (Mascher *et al.* 2013), and this lacks a size selection step and consequently is less customizable. We therefore developed a new approach for ddRADseq on the Ion platform.

Ion Torrent has the potential to compete with and even exceed Illumina in time and cost efficiency (Glenn 2011; Liu *et al.* 2012). In contrast to optical detection of DNA bases by Illumina sequencers, Ion Torrent technology (Ion Proton™ Sequencer or Ion PGM™ [Personal Genome Sequencer]) uses sensor chips to detect hydrogen ions (H^+) that are released during polymerization as a complementary strand of DNA is synthesized (Rothberg *et al.* 2011). Prior to sequencing, DNA libraries are clonally amplified by emulsion PCR and then loaded onto the Ion chip containing millions of wells, each holding one bead covered with homogeneous DNA fragments. DNA synthesis is initiated by sequentially flooding each of the four native nucleotides on the template DNA, while complementary integration of one of the nucleotides results in a biochemical reaction and hydrogen ion release that is detected by the semiconductor sensor under each well on the chip. The number of bases in homopolymer sequences is detected by the relative change in pH, which occasionally leads to erroneous insertions or deletions (Loman *et al.* 2012). Ion Torrent sequencers generate exclusively single-end reads that are of variable lengths normally distributed around a median value. Strengths of Ion Torrent sequencing are the relatively low purchasing cost of the platform, low cost of sequencing per chip and fast run-time of sequencing (e.g. 2–4 h for a Proton run).

Here, we present a protocol for the generation of SNPs using double-digest RAD sequencing for Ion Proton semiconductor sequencers (ddRADseq-ion). This method is a modification of the ddRADseq methodology for Illumina (Peterson *et al.* 2012) involving revised library preparation, newly designed adapters for Ion Torrent platforms, revisions to the Ion sequencing protocol and minor modifications to the standard bioinformatics pipeline (Fig. 1). We establish the methodology and demonstrate its utility on populations and replicates of three nonmodel organisms with complex genomes – two salmonid fish species and one lizard species. Combining species with similarities (because of phylogenetic history) and differences in genome size and properties demonstrates the robustness of our protocol. We show that ddRADseq-ion is a rapid, robust and cost-effective method for SNP genotyping even without prior genomic resources.

Material and methods

Model species

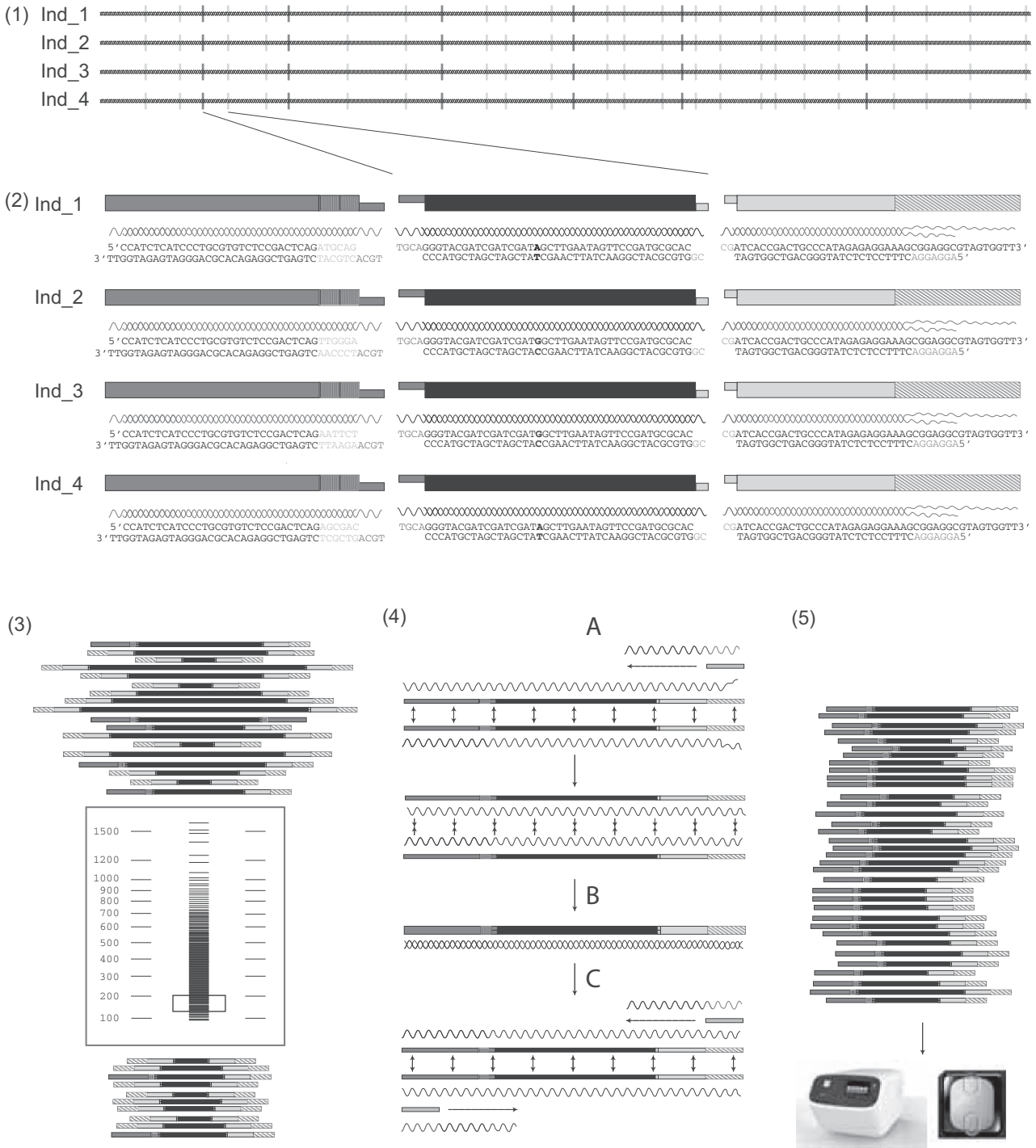
We used three different exemplar species to construct genomic libraries: Arctic charr (*Salvelinus alpinus*), European whitefish (*Coregonus lavaretus*) and common lizards (*Zootoca vivipara*). The former two are fish species with genome sizes around ~3 Gb, and the latter is a European lizard with a genome size of ~1.4 Gb (Gregory 2014) (Appendix S1, Supporting information). To biologically validate the sequencing method, we used two approaches: (i) population genetic analysis of a whitefish hybrid cross from two postglacial and geographically

distinct lineages and (ii) a phylogenetic analysis of Arctic charr and whitefish lineages.

Adapter and primer design for ddRADseq-ion

The protocol follows the general principle of double-digest RAD sequencing that was optimized for Illumina sequencing platforms (Peterson *et al*. 2012; Recknagel *et al*. 2013) with modifications for Ion Torrent sequencing technology (Fig. 1), which differs in chemistry, adapters, amplification primers and sequence output.

We designed new ddRADseq adapters to match the requirements of Ion sequencing platforms. Ion uses a P1 adapter at one end of fragmented DNA and an



A-adapter at the other end. The A-adapter can be barcoded for multiplexing individuals to be sequenced on a single chip. In our modification, the barcoded A-adapter consisted of a four base pair overhang (TGCA, restriction site for enzymes *NsiI*, *PstI* and *SbfI*), a six base pair barcode with a minimum difference of three base pairs between adapters and an Ion Torrent platform sequencing primer site (RADion-A-XXXXX-top: 5′-CCATCTCATCCCT GCGTGTCTCCGACTCAGXXXXXTGCA-3′; RADion-A-XXXXX-bottom: 5′-XXXXXCTGAGTCGGAGACACGCAGGGATGAGATGG*T*T-3′; where XXXXX stands for the unique 6-bp barcode and asterisks denote phosphorothioate bonds to prevent degradation by nucleases). We developed fifty of these uniquely barcoded adapters (Appendix S2, Supporting information).

The second, global adapter (modified P1-adapter for Ion Torrent) consisted of a two-base pair overhang (GC, restriction site for *Msp*I) and the Ion Torrent platform-specific primer binding site (RADion-P1-top: 5′-AGGAGGACTTTCCTCTCTATGGGCAGTCGGTGAT-3′; RADion-P1-bottom: 5′-CGATCACCGACTGCCCATAGAGAGGAAAGCGGAGGCGTAGTGG*T*T-3′). This modified P1 was designed as a Y-divergent adapter, so that during later PCR, only those fragments that contain a barcoded and a global adapter on opposing ends would amplify (see Baird *et al.* 2008).

After ligation and size selection (see 'Library construction'), the fragments are enriched by PCR. We designed ddRADseq-ion forward and reverse amplification PCR primers with sequence (RADion-for-primer) 5′-CCACTACGCCTCCGCTTTCC-3′ and (RADion-rev-primer) 5′-CCATCTCATCCCTGCGTGTCT-3′, respectively (see Fig. 1 for illustration of the method). The enriched fragments are then amplified with an emulsion PCR using, for example, the Ion OneTouch kit (Life Technologies) and loaded onto the Ion chip for sequencing.

Library construction

ddRADseq-ion library construction involves DNA extraction, digestion with enzymes, adapter ligation, size selection and finally fragment enrichment before Ion sequencing. A detailed workbench protocol for the library construction is available as part of the supplementary material (Appendix S3, Supporting information). Complete library preparation from DNA extraction takes approximately 14 h of hands-on laboratory bench work over 4 days for constructing libraries of 30 individuals.

Briefly, high-quality genomic DNA was extracted from fin clips (fish) or tail muscle tissue (lizards), integrity assessed by visualization after electrophoresis on an ethidium bromide agarose gel and quantified using a Qubit Fluorometer with the dsDNA BR Assay (Life Technologies). Each sample was normalized to a total amount of 1 μg of DNA with a minimum concentration of 25 ng/μL and digested using two restriction enzymes, a rare-cutting (*PstI-HF*, 20 units) and a frequent-cutting enzyme (*MspI*, 20 units) in combination with the CutSmart buffer (New England Biolabs) (Fig. 1). These particular restriction enzymes were chosen due to their frequent use for reduced representation library constructions (e.g. Mascher *et al.* 2013; Recknagel *et al.* 2013; Henning *et al.* 2014), but with the same adapters other enzyme combinations would be possible, if associated with TGCA (e.g. enzymes *NsiI*, *PstI* and *SbfI*) and GC (e.g. enzymes *HpaII*, *AciI* and *HpyCH4IV*) restriction sites. The digest was incubated at 37 °C for 3 h in a PCR thermocycler. The digested samples were each cleaned using a MinElute Reaction Cleanup Kit (Qiagen) and then postcleanup sample DNA concentrations were measured using a Qubit Fluorometer with the dsDNA BR Assay.

After adapter annealing, barcoded and global adapters were ligated to the DNA in a single reaction with

Fig. 1 Overview of the ddRADseq-ion methodology. After DNA extraction, the whole genome is fragmented by two restriction enzymes (see Peterson *et al.* 2012). In (1), a representative region on a homologous chromosome is shown for four individuals. Vertical lines represent cutting sites for a frequent-cutting (light grey) and a rare-cutting (dark grey) enzyme. After enzymatic digestion, adapters are ligated to the overhang of the sheared genomic sequences (2). The RADion-A-adapter binds to the rare-cutting restriction site (shown in dark grey). RADion-A-adapters contain a specific 6-bp barcode (shown as vertically striped) for tagging individuals when multiplexing. The global RADion-P1-adapter binds to the more frequently cut overhang (shown light grey), so that two adapters flank the genomic fragment. The global adapter is Y-divergent (shown as diagonally striped) (see Baird *et al.* 2008), meaning that it is not complementary on its end. This design prevents later amplification of fragments that are bound to only one type of adapter on both restriction site ends. Representative sequences of adapters and genomic fragments are shown and a SNP in the genomic fragment is shown in bold. Adapter-ligated fragments vary substantially in length before size selection. Only a specific range of fragments (130–200 bp in this illustration) is size selected to further reduce the genomic representation to be sequenced (3). Prior to size selection, all barcoded individuals are pooled. The genomic DNA fragment may be flanked by the same adapter on both ends or by different adapters. Only those fragments that have the global adapter on one end and the barcoded adapter on the other end are amplified during PCR (4). During the first amplification step, the RADion-for primer binds to the P1-adapter (A) and builds the complementary strand (B). The RADion-rev primer binds to the A-adapter end of the complementary strand resulting from the previous amplification step (C). The amplified fragments are then sequenced using the Ion Torrent platform (5). Ind = Individual.

Table 1 Summary statistics of ddRADseq-ion libraries. Libraries varied in biological levels (species, lineages or regional populations, and genome size), number of individuals per library and target size selection range (library size). Sequencing outputs are specified as median read length and the number of total reads per library. The number and percentage of reads retained after quality filtering and read trimming are given. bp = base pairs

Library ID	Species	Region/cross	C-value (mean)†	Library size (bp)	Median read length (bp)	Total reads	Retained reads	% Retained reads
1	*Zootoca vivipara*	Eurasia	1.38	175–225	96	76 871 978	60 998 276	79.4
2	*Coregonus lavaretus*	Scotland	3.04	130–200	73	66 376 198	33 942 046	51.1
3	*Coregonus lavaretus*	Scotland	3.04	130–200	85	51 892 209	40 701 296	78.4
4	*Coregonus lavaretus*	Scotland	3.04	250–320	85	68 872 227	49 690 064	72.1
5	*Coregonus lavaretus*	Alpine × Baltic	3.04	130–200	72	73 343 457	43 617 408	59.5
6	*Coregonus lavaretus*	Alpine × Baltic	3.04	130–200	61	77 307 832	28 767 867	37.2
7	*Salvelinus alpinus*	Russia	3.33	130–200	81	85 041 194	57 893 374	68.1
8	*Salvelinus alpinus*	Scotland	3.33	130–200	90	87 534 089	65 339 925	74.6
9	*Salvelinus alpinus*	Scotland	3.33	130–200	78	70 174 587	49 669 770	70.8
10	*Salvelinus alpinus*	Iceland	3.33	130–200	83	83 492 040	47 069 798	56.4
11	*Salvelinus alpinus*	Scotland	3.33	130–200	70	81 150 442	58 539 993	72.1
12	*Coregonus/Salvelinus*	Europe	~3.18	130–200	71	61 210 349	36 267 995	59.3
13	*Coregonus/Salvelinus*	Europe	~3.18	130–200	75	78 768 073	52 421 636	66.6

†Gregory (2014).

0.125 μM of RADion-A-adapter (unique for each individual) and 0.125 μM of RADion-P1-adapter per individual, T4 ligase (1000 units) and 10× T4 ligation buffer. Ligation reactions were incubated under the following conditions: (i) ligation for 30 min at 25 °C, (ii) heat kill at 65 °C for 10 min and (iii) cool down to room temperature (2 °C per 90 s).

Tagged individuals were pooled at equal concentration into multiplexed libraries (6–30 individuals). These multiplexed libraries were individually size selected using a Pippin Prep (Sage Science) targeted, automated size selection machine using 2% dye-free gel cassettes. DNA libraries were size selected in a target range of 130- to 200-bp fragments, 175- to 225-bp fragments or 250- to 320-bp fragments (see Table 1). Marker E was used as a reference for the size selection. Size selecting ddRADseq libraries with automation has been found to considerably reduce interlibrary sequencing variability compared with manual size selection from agarose gels (Peterson *et al.* 2012), which ultimately maximizes the number of shared markers and sequencing efficiency. The size-selected libraries were subsequently quantified using a Qubit Fluorometer with the dsDNA HS Assay.

Following the size selection, an enrichment PCR was performed to amplify the libraries. Four to seven PCRs were performed for each library to reduce PCR bias using 5–10 ng of library DNA, depending on the amount of DNA available (concentration ranged from 0.40 to 1.34 ng/μL). The PCR mix consisted of 0.4 μL dNTPs, 0.2 μL taq polymerase, 4.0 μL buffer HF, 1 μL each of forward and reverse RADion primers (10 μM each) and template DNA. Each PCR was topped up to 20 μL with ddH_2O. Thermal conditions were set as follows: 30 s 98 °C, 10X [10 s 98 °C, 30 s 65 °C, 30 s 72 °C], 5 min 72 °C. A small amount of each PCR product was run out on an agarose gel next to the library template to check whether the libraries were amplified. The multiple separate PCRs for each library were then combined and cleaned using the MinElute Reaction Cleanup Kit (Qiagen).

Following the clean-up, the libraries were electrophoresed on a 1.25% agarose gel to remove any remaining adapter dimers and fragments outside the size range selected by the Pippin Prep. SYBRSafe (Life Technologies) was used for gel staining because ethidium bromide may interfere with downstream sequencing protocols. The bands in the size range chosen during size selection were cut out manually, and the library DNA was extracted from the matrix using a MinElute Gel Extraction Kit (Qiagen). Following the gel extraction, DNA was quantified using a Qubit Fluorometer with the dsDNA BR Assay. To determine the DNA quality, exact size distribution and molarity, the libraries were analysed using a BioAnalyzer or TapeStation (Agilent Technologies). Final ddRADseq-ion libraries had a concentration of 4.56–5.54 ng/μL and bell-shaped size distribution around a mean that depended on the selected size range.

To generate a variable data set of genomic libraries (e.g. number of loci, coverage and SNPs) in order to optimize multiplex and library construction parameters, we analysed several species. In addition, we used different

size selection ranges (between and within species, see Table 1) to increase variation between libraries and assess its impact.

Ion Proton sequencing

ddRADseq-ion libraries were prepared and sequenced with minor modifications to the manufacturer's protocol for Ion Proton genome sequencing. Emulsion PCR was performed with a final concentration of 0.336 pM of library DNA (reduced from manufacturer's suggestion of 0.417 pM) using the Ion OneTouch 2 instrument and the Ion PI Template OT2 200 Kit v3. Decreasing the library concentration in the emulsion PCR causes a lower percentage of template-positive and polyclonal Ion sphere particles (ISPs), which therefore maximized the number of usable reads per chip. Following emulsion PCR, a Qubit Ion Sphere Control Assay was performed to control the percentage of template-positive ISPs to a range of 10–25% per manufacturer's instructions. Libraries were sequenced at Glasgow Polyomics using an Ion PI Sequencing 200 Kit v3 on an Ion Proton PI chip.

Bioinformatic processing for ddRADseq-ion

The most commonly employed SNP identifying software for RADseq analysis is the programme *Stacks* (Catchen *et al.* 2011), which was developed for Illumina sequences and requires a common read length for all individuals in order to call individual genotypes. We made minor modifications to be able to analyse the ddRADseq-ion data (which is of variable length around a median) in the *Stacks* v1.20 pipeline. All reads were trimmed to a length of 60 bp. We selected 60 bp after optimizing to balance read length and the number of reads retained (Appendix S4, Supporting information). RAD fragments were demultiplexed based on their barcodes using the *Stacks* script 'process_radtags'. Reads shorter than the trimming threshold were discarded during this step.

The trimmed and grouped reads were further processed using the *Stacks* 'denovo_map.pl' pipeline. This pipeline executes three different *Stacks* scripts to build loci and call SNPs in each sample (ustacks), create a catalogue of all loci for the samples (cstacks) and finally to match the loci of the samples against the catalogue (sstacks). The parameters for the 'denovo_map.pl' pipeline were set to a maximum genetic distance within an individual locus of $m = 2$ and between individuals to a single base pair ($n = 1$). The minimum coverage depth to create a stack was set to $m = 3$, and the number of mismatches allowed when building aligning secondary reads was set to $N = 3$. Furthermore, the removal and separation of highly repetitive RAD fragments was enabled in the 'ustacks' script (−t option). The SNP model using a maximum-likelihood framework implemented in *Stacks* was chosen to call a homozygote or heterozygote.

The *Stacks* script 'populations' was used to export loci for further downstream analyses. The coverage threshold for population genetic estimates of genetic diversity and all other subsequent analyses was set to eight reads per individual locus ($m = 8$). To be counted as a shared locus, we set that a locus had to be present in at least 75% of all individuals within a catalogue (−p option).

To minimize genotyping errors, the 'rxstacks' script was used after running the 'denovo_map.pl' pipeline. The 'rxstacks' script applies four different types of corrections to a *Stacks* analysis: SNP model correction, log-likelihood filtering, a confounded locus filter and haplotype pruning. After running 'rxstacks', 'cstacks' and 'sstacks' were then run again to build and match the filtered loci and haplotypes with the corrected SNP calls to the catalogue.

Error estimates

To estimate repeatability and error rates resulting from library preparation, sequencing or bioinformatic analyses of ddRADseq-ion, we used two technical replicates of the lizard samples from DNA digestion with enzymes through to SNP calling. The replicates were sequenced in the same library, excluding any variation that might result from different sequencing runs. We calculated the SNP error rate because it is the relevant measure for subsequent genetic analyses. We applied an R script recently published by Mastretta-Yanes *et al.* (2014) to calculate SNP error rates. Basically, the replicate genotype files (extracted as plink files from the 'populations' script in *Stacks*) are compared, and the number of SNP mismatches is counted and calculated as a ratio over all compared loci (excluding loci with missing data). In addition to comparing the error rates of the two replicate samples at eightfold coverage, we tested the effect of increasing the coverage to 16×.

Population and phylogenetic validations

To further validate our ddRADseq-ion sequencing data in biological context, we used population genetic and evolutionary analyses. The first approach quantified genomic compositions of a whitefish hybrid cross and their offspring. The two parents originated from distinct evolutionary lineages (Alpine region or Baltic region) (Hudson *et al.* 2011). Their offspring ($N = 14$) should contain about half of each parental genome. To test this, we used population genetic analyses, extracting loci with one or two SNPs in any individual (Table 2). If a locus had two SNPs, only the first SNP was recorded to avoid

Table 2 Library statistics after data processing. Libraries were analysed in *Stacks* (Catchen *et al.* 2011) and include filtering steps, the resulting number of SNP markers and a population genetic estimate of nucleotide diversity. Catalogue loci represent the total loci per library, of which the number and per cent shared by at least 75% of individuals per library are given; these determine the number and per cent of reads used per library. 'Analysis' specifies for which validation analysis samples of a particular library were used. All libraries were used to calculate summary statistics

Library ID	*N* individuals	Mean coverage	Stdev coverage	Catalogue loci	Shared (>75%)	% Shared	Used reads	% Used reads	*N* loci SNPs*	Nucleotide diversity**	Analysis
1	27	18.3	5.08	625 306	29 433	4.7	20 073 753	32.9	10 870	0.0078	–
2	30	16.4	2.42	187 748	30 157	16.1	12 972 728	38.2	6225	0.0015	–
3	6	39.8	6.06	378 077	87 908	23.3	17 510 130	43.0	17 444	0.0020	–
4	6	20.4	2.12	538 133	102 959	19.1	10 437 148	21.0	26 157	0.0029	–
5	30	16.0	1.6	456 169	32 534	7.1	13 916 052	31.9	12 267	0.0014	Str
6	30	15.8	1.2	165 671	24 299	14.7	9 778 701	34.0	8049	0.0022	Str
7	30	23.5	2.05	266 807	70 849	26.6	30 122 343	52.0	8076	0.0009	Phy
8	30	24.0	2.36	478 996	76 419	16.0	31 548 570	48.3	15 082	0.0012	–
9	6	42.1	6.02	361 057	109 278	30.3	27 585 487	55.5	20 397	0.0016	–
10	30	16.0	2.63	259 560	45 290	17.4	22 666 196	48.2	7282	0.0011	–
11	30	17.6	1.54	367 308	59 869	16.3	29 985 128	51.2	8153	0.0012	–
12	16	22.3	3.35	409 283	31 921	7.8	7 310 063	20.2	12 651	0.0039	Phy + Str
13	17	26.8	2.04	446 356	6696	1.5	2 471 622	4.7	3134	0.0092	Phy + Str

Stdev = standard deviation; bp = base pairs; Str = population structure; Phy = phylogenetics. One/two asterisks specify options used in *Stacks*.

*With 75% of shared loci across all individuals and 1–2 SNPs.

**With 75% of shared loci across all individuals and an eightfold coverage per locus.

genetically linked SNPs in the data set. Only loci that had at least an eightfold coverage were retained for analysis. Population structure was inferred from fixed SNPs between parents (following standard genetic mapping procedure and allowing missing data at loci in offspring) using *Structure* (Pritchard *et al.* 2000) with three runs of $K = 2$ (one cluster for each parental genome) using the admixture model with correlated allele frequencies among populations (standard settings) for 50 000 generations after a burn-in length of 5000.

The second approach intended to resolve phylogenetic relationships of three whitefish and three Arctic charr individuals from different regions in Europe. Loci had to be present in 75% of all individuals and a minimum coverage of 8× per locus to be extracted from *Stacks*. In addition, only loci with one to three SNPs that were variable between and/or within individuals were retained. Choosing the right SNP boundaries depends on the genetic divergence within and, if including other species, also between species and should be customized for each project. For example, if interspecific phylogenies are created, a larger genetic distance (more SNPs per locus) might be allowed between individuals. However, this increases the chance of confounding homologous loci with paralogous loci between species. In general, longer reads with few SNPs (~1–3 SNPs) will be most robust against confounding paralogous loci. We analysed our SNP data set using the maximum-likelihood software *RAxML* (Stamatakis 2006).

Results

Library depth and sequencing coverage

Across the thirteen ddRADseq-ion libraries sequenced on Ion Proton, read length was normally distributed around 80 bp (mean: 78.5 bp; median: 78 bp; range: 61–96 bp; standard deviation: 10.0 bp) (Table 1; Appendix S5, Supporting information). A total of ~962 M reads were generated, on average ~74 M reads per library (standard deviation: 10.5 M).

To analyse the data bioinformatically, reads were trimmed to a common length. Setting of the trimming threshold should be optimized to maximize the total number of base pairs retained. Decreasing the threshold will increase the number of reads (as moving the threshold to the left end of the distribution will retain shorter reads), but decrease the read length. We generated a simple R script to determine at which length the number of base pairs of a raw FASTQ file is maximized (provided in Appendix S4, Supporting information).

In this study, we used a common threshold of 60 bp to consistently compare the different libraries. Trimming reads to 60 bp resulted in 65.0% of reads being retained on average after the read quality filtering step in *Stacks* (on average 48 070 727 reads per library). Trimming the reads to 50 bp resulted in more retained reads (~16.9% more reads and 15.1% higher coverage per locus), but depending on the species in either slightly fewer SNPs

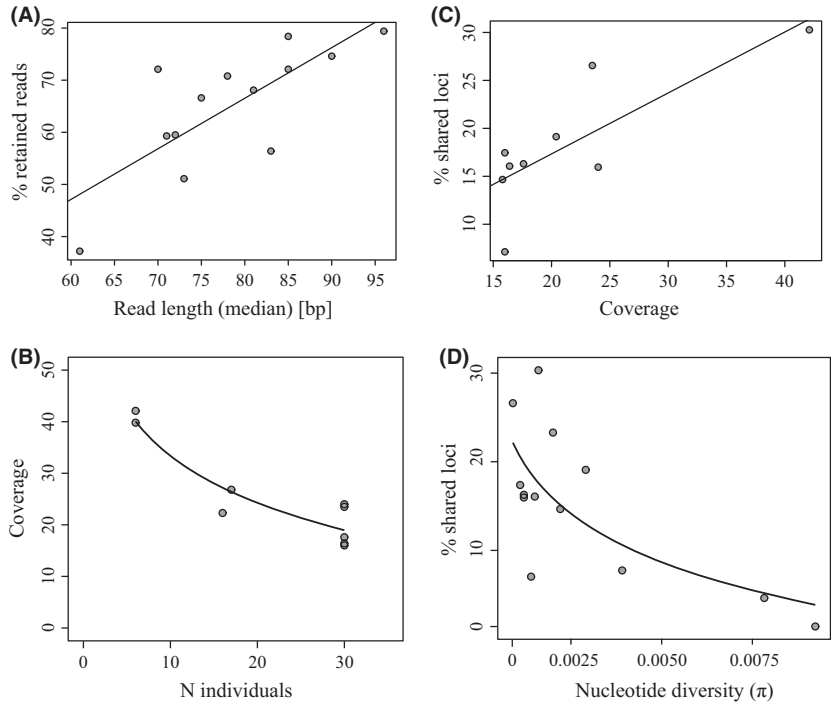


Fig. 2 Analysis of the relationships between median sequencing read length, coverage and data retained of the sequenced ddRADseq-ion libraries. When using a fixed trimming threshold (e.g. 60 bp), the number of retained reads increases with the raw median read length of the sequenced library (A). Read coverage per individual generally decreases if more individuals are included in the library (B). After data processing and generating catalogue loci, the percentage of loci shared by at least 75% of all individuals increases with read coverage per individual (C) and decreases with nucleotide diversity of the library (D).

(because reads were shorter) or more SNPs (because more reads are retained). Using reads trimmed to 70 bp resulted in fewer reads and also fewer SNPs compared to reads trimmed to 60 bp (13.9% fewer SNPs and 14.1% less coverage). In general, libraries with longer reads (read length distribution is shifted to the right; Appendix S5, Supporting information) resulted in a larger number of retained reads ($R^2 = 0.582$; Fig. 2A) because more reads pass the trimming threshold. We suggest the read length should be optimized for each study depending on the sequencing technology (e.g. PGM or Proton) and median read length (dependent on sequencing quality and chip used).

Mean coverage per unique read per individual ranged from 15.8× to 78.4× between libraries, depending on species, number of individuals and size selection range used. As expected when using similar size selection parameters and species with similar genome sizes, the coverage decreases when more individuals are included in a library (here a library equals a sequencing chip) (Fig. 2B). Coverage per unique read was consistent across individuals evidenced by the standard deviation of coverage ranging from 1.2 to 6.1 ($N = 13$; mean = 3.0×; Table 2). The high coverage (>15×) and low variation in coverage (max. 6.1×) suggest that our data are suitable for (most types of) genetic analyses.

Testing library construction of size selection parameters

A key component of ddRADseq is that the combination of enzymes and size selection should allow researchers to minimize the number of loci (to sequence more individuals per chip) or maximize the number of loci (if sequencing is not limited) depending on the study. Our assessment of this balance was based on two libraries of the same biological samples of European whitefish but for which we selected different size ranges (Pippin Prep size selection) of 70-bp fragments: one selecting for 130–200 bp and the other for 250–320 bp. The longer library (250–320 bp) contained more loci (538 133 vs. 378 077 from the shorter library) and therefore a lower overall coverage per locus (20.4× vs. 39.8×). As a result of the larger number of loci, the longer library also contained more shared loci (102 959 vs. 87 908) and SNPs (26 157 vs. 17 444), but a lower percentage of shared loci (19.1% vs. 23.3%), presumably due to the lower coverage. The larger number of loci and SNPs was expected for the longer library, as *MspI* cuts every 256 bp on average (assuming equal base frequencies within the genome); in principle, the closer the size selection range is to the average enzyme cutting, the more loci will be obtained. In this way, the final amount of loci can be altered and optimized, in addition to using a larger or smaller size selection band (note Pippin Prep cannot excise narrower than a 50 bp range of fragments) or different restriction enzymes.

Quantifying shared loci across biological levels

Across each single-species library, on average 17.4% of catalogue loci were shared among at least 75% of individuals (a mean of 60 818 loci with a standard deviation

of 30 827 across libraries; Table 2). These 17.4% of loci contained an average of 41.5% of the total retained reads in a library. Excluded reads were associated with loci that were shared by fewer than 75% of the individuals.

When combining all 126 individuals of Arctic charr from five libraries (libraries 5–9), 55 647 loci were shared. This is lower than the average of 72 341 shared loci when fewer individuals (N = 30 in libraries 5–8, N = 6 in library 9) were combined and each library analysed separately. These 55 647 loci of all combined charr contained 43.1% of all retained reads from the five combined libraries.

When a similar number of loci are sequenced across libraries (e.g. when using the same size selection range and species with similar or comparable genome sizes), increasing the coverage should maximize the percentage of shared loci. As predicted, we found that higher coverage of reads per individual in a library resulted in a higher percentage of shared reads (R^2 = 0.630) when libraries that were size selected for the same fragment size range (130–200 bp) and with species exhibiting a similar genome size (*Salvelinus* and *Coregonus*) were considered (N = 10, Fig. 2C).

As expected because of the more distinct genomes, the two libraries (12 and 13) based on multiple species shared a substantially lower percentage of loci across 75% of individuals (total shared loci: 31 921 and 6696; 7.8% and 1.5% shared loci, respectively). Overall nucleotide diversity in a given library was generally higher when distantly related lineages (Libraries 1, 12, 13: *Zootoca* lineages or *Salvelinus* and *Coregonus* combined) were included, as would be expected (Table 2). Hence, libraries that contained a higher overall nucleotide diversity (more genetically distinct lineages) shared a lower percentage of loci (Fig. 2D). Similarly, when combining all Arctic charr libraries that contained individuals from distinct geographic regions, the number of shared loci was lower across all libraries than within libraries (across libraries: 55 647; average within libraries: 72 341), as individuals within libraries were generally genetically closer to each other than between libraries.

In summary, sequencing with high-coverage and minimizing nucleotide diversity within and across libraries will maximize the number of shared loci (Fig. 2C,D). This should be taken into consideration if, for example, different libraries contain individuals from different evolutionary lineages of a species; the number of shared loci will be lower across libraries than within, as genetic distance between individuals increases when combining the two libraries. Similarly, coverage should be high across libraries to maximize the percentage of shared loci.

ddRADseq-ion SNP error rate

We calculated SNP error rates based on two technical replicates of two lizard individuals analysed in pairwise comparisons. Replicate samples were given different individual barcodes in a library and sequenced on a single chip. The SNP error rates for 3944 polymorphic loci were 1.8% for one and 2.2% for the other individual. Increasing the coverage from 8× to 16× reduced the number of loci retained in the analysis to 1119 and decreased the error rate only slightly, to 1.6% in one and not at all (2.2%) in the other individual.

Method validation via evolutionary analyses

The analysis of the whitefish cross of Alpine and Baltic lineages produced the admixture results expected from parental to F1s. The data set consisted of 21 317 SNP loci with a mean coverage of 23.5x ± 11 per individual and a nucleotide diversity (π) of 0.0046. *Structure* analyses of parents and offspring (focusing on N = 1356 SNPs fixed between parents) clustered offspring genomes 50.3% to the Alpine and 49.7% to the Baltic lineage (standard deviation of Q-value across individuals = 1.6%), with membership values ranging from 0.463 to 0.537 (Fig. 3).

The phylogenetic reconstruction of charr and whitefish lineages was composed of 6036 variable SNPs in 4060 loci. Both species are clearly separated by long branches and high support (Fig. 4, full length of branches not shown). Intraspecific relationships show

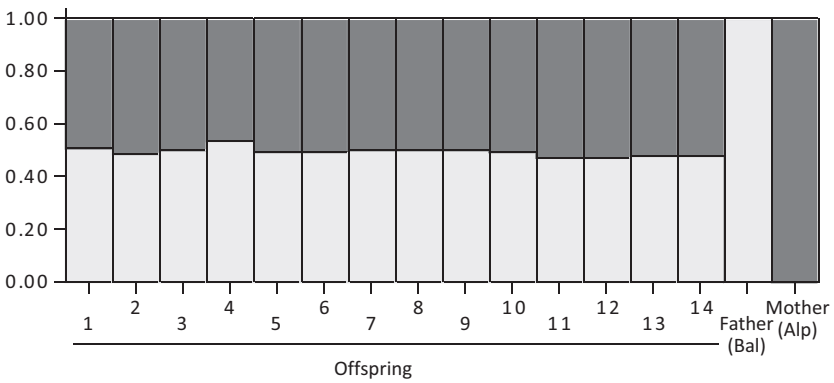


Fig. 3 Genomic structure of an interlineage cross between Alpine and Baltic European whitefish (*Coregonus lavaretus*). Each bar represents a single individual. Different colours illustrate Alpine (dark grey) or Baltic (light grey) genomic ancestries. Respective membership values (Q-value) for each individual are shown on the y-axis.

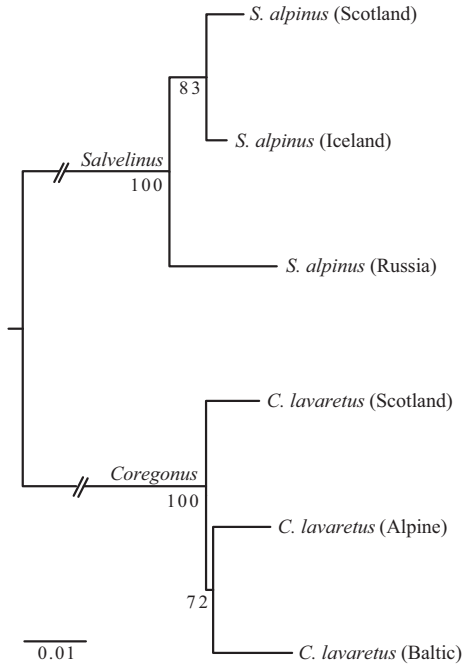


Fig. 4 Phylogenetic relationships between the salmonid fishes Arctic charr (*Salvelinus alpinus*) and European whitefish (*Coregonus lavaretus*) from different geographic regions. Nodal supports are shown as bootstrap values derived from maximum-likelihood analysis. The phylogeny was generated from 6036 variable SNPs.

clear divisions between individuals from different geographic regions and nodal bootstrap supports are generally high (72–100).

Discussion

Here, we developed and demonstrated double-digest RAD sequencing using Ion Torrent semiconductor sequencing platforms (ddRADseq-ion) and showed that it works robustly in population genetic and phylogenetic frameworks. ddRADseq on the Illumina platform has been well established as a next-generation sequencing genotyping protocol at a breadth of biological scales (Peterson *et al.* 2012), and our approach is a new variation. Because the Ion Torrent chemistry and sequencing technology differ from Illumina, here we outlined a series of modifications to library preparation and subsequent bioinformatic pipelines. Our protocol results in a rapid, robust and cost-effective genotyping protocol for non-model organisms.

Characteristics of ddRADseq-ion: read length

Ion Torrent sequencing generates reads of different length around a median value (Fig. 2A). For efficient SNP calling using existing pipelines, all reads need to be of a similar length or else they cannot be compared and overlapped as a locus. Therefore, ddRADseq-ion reads need to be trimmed to a common length before calling loci and SNPs, ideally by user-determined and project-specific parameters that maximize read length and number of reads retained. If the reads are trimmed to a very short length, the probability of confusing a paralogous locus with a polymorphic locus increases. With increasing read length, this probability decreases because of the increasing amount of comparable sequence (e.g. Li *et al.* 2001). However, with the ddRADseq-ion approach, the coverage decreases when trimmed read length is increased because more reads will not pass the trimming threshold, leading to the exclusion of loci.

To compare the different libraries, we chose a read length of 60 bp (genomic DNA excluding the barcode), although in principle the trimming threshold should be adjusted based on the maximum number of base pairs obtained, coverage and possibly also on the expected genetic distance between the analysed individuals. Because of the discussed variable read length from Ion Torrent sequencing platforms, inevitably more reads are lost during initial filtering steps using ddRADseq-ion compared with Illumina-based RADseq; in our case, ~70% reads were retained, while usually more than 80% of all reads are retained in Illumina-based RADseq (Carmichael *et al.* 2013; Palaiokostas *et al.* 2013; Recknagel *et al.* 2013; Vandepitte *et al.* 2013). The possibility of Ion Torrent technology sequencing longer reads could minimize these size distribution effects, for example, with PGM chips which use the same sequencing chemistry and adapters as outlined here.

Characteristics of ddRADseq-ion: error rates

The wholesale generation of sequencing data increases the chance of generating errors compared with more small-scale approaches, starting from the library preparation up to the final genotype calling. Fortunately, there are several steps during the bioinformatic processing that allow us to estimate error rates and filter these out (e.g. Henning *et al.* 2014; Mastretta-Yanes *et al.* 2014). Here, we found an average SNP error rate of 1.8–2.2% from ddRADseq-ion, calculated from technical replicates that span the entire pipeline from library preparation to SNP calling.

The sequencing error rate for Ion Torrent is reported to be between 1 and 2%, while their most common errors are insertions or deletions (indels) as opposed to base substitutions common in Illumina sequencing (Glenn 2011; Loman *et al.* 2012; Quail *et al.* 2012). While substitutions may be called as SNPs during bioinformatic processing, indels create frameshifts and might be called as a different locus. Hence, the way these different types of

errors affect the genomic data set and marker calling also differs. High coverage usually increases the chance of avoiding substitutions induced by sequencing errors (Henning *et al.* 2014; Mastretta-Yanes *et al.* 2014). For indels, the same is true: a locus that results from an indel sequencing error would have a lower coverage than true loci. In addition, as the locus needs to be shared by a certain proportion of individuals, the chance of including such wrong indel loci further decreases dramatically (note that this is not the case for substitutions). Therefore, while the sequencing error rate is generally viewed as being relatively high using Ion Torrent sequencing, indels should have a negligible effect on the SNP error rate in ddRADseq-ion. We suggest a minimum coverage of eightfold per locus should be sufficient to ensure low error rates and that increasing coverage to 16× does not substantially improve SNP confidence.

Studies using Illumina-based RAD sequencing have calculated sequencing error rates between 0.2 and 3.7% (Emerson *et al.* 2010; Peterson *et al.* 2012); however, those estimates did not include errors originating during the library preparation (e.g. PCR bias) and the bioinformatic processing (e.g. SNP calling) of the sequence data (Mastretta-Yanes *et al.* 2014). Using technical replicates, a study by Mastretta-Yanes *et al.* (2014) found SNP error rates between 2.4 and 5.8% across the entire molecular and bioinformatics genotyping protocol, using *Stacks* pipelines on Illumina-sequenced RAD libraries. Their error rates are slightly higher than our 1.8–2.2% from ddRADseq-ion and emphasize the importance of including technical replication in the experimental design. In summary, ddRADseq-ion produces more 'waste' in the form of discarded reads, but the retained loci should be of high confidence after filtering for coverage and shared loci.

Characteristics of ddRADseq-ion: sequencing effort

Genomic studies endeavour to maximize the number of markers, the confidence of those markers and to minimize missing data. Missing data can be problematic in downstream statistical analyses. Maximizing the read coverage of each individual minimizes missing data and simultaneously improves the reproducibility of loci across individuals (Fig. 2C). It is a common problem in NGS-generated data – and particularly in reduced representation methods – that a large proportion of sequenced loci are not shared between all individuals (e.g. Lemmon & Lemmon 2013; McCormack *et al.* 2013). Further, a large proportion of loci that are shared will be invariant (ranging from 53% to 89% in this study; even 90% in Recknagel *et al.* 2013), because a ddRADseq polymorphism will only be detected at the rate of background mutation. This is a psychological shift from earlier genotyping protocols, in which variability was determined *a priori* or in pilot studies, such as when screening microsatellite libraries or generating SNP chips.

In our study, the percentage of loci shared across individuals is relatively low at first view; 5–30% (20% on average) across >75% of individuals, depending on the genetic distance among individuals in the library (Fig. 2D). However, per chip, the 20% shared loci represent almost half (~42%) of all sequenced reads that passed initial filtering thresholds. Therefore, while much data are dispensed with before analysis, the retained high-coverage polymorphic loci generated by ddRADseq-ion are considerable and, because of the low financial cost and fast library preparation and sequencing time, the method is overall very efficient.

This high percentage of discarded data results in part from the library preparation: during the size selection step, the margins of the fragment range will by chance contain some fragments that are slightly shorter or longer than the target range. These fragment will then be amplified, presumably at lower coverage than other fragment that are 'truly' within the size range (Mastretta-Yanes *et al.* 2014). This is a common issue in ddRADseq (Peterson *et al.* 2012; Mastretta-Yanes *et al.* 2014) and not specific to ddRADseq-ion.

The average coverage per unique read was quite consistent across individuals sequenced in ddRADseq-ion libraries (Table 2), evidenced by a standard deviation relatively low (ranging from 1.2× to 6.1×) compared to published Illumina RADseq libraries (e.g. Baxter *et al.* 2011; (10.6×); Liu *et al.* 2013 (6×); Lexer *et al.* 2014 (10.8×) [our calculations from their tables]). However, in part, this may also reflect differences between single-digest and double-digest prepared RAD libraries (Peterson *et al.* 2012; Davey *et al.* 2013; Puritz *et al.* 2014) rather than differences between Illumina and Ion sequencing platforms.

Cost efficiency of sequencing effort is probably the major factor on deciding which sequencing platform to use for genotyping. From our experience, the ddRADseq-ion library preparation costs approximately US$23 per sample, assuming a range of a few to tens of individuals. Per sample costs may decrease when more samples are used. Barcoded RADion-A-adapters cost approximately US$80 (e.g. US$2400 for 30 individuals), and the single global RADion-P1 pair costs approximately US $130. At its current incarnation, the Ion Torrent reagents and sequencing cost approximately US$1000 per Ion Proton PI chip (Glenn 2011, updated data available at http://www.molecularecologist.com/next-gen-fieldguide-2014/). Reagent costs per Gb of genetic data generated show that Ion Proton PI chips are cheaper (US$81.63) than Illumina GAIIx (min. US$97.54) and MiSeq (min. US $109.24); however, Ion Proton PI chips are more expensive

than NextSeq 500 (US$33.33–US$50.00), HiSeq 2500 (US $29.90–US$90.00) and HiSeq X (US$7.08). Forecasts of the PII (US$20.41) and PIII (US$11.43) chips dramatically improve the cost per Gb for Ion Proton and are comparable to or even exceed most current Illumina specifications.

However, one might consider using criteria other than cost per Gb to choose the optimal sequencing platform. For example, if conducting a pilot study or small-scale NGS genomic analysis, Ion Torrent has the lowest per run cost (Glenn 2011, updated data available at http://www.molecularecologist.com/next-gen-fieldguide-2014/). Here, we showed that ddRADseq-ion from one Ion Proton PI chip could generate up to 26 000 shared polymorphic SNPs for six individuals or ~7000 (6225–15 082, depending on genetic diversity) shared polymorphic SNPs for 30 individuals. The sequencing cost per genotype is then approximately 14 cents. An advantage of the Ion Torrent platform is the very short time needed for sequencing (2–4 h, e.g. Glenn 2011; Liu *et al.* 2012) and the customizable amount of data generated from the various PGM or Proton chips. This makes ddRADseq-ion particularly well qualified for pilot and small-scale genomic studies at its current state, and pending the availability of the PII and PIII chips, also for large-scale genomic studies.

Method validation via evolutionary analyses

We validated our ddRADseq-ion method using two approaches. In the first, we visualized the genomic composition of an interlineage cross of Alpine and Baltic whitefish and their offspring. The genomes of the two parents were identified as genetically distinct and their offspring contained approximately 50% of each parents' genome, as would be predicted (Fig. 3). This data set is based on 21 317 SNPs (of which 1356 were fixed between parents) generated from the equivalent of half (N = 16 individuals, mean coverage = 20.2) of a PI chip of Ion Proton sequencing. This analysis shows that the ddRADseq-ion methodology can effectively and efficiently characterize genetic variation at fine scales and with high resolution.

The second approach was to reconstruct phylogenetic relationships among whitefish and Arctic charr individuals from different regions. This evidences the ability of ddRADseq-ion to resolve higher level phylogenetic relationships (whitefish and charr are both in the family Salmonidae and are approximately 50 MY divergent; Crête-Lafrenière *et al.* 2012). The phylogeny conclusively separated both genera and elucidated intraspecific relationships between individuals from different regions in Eurasia (Fig. 4). This data set is based on 6036 SNPs generated from the equivalent of one-third of a PI chip (N = 6 individuals, mean coverage = 31.5) sequenced on Ion Proton. Bootstrap support for intraspecific relationships was slightly lower compared with the interspecific separation of charr and whitefish. Within Arctic charr, the Russian lineage was placed sister to the other two European lineages, as has been demonstrated previously based on mitochondrial DNA (Brunner *et al.* 2001). Phylogenetic resolution of European whitefish has been shown to be problematic previously (Østbye *et al.* 2005). Here, we found the Scottish lineage was sister to the Alpine and Baltic lineages. Additional biological sampling will be required to resolve the relationships among European whitefish.

In summary, both these approaches show that our ddRADseq-ion method produces data that can be used to address biological questions. The sequence data are robust, efficient, inexpensive and repeatable.

Recommendations

All next-generation sequencing platforms have a defined number of reads per sequencing job. Balancing the number of individuals, the number of loci and data quality is the most crucial step in designing any genomic project. These parameters should be adjusted in a trade-off depending on the type, quality and quantity of data sought. With the ddRADseq-ion protocol we present here and a typical 1–3 Gb vertebrate genome without a genomic reference, from a single PI chip one can expect ~30 000–80 000 catalogue loci at a minimum 15-fold coverage in 30 individuals and shared by at least 75% of individuals (Fig. 2).

Having an estimate of nucleotide diversity (or phylogenetic divergence between samples) may be helpful in designing the project, as it influences the number of SNP markers and number of homologous loci that can be identified between individuals per unit of sequencing effort. In optimizing library preparation, the number of loci can be varied depending on enzyme combinations (which we did not do here) and breadth and target of size selection.

The length to which reads should be trimmed during data processing to maximize data information can be estimated via the R script we provide (Appendix S4, Supporting information). However, in cases where sequencing coverage per locus (i.e. SNP confidence) might be more crucial than the absolute numbers of SNPs, shorter reads should be used. In other cases, it might be more useful to have longer reads to minimize confusion of homologous and paralogous sequences.

We suggest that, as in all genetic investigations, technical replicates should be included to estimate errors and to choose optimal bioinformatic pipeline parameters. The number of replicates to be included is of course a trade-off that can be balanced with the

number of biological replicates in the study, the stress on a particular question might have on minimizing genotyping errors, and the experience of the researchers with the laboratory and informatics pipeline.

Ion Torrent sequencing technologies have relatively low per base pair sequencing costs. Although this is somewhat offset by the high amount of data that are discarded, as explained above, the overall cost is nonetheless only pennies a genotype. A further benefit is that the Ion Torrent system is very scalable using identical chemistry and adapters: from ~5 M reads per run on a PGM to ~80 M reads on a Proton with PI chip. The pending PII chip is promising to deliver threefold more reads for similar cost as PI and would make Proton less expensive than most current platforms per Gb of data (http://www.molecularecologist.com/next-gen-fieldguide-2014/), but to date the PII's release has been much delayed. Nonetheless, at present, a benefit of Ion Torrent's scalability is that pilot assessments can be cost effectively tested on small numbers of individuals. This is especially powerful for researchers who have local access to the platform.

The expense of any reduced representation sequencing is a combination of adapter cost (determined by level of multiplex required) and per base pair sequencing cost, relative to the amount of data retained. We do not outline these costs in detail here, as they are constantly changing and highly regionally dependent. From our experience to date, ddRADseq-ion is overall a similar cost to Illumina RADseq per informative locus at modest scales, with the benefit that fewer costly adapters need to be purchased upfront because multiplex pools are smaller (at least for PI chips). A disadvantage of ddRADseq-ion is that sequencing cannot be paired end.

We consider ddRADseq-ion's strongest current applications to be for pilot assessments and quick data return, for example, to optimize restriction enzyme combinations and/or size selection parameters, to assess nucleotide diversity in a population to efficiently design a larger scale experiment or to conduct small projects such as undergraduate and master's research on short-time scales and limited budgets. In many cases and for many researchers, optimizing and pilot projects on high-throughput platforms is neither particularly feasible nor time effective. Library optimizing parameters (enzymes, target size) will be robust to changes in platform, so that one could pilot with ddRADseq-ion and then implement the full study on a high-throughput platform such as Illumina HiSeq. Because of its scalability and potential for rapid in-house optimizing, we expect the ddRADseq-ion method will be especially useful to scientists who already have an Ion Torrent platform on hand.

Here, we show that the ddRADseq-ion method is a valuable and useful addition to the molecular ecologist's toolkit. Our method can successfully genotype—for example for genetic mapping, population genomics and phylogenomics—quickly, robustly and cost effectively.

Acknowledgements

We thank J. Wanzenboeck, M. Garduño-Paz, C. Adams, N. Gordeeva, S. Aleksyev, W. Mayer, B. Murphy, H. Thorarensen and O. Hooker for providing samples, and K. Gharbi, T. Casci and B. Mable for helpful discussions. We also thank J. Galbraith, J. Wang and A. Adam for help with library preparation and sequencing. This work was funded by a Marie Curie CIG (grant no. 32199), University of Glasgow start-up funding, a John Robertson Bequest Fund of the University of Glasgow and a Glasgow Polyomics ISSF Pilot Grant to KRE.

References

Andolfatto P, Davison D, Erezyilmaz D *et al.* (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research*, **21**, 610–617.

Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, **3**, e3376.

Barrett RDH, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews Genetics*, **12**, 767–780.

Baxter SW, Davey JW, Johnston JS *et al.* (2011) Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS One*, **6**, e19315.

Brunner PC, Douglas MR, Osinov A, Wilson CC, Bernatchez L (2001) Holarctic phylogeography of Arctic charr (*Salvelinus alpinus* L.) inferred from mitochondrial DNA sequences. *Evolution*, **55**, 573–586.

Carmichael SN, Bekaert M, Taggart JB *et al.* (2013) Identification of a sex-linked SNP marker in the salmon louse (*Lepeophtheirus salmonis*) using RAD sequencing. *PLoS One*, **8**, e77832.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*, **1**, 171–182.

Crête-Lafrenière A, Weir LK, Bernatchez L (2012) Framing the Salmonidae family phylogenetic portrait: a more complete picture from increased taxon sampling. *PLoS One*, **7**, e46662.

Davey JW, Hohenlohe PA, Etter PD, *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Davey JW, Cezard T, Fuentes-Utrilla P *et al.* (2013) Special features of RAD sequencing data: implications for genotyping. *Molecular Ecology*, **22**, 3151–3164.

Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, **6**, e19379.

Emerson KJ, Merz CR, Catchen JM *et al.* (2010) Resolving post-glacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences USA*, **107**, 16196–16200.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, **11**, 759–769.

Gregory TR (2014) Animal Genome Size Database. http://www.genomesize.com.

Henning F, Lee HJ, Franchini P, Meyer A (2014) Genetic mapping of horizontal stripes in Lake Victoria cichlid fishes: benefits and pitfalls of using RAD markers for dense linkage mapping. *Molecular Ecology*, **23**, 5224–5240.

Hudson AG, Vonlanthen P, Seehausen O (2011) Rapid parallel adaptive radiations from a single hybridogenic ancestral population. *Proceedings of the Royal Society of London Series B*, **278**, 58–66.

Lemmon EM, Lemmon AR (2013) High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, **44**, 99–121.

Lexer C, Wüest RO, Mangili S *et al.* (2014) Genomics of the divergence continuum in an African plant biodiversity hotspot, I: drivers of population divergence in *Restio capensis* (Restionaceae). *Molecular Ecology*, **23**, 4373–4386.

Li WH, Gu Z, Wang H, Nekrutenko A (2001) Evolutionary analyses of the human genome. *Nature*, **409**, 847–849.

Liu L, Li Y, Li S *et al.* (2012) Comparison of next-generation sequencing systems. *BioMed Research International*, **2012**, 251364.

Liu MM, Davey JW, Banerjee R *et al.* (2013) Fine mapping of the pond snail left-right asymmetry (chirality) locus using RAD-seq and fibre-FISH. *PLoS One*, **8**, e71067.

Loman NJ, Misra RV, Dallman TJ *et al.* (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, **30**, 434–439.

Mascher M, Wu S, Amand PS, Stein N, Poland J (2013) Application of genotyping-by-sequencing on semiconductor sequencing platforms: a comparison of genetic and reference-based marker ordering in barley. *PLoS One*, **8**, e76925.

Mastretta-Yanes A, Arrigo N, Alvarez N *et al.* (2014) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, **15**, 28–41.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66**, 526–538.

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, **22**, 2841–2847.

Østbye K, Bernatchez L, Næsje TF, Himberg M, Hindar K (2005) The evolutionary history of European whitefish (*Coregonus lavaretus* L.) as inferred from mtDNA phylogeography and gillraker numbers. *Molecular Ecology*, **14**, 4371–4387.

Palaiokostas C, Bekaert M, Davie A *et al.* (2013) Mapping the sex determination locus in the Atlantic halibut (*Hippoglossus hippoglossus*) using RAD sequencing. *BMC Genomics*, **14**, 566.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*, **7**, e37135.

Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. *The Plant Genome*, **5**, 92–102.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Puritz JB, Matz MV, Toonen RJ *et al.* (2014) Demystifying the RAD fad. *Molecular Ecology*, **23**, 5937–5942.

Quail MA, Smith M, Coupland P *et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.

Recknagel H, Elmer KR, Meyer A (2013) A hybrid genetic linkage map of two ecologically and morphologically divergent Midas cichlid fishes (*Amphilophus* spp.) obtained by massively parallel DNA sequencing (ddRADSeq). *G3: Genes, Genomes, Genetics*, **3**, 65–74.

Rothberg JM, Hinz W, Rearick TM *et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348–352.

Rowe HC, Renaut S, Guggisberg A (2011) RAD in the realm of next-generation sequencing technologies. *Molecular Ecology*, **20**, 3499–3502.

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

Stapley J, Reger J, Feulner PGD *et al.* (2010) Adaptation genomics: the next generation. *Trends in Ecology & Evolution*, **25**, 705–712.

Vandepitte K, Honnay O, Mergeay J *et al.* (2013) SNP discovery using Paired-End RAD-tag sequencing on pooled genomic DNA of *Sisymbrium austriacum* (Brassicaceae). *Molecular Ecology Resources*, **13**, 269–275.

K.R.E., P.H. and H.R. designed the study. H.R. designed the adapters. P.H., H.R. and A.J. did the sequencing. H.R. and A.J. performed protocol optimizations, library preparations and analysed the data. H.R., A.J. and K.R.E. wrote the manuscript. All authors contributed to and finalized the manuscript.

Data accessibility

Raw read sequence files (in fastq format) for each library can be accessed from the sequence read archive (SRA) on NCBI (PRJNA276094). Table S1 lists the respective accession number, barcode and individual information needed to demultiplex libraries. SNP files for each library, the STRUCTURE data, the phylogenetic data and tree, and plink files of technical replicates are available through DRYAD (doi:10.5061/dryad.7tb72).

Supporting Information

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

Appendix S1 Information on individuals used for library preparation.

Appendix S2 Adapter and primer sequences used for ddRADseq-ion.

Appendix S3 Detailed bench protocol for ddRADseq-ion library preparation and sequencing.

Appendix S4 R script to estimate optimal (highest number of retained DNA base pairs) trimming threshold for individual ddRADseq-ion libraries.

Appendix S5 The distribution of fragment lengths for 13 ddRADseq-ion libraries sequenced on the Ion Proton semiconductor sequencer with P1 chips.