MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2013) **13**, 899–907 doi: 10.1111/1755-0998.12137

Mass production of SNP markers in a nonmodel passerine bird through RAD sequencing and contig mapping to the zebra finch genome

YANN X. C. BOURGEOIS,* EMELINE LHUILLIER,†‡ TIMOTHÉE CÉZARD,§ JORIS A. M. BERTRAND,* BORIS DELAHAIE,* JOSSELIN CORNUAULT,* THOMAS DUVAL,¶ OLIVIER BOUCHEZ,‡** BORJA MILÁ†† and CHRISTOPHE THÉBAUD*

**Laboratoire Évolution et Diversité Biologique, UMR 5174 CNRS - Université Paul Sabatier – ENFA, 118 route de Narbonne, Bâtiment 4R1, F-31062 Toulouse Cedex 9, France, †INRA, UAR 1209 Département de Génétique Animale, INRA Auzeville, F-31326 Castanet-Tolosan, France, ‡GeT-PlaGe, Genotoul, INRA Auzeville, F-31326 Castanet-Tolosan, France, §The GenePool, Ashworth Laboratories, The University of Edinburgh, The King's Building, Edinburgh EH9 3JT, UK, ¶Société Calédonienne d'Ornithologie Nord, BP 236, F-98822 Poindimié, Nouvelle Calédonie, France, **INRA, UMR 444 Laboratoire de Génétique Cellulaire, INRA Auzeville, F-31326 Castanet-Tolosan, France, ††Museo Nacional de Ciencias Naturales, CSIC, José Gutiérrez Abascal 2, Madrid 28006, Spain*

Abstract

Here, we present an adaptation of restriction-site-associated DNA sequencing (RAD-seq) to the Illumina HiSeq2000 technology that we used to produce SNP markers in very large quantities at low cost per unit in the Réunion grey white-eye (*Zosterops borbonicus*), a nonmodel passerine bird species with no reference genome. We sequenced a set of six pools of 18–25 individuals using a single sequencing lane. This allowed us to build around 600 000 contigs, among which at least 386 000 could be mapped to the zebra finch (*Taeniopygia guttata*) genome. This yielded more than 80 000 SNPs that could be mapped unambiguously and are evenly distributed across the genome. Thus, our approach provides a good illustration of the high potential of paired-end RAD sequencing of pooled DNA samples combined with comparative assembly to the zebra finch genome to build large contigs and characterize vast numbers of informative SNPs in nonmodel passerine bird species in a very efficient and cost-effective way.

Keywords: next-generation sequencing, passerine, pooled DNA, SNP detection, zebra finch genome, *Zosterops*

Received 26 March 2013; revision received 24 May 2013; accepted 4 June 2013

Introduction

The rapid development of new sequencing technologies has brought much hope to identify the allelic variants that underlie phenotypic variation and divergence in natural populations of nonmodel species (Davey *et al.* 2011; Radwan & Babik 2012; but see Bierne *et al.* 2011; Rockman 2012; Travisano & Shaw 2013). However, unravelling the molecular causes of phenotypic changes, reconstructing the demographic histories of multiple populations or quantifying fine-scale gene flow require genome-wide sequence data from multiple individuals, something which remains difficult to achieve for most nonmodel species. Strategies based on genome reduction are thus of great interest when attempting to link genetic and phenotypic variants, as they can provide substantial population genomic data for a large number of individuals at a reasonable cost (Davey *et al.* 2011).

The first protocols of genome reduction have included the development of Reduced Representation Libraries (RRLs) or RNA sequencing (Altshuler *et al.* 2000; Wang *et al.* 2009). One problem with the RRL approach is that it usually requires extensive testing prior to identifying orthologous single-copy loci that can be compared between different individuals or populations (van Tassell *et al.* 2008; van Bers *et al.* 2010). RNA sequencing can yield much information about coding mutations and relative expression levels, especially in combination with other approaches (Wang *et al.* 2009; Hawkins *et al.* 2010). However, it can be difficult to implement in nonmodel organisms because obtaining and preserving RNA can be challenging, especially in field conditions. It also excludes noncoding regions that can be of interest.

In this context, Restriction-site-Associated DNA sequencing or RAD sequencing has become a method of

Correspondence: Yann X. C. Bourgeois, Fax: +33 (0)5 61 55 73 27; E-mail: yann.x.c.bourgeois@gmail.com

choice for high-density single-nucleotide polymorphism (SNP) discovery and genotyping across many individuals in many populations (Davey & Blaxter 2010). RAD sequencing allows reducing genome complexity by sequencing the same loci across the genome in several individuals, facilitating among-individual comparisons and limiting sequencing investment. The approach consists in cleaving double-stranded genomic DNA with a restriction enzyme chosen to obtain an appropriate sequencing depth. Then, DNA fragments are randomly sheared to a specific length that varies depending on which next-generation sequencing (NGS) platform is used. Using specific adapters, it is thus possible to selectively amplify the regions flanking the restriction sites. Paired-end reads for each RAD tag can then be assembled into long contiguous sequences (Hohenlohe *et al.* 2011). While RAD sequencing was initially designed for microarray (Miller *et al.* 2007), it has been quickly adapted to NGS technology (Baird *et al.* 2008) and has opened up a wide range of applications in evolutionary genomics (e.g. Emerson *et al.* 2010; Gagnaire *et al.* 2012; Hess *et al.* 2012; Keller *et al.* 2012; Takahashi *et al.* 2012; Wang *et al.* 2012), most notably in species that have a reference genome.

Here, we present an efficient and cost-effective protocol for developing RAD markers by adapting previous protocols (Baird *et al.* 2008) to the Illumina Hiseq2000 technology, which allows the sequencing of 150 to 180 million paired-end reads per lane for a reduced price per base pair, compared with the 40 million produced by a Genome Analyzer IIx. Technological advances now allow sequencing at substantial depth tens to hundreds of libraries per sequencer run (Davey & Blaxter 2010) and have been shown to be of interest in most recent studies using RAD sequencing (Davey *et al.* 2012; Keller *et al.* 2012; Peterson *et al.* 2012; Wagner *et al.* 2013).

The present study aimed at generating a very large number of SNP markers in a small passerine bird endemic to Réunion (Mascarene Islands, southwestern Indian Ocean), the Réunion grey white-eye (*Zosterops borbonicus*). To deal with small genomic DNA quantities that are typically obtained from field-collected blood samples, while estimating allele frequencies at a genome-wide scale and keeping cost as low as possible, we sequenced multiple libraries, each corresponding to pools of individuals. Pooling has been shown to be a very cost-effective approach to estimate allele frequencies at a large number of SNPs for many individuals in multiple populations (Futschik & Schlötterer 2010), when using whole genome sequencing (Kolaczkowski *et al.* 2011; Turner *et al.* 2011; Boitard *et al.* 2012) or reduced representation libraries (van Tassell *et al.* 2008; Pérez-Enciso & Ferretti 2010). Combining pooling and reduced representation is thus an affordable way to obtain a large number of informative SNPs.

Because information about distribution of SNPs across the genome is important for further population genomic analyses, mapping contigs to a reference genome remains critical. By taking advantage of the high degree of genome stability in birds (Backström *et al.* 2008; Griffin *et al.* 2008; Warren *et al.* 2010), we were able to build and map white-eye contigs to the zebra finch (*Taeniopygia guttata*) genome (Warren *et al.* 2010). In this note, we describe our protocol from library construction to contig mapping, with an emphasis on how to fully use information from paired-end reads, and evaluate its performance with regard to mass-producing SNP markers in our study species.

Methods

Library construction

DNA was extracted from individual blood samples using the QIAGEN DNeasy® Blood and Tissue kit, following manufacturer's instructions. Six pools were prepared, each including genomic DNA from 18 to 25 individuals and representing two replicates with different individuals from three distinct localities (named 'Bois Ozoux', 'Tévelave' and 'Pas de Bellecombe'). To minimize the risk of high variance in the number of reads per individual within a pool, we assessed double-stranded DNA concentration for each individual sample using the Quant-iT™ dsDNA Assay Kit (Invitrogen) and made the necessary adjustments to bring each individual DNA in the pool to equal molar concentration. Each library was built with equally represented samples for a total amount of 3 μg of total genomic DNA in a final volume of 75 μL.

To prepare RAD libraries, genomic DNA was digested with *EcoRI*, a widely used enzyme, cutting frequently enough so that a good coverage of the *Z. borbonicus* genome (around 300 000 restriction sites) could be obtained without compromising sequencing depth per locus. Each 75 μL-library was digested using the Promega *EcoRI* restriction reagents. For each reaction, 12 μL of H buffer, 30 μL of pure water and 3 μL (36 units) of enzyme were added (final volume: 120 μL). The reaction mix was divided into three 40 μL aliquots, and digestion was performed at 37 °C overnight, ending with a 20-min deactivation step at 65 °C. The three aliquots from each library were then progressively cooled at 4 °C and pooled. Based on the protocol by Baird *et al.* (2008) and customizing the sequences given by Illumina (oligonucleotide sequences © 2007-2012 Illumina, Inc., all rights reserved), we built new P1 and P2 adapters compatible with the paired-end technology currently supported by the Illumina HiSeq2000 system (whereas the adapters by

Baird and collaborators were designed for a previous version of single-read technology on the Genome Analyzer IIx system). Our P1 adapter, which includes the *EcoRI* restriction site, contains reverse amplification and Illumina sequencing primer sites, as well as a six base-long barcode sequence for sample identification, following the design of the TruSeq indexed adapters of Illumina (Table 1). Barcodes differed by at least four nucleotides to avoid misidentification of samples. This design allows the barcode to be read independently of the two reads of the genomic insert, in contrast to previous protocols in which the barcode was read together with the genomic fragment (*e.g.* Etter *et al.* 2011), avoiding subsequent barcode trimming from the raw reads. Barcodes were chosen among the 24 Illumina TruSeq Barcodes in order to be used with the Illumina TruSeq PCR Kit and to be easily read by the HiSeq2000 during the barcode read of the run. Our P2 adapter contains forward amplification and Illumina sequencing primer sites, following the design of the Illumina TruSeq Universal Adapter. As previously described (Baird *et al.* 2008), an asymmetric design of the P2 adapter ensured that only P1-ligated fragments could be amplified during the final amplification step (Table 1).

For each adapter, stocks of nonannealed oligonucleotides (Table 1) were diluted at 100 μM in 1× elution buffer (10 mM Tris-Cl, pH 8.5). Then, the pairs of forward and reverse oligonucleotides for each adapter were combined at 10 μM in 1× AB buffer (10× AB: 50 mM NaCl, 10 mM Tris-Cl, pH 8.0). Each stock of adapters was denatured 2 min at 95 °C and slowly cooled for 45 min at room temperature to obtain double-stranded adapters. P1 adapters were then diluted at a final concentration of 100 nM in 1× AB buffer. P2 adapters were used at a 10 μM concentration. 40 μL ligations were performed using the Promega High Concentration T4 DNA ligase, adding 18 μL of T4 DNA ligase Buffer (10×), 6 μL of NaCl (500 mM) and 18 μL purified water. Salt was added to ensure double-stranded adapters stability. After a 15 min incubation step at room temperature, 15 μL of 100 nM P1 adapters and 30–60 units of HC T4 DNA ligase (3 μL) were added to the mix. The reaction mix was divided into 60 μL aliquots and incubated at 22 °C for three hours, then deactivated for 10 min at 70 °C. For each library, aliquots were then pooled, purified using the QiaQuick PCR purification kit and eluted in a final volume of 100 μL.

Fragmentation of digested DNA was performed by sonication on a Bioruptor (Diagenode), using 10 cycles of 30 s on and 90 s off, in high mode. Control of sonication was done by checking fragment sizes with Bioanalyzer. Purification on Agencourt AMPure XP beads (Beckman-Coulter) was then performed, and the DNA of each library was eluted in 25 μL Resuspension Buffer (Illumina). Fragments around 500 bp (± 150 bp) were selected on an E-Gel system (E-Gel® CloneWell 0.8% SYBR Safe™ gel, Life Technologies) and retrieved in 25 μL Resuspension Buffer. Fragment end repair and adenylation were performed using Illumina TruSeq DNA Sample Preparation kit and guidelines. After another AMPure purification and elution in 45 μL EB buffer, the P2 ligation was performed by adding 5.8 μL of T4 DNA ligase buffer (10×), 5.8 μL of NaCl (500 mM), 1 μL of P2 adapter (10 μM) and 10 units of high

Table 1 Modified Illumina© adapters (a) used in this study (Oligonucleotide sequences © 2007-2012 Illumina, Inc., all rights reserved). In the P1 oligos: underlined nucleotides correspond to the overhanging end of the *EcoRI* restriction products; '**XXXXXX**' ('**YYYYYY**' for reverse strand) refers to the index-sequence or barcode. The sequences of barcodes used are given in (b) with corresponding localities. P2 adapter is designed to allow the amplification only of P1-linked DNA fragments. [PHO] designs the addition of a phosphate group in 5', * indicates the addition of a phosphorothioate bond to enhance nuclease resistance

Oligonucleotide	Sequence
(a)	
P1 forward	[PHO]AATTAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC**XXXXXX**ATCTCGTATGCCGTCTTCTGCTTG
P1 reverse	CAAGCAGAAGACGGCATACGAGAT**YYYYYY**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
P2 forward	[PHO]GATCGGAAGAGCGTCGTG
P2 reverse	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T

Barcode	Locality	Latitude (degrees)	Longitude (degrees)	Number of individuals
(b)				
ATCACG	Bois Ozoux	−21.198	55.647	18
TTAGGC	Bois Ozoux	−21.198	55.647	25
ACTTGA	Tévelave	−21.169	55.387	24
GATCAG	Tévelave	−21.169	55.387	20
TAGCTT	Pas de Bellecombe	−21.217	55.688	25
GGCTAC	Pas de Bellecombe	−21.217	55.688	25

concentration ligase (0.5 μL). After two other AMPure purifications and concentration in 20 μL Resuspension Buffer, an aliquot was kept as a control on Bioanalyzer.

An underappreciated issue of RAD sequencing concerns the amount of genomic material that is needed for library construction because *ca*. 1 μg of good quality genomic DNA is typically required per library (Baird *et al.* 2008). Using the Illumina HiSeq2000 technology requires even greater DNA quantities (3 μg per library) in order to satisfy quality standards. While this could be seen as an inconvenience, it makes possible to reduce the number of PCR cycles needed to obtain usable libraries, thereby reducing PCR amplification biases that can be especially problematic when dealing with samples with low DNA concentration. Thus, an enrichment step was performed for each library, consisting in just 12 cycles of PCR amplification, using TruSeq Sample Prep PCR Kit and guidelines from Illumina. After AMPure purification, a final step of size selection was then performed on the libraries to remove the remaining adapters, using the E-Gel system again.

Library profiles were controlled on a BioAnalyzer High Sensitivity chip. Finally, quantities of usable material for each of the six libraries were estimated by qPCR (KAPA Library Quantification Kit–Illumina Genome Analyzer-SYBR Fast Universal) and then normalized and pooled. The quality of the pool was then checked using qPCR and immediately followed by sequencing on the HiSeq2000 platform (Plateforme Génomique - Genopole Toulouse Midi-Pyrénées), using TruSeq PE Cluster Kit v3 (2 × 100 pb) and TruSeq SBS Kit v3.

Assembling contigs and SNP detection

Large contigs are required to perform accurate alignments to a related genome. In the case of the white-eye, the divergence time to the zebra finch is estimated to be 40 million years (Barker *et al.* 2004), which prevents the use of short-read alignment tools such as BWA (Li & Durbin 2009). Therefore, we used a pipeline aimed at assembling large contigs (300–500 bp), using information from both paired-end reads. First ustacks (version 0.9995) and then cstacks (Catchen *et al.* 2011) were used to group reads immediately flanking restriction sites (reads 1) for each of the six libraries, allowing up to 3 mismatches between stacks (ustacks options: -m 4 -M 3). Loci not found in at least 5 of the 6 libraries were discarded. Both reads 1 and reads 2 were aligned with BWA (version 0.7.0) on this catalogue of stack, forcing the alignment of the second read by using a custom python script (RAD_assign_reads_to_consensus.py, all scripts available at the address https://github.com/tcezard/RADmapper). The resulting .bam file was then translated into several .fastq files, each corresponding to one stack, using another python script (RAD_bam_to_fastq.py). For each locus, a consensus of reads 2 was then assembled with the consensus of reads 1, using a third python script (RAD_assemble_read2.py) making use of the IDBA_UD assembler (Peng *et al.* 2012; version 1.0.9), which is a fast assembler originally designed for single-cell assembly, not relying upon an even coverage and using several k-mer lengths. Reads 1 and reads 2 consensuses were merged into a single contig with EMBOSS (version 6.4.0.0) merger (Rice *et al.* 2000). Consensuses that did not overlap were forced into one contig by adding ten ambiguous bases ('N') between them. Contigs with the best assembly score were then extracted and used as a reference for mapping back reads and eliminating PCR duplicates with SAMTOOLS (Li *et al.* 2009, version 0.1.19). Options for the java script MarkDuplicates were as follows: VALIDATION_STRINGENCY=LENIENT, MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=100, CREATE_INDEX=true, AS=true.

The mpileup2sync.jar java script (version 1.201) from POPOOLATION2 (Kofler *et al.* 2011b) was used to construct files with allele counts (synchronized 'pileup' files) with the following options: –fastq-type sanger –min-qual 20 –threads 8. To call biallelic SNPs, we applied a standard quality threshold of 20 (corresponding to an error probability of less than 1%), a global minimum allele count (MAC) of 2 or 3, a minimum sequencing depth of 10× or 20× by library (60× or 120× overall) and a maximum depth of 500×.

We checked that contigs did not contain an excess of SNPs because poor assembly and bad alignment of the reads can lead to a heterogeneous distribution of SNPs over contigs. Finally, we further assessed whether those SNPs could be due to massive sequencing errors by testing whether the same SNPs could be found in multiple libraries and in the two replicates from each locality. We performed SNPs calling by applying a MAC of three and a minimal sequencing depth of 20×.

LASTZ alignment

We assessed the quality of contig assembly by testing whether the contigs that were obtained could be mapped easily onto the zebra finch genome. To this end, all the contigs assembled from paired-end reads were aligned against the zebra finch genome (version July 2008, assembly WUGSC v.3.2.4). Alignment was performed using LASTZ (Harris 2007), an improved version of BLASTZ (Schwartz *et al.* 2003) using default parameters, except for 'ambiguous=n' and ydrop=7000, mainly to allow large gaps (up to 220 bp) inside contigs because consensuses from paired reads sometimes did not overlap. A minimum identity of 60% and coverage of 70% were required.

Results

Mapping of reads on a related genome

More than 154 million usable paired reads (2 × 100 bp) were generated from the six libraries using only one Illumina HiSeq2000 lane, with a mean sequencing depth of 27× per library at the end defined by the restriction site (SD between libraries = 2.83). From these reads, we generated a total of 606 725 contigs, 99% ranging in size from 101 to 612 bp (mean= 396 bp, median = 384 bp; number of contigs larger than 100 bp = 592 712; number of contigs larger than 300 bp = 582 193; N50 = 402 bp). Eighty-six per cent of these contigs could be mapped to the zebra finch genome (Table 2). After excluding all sites associated with a chromosome but with no known position ('random' chromosomes) and nonmapped repetitive sequences ('Unknown' chromosome), we found that 398 793 contigs were unique hits, and 386 841 (63.8% of all contigs) could be positioned unambiguously. A fraction of the contigs (14.5%) mapped to only two distinct locations in the zebra finch genome. Nearly two-thirds of these contigs mapped onto a known chromosome and also onto the 'Unknown' chromosome, suggesting a possible location in repetitive regions.

Having started with a relatively low-identity requirement (60.0%), we finally obtained a large set of contigs unambiguously aligned displaying between 80.0 and 100.0% identity (mean: 89.9%; SD: 3.5%) with the corresponding zebra finch sequence. In birds, substitution rates in autosomes have been estimated at 3.6×10^{-9} substitution per year per site (Axelsson *et al*. 2004). Using a divergence time of ca. 40 MYA (Barker *et al*. 2004), we would expect a sequence divergence of 14.0% between zebra finch and *Zosterops*, a figure that is consistent with our data.

Table 2 Counts of contigs mapped onto the zebra finch genome. Repartition of double hits and unique hits onto the zebra finch genome are detailed

Category of hits	Count	Percentage
Total	606 725	100
Mapping onto zebra finch	523 744	86.32
More than two hits	37 113	6.12
Total of double hits	87 838	14.48
Double hits with one single hit on unknown chromosome	60 539	9.98
Total of unique hits	398 793	65.73
Known position	386 841	63.76
Known chromosome (random)	9521	1.57
Unknown chromosome	2431	0.40

Distribution of contigs across chromosomes

We observed a strong correlation between chromosome size and the number of contigs mapped unambiguously (Fig. 1, $R^2 = 0.987$, P-value <0.0001). We did not observe any obvious outliers and were able to map contigs even on the smallest assembled chromosomes, such as chromosomes 16 or 1B. This indicates that contigs and associated SNPs were evenly distributed across the white-eye genome.

SNP calling

The number of reads that qualified as PCR duplicates accounted for less than 23.0% of the whole data set, a figure that remains low compared to some other RAD-sequencing studies (M. Gautier, personal communication). This is probably due to the fact that we used relatively large amounts of genomic DNA (~3 μg) and performed no more than 12 amplification cycles. Also, we used several stringency criteria to call SNPs (Table 3). Using a MAC of three and a minimum sequencing depth of 20× in each library after quality and duplicate filtering, we were able to identify 133 958 SNPs. Of these, 81 246 SNPs (60.7%) could be mapped unambiguously, which is consistent with the proportion of contigs with a unique hit at a known position (63.8%) on the zebra finch genome.

Minor allele frequencies (MAF) for the whole data set were mainly below 0.2, with a mean frequency of 0.103 (SD: 0.116) when using the most stringent conditions for SNP calling. Reducing sequencing depth to at least 10× per library did not drastically change the MAF distribution (mean frequency: 0.114, SD: 0.118). When

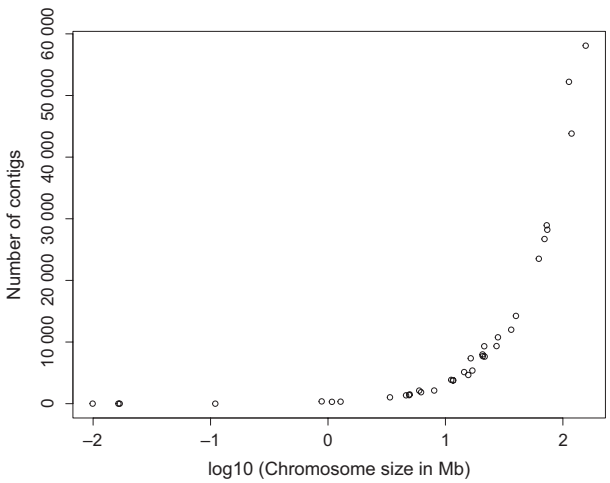

Fig. 1 Correlation between chromosome size and number of unambiguous contig hits.

considering each library, no obvious differences in MAF distributions could be observed between replicates from the same locality (Fig. 2a). Differences in allele frequencies were low (means of 0.061, 0.067 and 0.066 for 'Bois Ozoux', 'Tévelave' and 'Pas De Bellecombe' libraries, respectively) and in the range of sampling noise (Fig. 2b). A total of 118 683 (89%) SNPs were found in a minimum of two libraries (Table 4), and most loci found to be polymorphic in a given replicate were also polymorphic in the other replicate from the same locality. On average, each polymorphic contig contained 2 SNPs (Fig. 3; median = 1, SD = 2.36), even when considering SNPs sampled at a lower depth (mean = 2.54, median = 2, SD = 2.72). This further suggests that contigs were correctly assembled and that multiple SNPs on contigs were not due to poor alignment, but rather to contig length.

Table 3 Number of SNPs called following several stringency criteria

Minimal sequencing depth by library	MAC	
All SNPs	2	3
10×	397 969	327 705
20×	163 086	133 958
SNPs at a known position		
10×	244 384	199 955
20×	99 821	81 246

MAC, minimum allele count required to call a SNP.

Discussion

There has been an increasing interest in the use of next-generation sequencing to address evolutionary questions in nonmodel species (Davey *et al.* 2011). However, the lack of a reference genome, the costs associated with the sequencing of many individuals and the need to compare homologous sequences across individuals have remained limiting when trying, for example, to associate SNPs to genes and traits of interest.

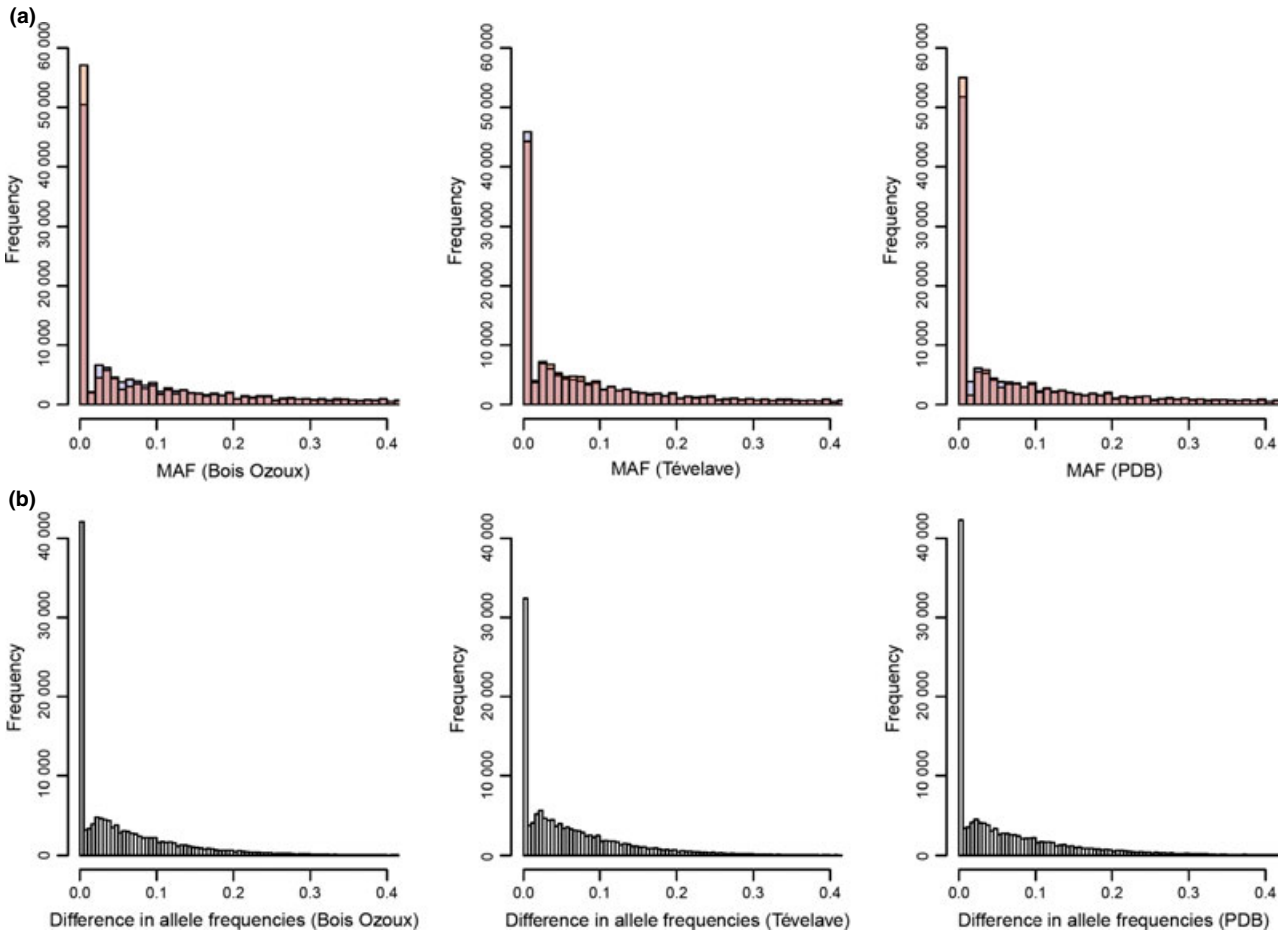

Fig. 2 Distribution of minimum allele frequencies (MAF) for each library used in this study, grouped by population (a). Changes in allele frequencies between libraries from the same population are also plotted (b). PDB: Pas de Bellecombe.

Table 4 Comparison of allele frequencies and count of shared polymorphisms in two replicates for each of the three populations. Estimates are based on SNPs obtained using a minimum sequencing depth of 20× and a MAC of 3

Population	Mean difference in allele frequencies	Median difference in allele frequencies	Standard deviation	Number of SNPs polymorphic in one library	Number of SNPs polymorphic in both libraries
Bois Ozoux	0.061	0.036	0.076	95 404	65 504
Tévelave	0.067	0.043	0.076	105 852	72 727
PDB	0.066	0.036	0.083	95 090	66 726

Our HiSeq2000-based RAD sequencing protocol, by making use of the very large and cost-efficient production of paired-end sequences for pools of individuals, has enabled us to build a very large number of 300–500-bp contigs that could be mapped unambiguously onto the zebra finch genome. This led to the discovery of more than one hundred thousand SNPs across 137 individuals sampled in three relatively close localities, separated by less than 35 km.

Through our experiment, we also confirm the usefulness of the RAD sequencing approach to identify the genomic position of markers in passerine birds, even when dealing with nonmodel species. This had been previously suggested by van Bers *et al.* (2010) in their study of the great tit (*Parus major*). However, comparing their results to ours, we note that we detected six times more SNPs and were able to map nearly 20 times more SNPs to unique locations distributed over the zebra finch genome. This was mostly due to the fact that we obtained many more contigs and that a much larger proportion of these contigs were greater than 100 bp (3.5% versus 97.7%). Thus, our approach based on pooled DNA samples, which keeps the cost of library construction and adapter preparation to a minimum, and HiSeq2000 technology provides a cost-effective strategy for SNP detection and mapping in a passerine bird species for which a sequenced genome is currently lacking.

While haplotype information is obviously not available, DNA pools produce a high number of informative SNPs that are ideal for characterizing variation in population samples or can subsequently be assayed in further experiments. Because synteny is remarkably conserved in birds (Derjusheva *et al.* 2004; Griffin *et al.* 2008), a large number of these SNPs can be readily identified as orthologous of known genes in the zebra finch genome, providing unprecedented opportunities to describe the molecular background of nonmodel passerine birds. Depending on the pooling strategy, the approach provides a wide range of applications that will enable students in ecology and evolution to have access to genome-wide allele frequency estimates, to compare patterns of differentiation on a genomic scale, to

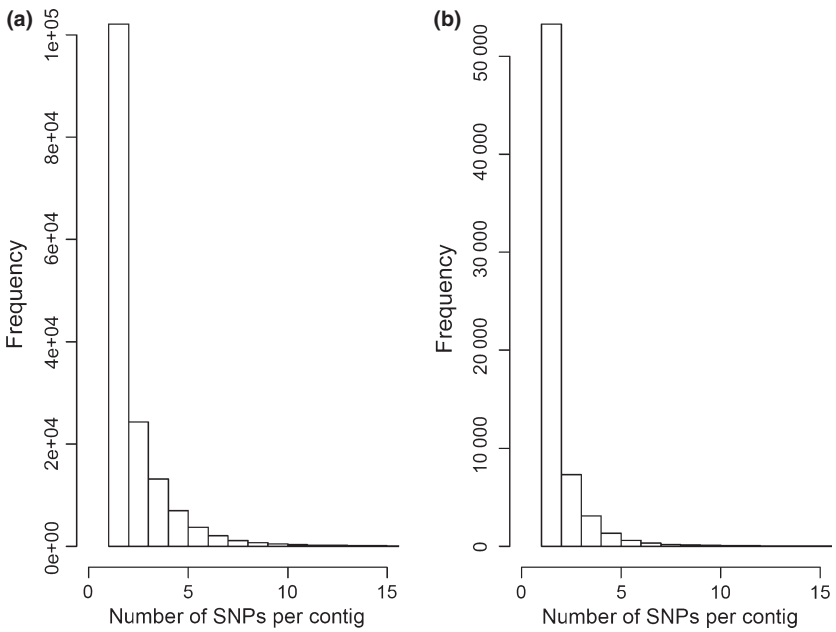

Fig. 3 Distribution of the number of SNPs per polymorphic contig for a sequencing depth of 10× and a minimum allele count (MAC) of 2 (a) and for a sequencing depth of 20× and a MAC of 3 (b).

characterize the demographic history of differentiated populations and detect selective sweeps, or even to investigate the genetic basis of ecologically significant traits using genome-wide association mapping (e.g. Boitard *et al.* 2012; Rubin *et al.* 2010; Willing *et al.* 2011; Kofler *et al.* 2011a; Futschik & Schlötterer 2010; Zhu *et al.* 2012; Gautier *et al.* 2013). This may pave the way to other approaches based on genotyping by sequencing that will then allow to get individual genotypes and to characterize molecularly precise variants within a population (see e.g. Garroway *et al.* 2013; Hagen *et al.* 2013).

Acknowledgements

Ben Warren, Guillaume Gélinaud, Dominique Strasberg, Juli Broggi, Magali Thierry, René-Claude Billot, Jean-Michel Probst, Isabelle Henry, Vincent Leconte, Marc Salamolard, Benoît Lequette, We gratefully acknowledge the Réunion National Park for permission to conduct fieldwork. We thank Mathieu Gautier for his insights about pooling experiments. Ulrich Knief and an anonymous reviewer provided valuable comments that greatly improved an earlier version of this manuscript. This work was supported by Institut Français de la Biodiversité (IFB), Agence Française pour le Développement (AFD) and ANR Biodiversity Program grants to CT, the Génopole Toulouse Midi-Pyrénées, the National Geographic Society and the 'Laboratoire d'Excellence' TULIP (ANR-10-LABX-41). YB, JB and BD were supported by MESR (Ministère de l'Enseignement Supérieur et de la Recherche) PhD scholarships.

References

Altshuler D, Pollara VJ, Cowles CR *et al.* (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, **407**, 513–516.

Axelsson E, Smith NGC, Sundström H, Berlin S, Ellegren H (2004) Male-biased mutation rate and divergence in autosomal, z-linked and w-linked introns of chicken and Turkey. *Molecular Biology and Evolution*, **21**, 1538–1547.

Backström N, Fagerberg S, Ellegren H (2008) Genomics of natural bird populations: a gene-based set of reference markers evenly spread across the avian genome. *Molecular Ecology*, **17**, 964–980.

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

Barker FK, Cibois A, Schikler P, Feinstein J, Cracraft J (2004) Phylogeny and diversification of the largest avian radiation. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 11040–11045.

van Bers NEM, van Oers K, Kerstens HHD *et al.* (2010) Genome-wide SNP detection in the great tit *Parus major* using high throughput sequencing. *Molecular Ecology*, **19**(Suppl 1), 89–99.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.

Boitard S, Schlötterer C, Nolte V, Pandey RV, Futschik A (2012) Detecting selective sweeps from pooled next-generation sequencing samples. *Molecular Biology and Evolution*, **29**, 2177–2186.

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

Davey JW, Blaxter ML (2010) RADSeq: next-generation population genetics. *Briefings in Functional Genomics*, **9**, 416–423.

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.* (2012) Special features of RAD Sequencing data: implications for genotyping. *Molecular Ecology*, **22**, 3151–3164.

Derjusheva S, Kurganova A, Habermann F, Gaginskaya E (2004) High chromosome conservation detected by comparative chromosome painting in chicken, pigeon and passerine birds. *Chromosome Research*, **12**, 715–723.

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

Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA (2011) Local *de novo* assembly of RAD paired-end contigs using short sequencing reads. *PLoS One*, **6**, e18561.

Futschik A, Schlötterer C (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, **186**, 207–218.

Gagnaire P-A, Normandeau E, Pavey SA, Bernatchez L (2012) Mapping phenotypic, expression and transmission ratio distortion QTL using RAD markers in the Lake Whitefish (Coregonus clupeaformis). *Molecular Ecology*, **22**, 3036–3048.

Garroway CJ, Radersma R, Sepil I *et al.* (2013) Fine-scale genetic structure in a wild bird population: the role of limited dispersal and environmentally based selection as causal factors. *Evolution*, doi: 10.1111/evo.12121.

Gautier M, Foucaud J, Gharbi K *et al.* (2013) Estimation of population allele frequencies from next-generation sequencing data: pooled versus individual genotyping. *Molecular Ecology*, **22**, 3766–3779.

Griffin DK, Robertson LB, Tempest HG *et al.* (2008) Whole genome comparative studies between chicken and turkey and their implications for avian genome evolution. *BMC Genomics*, **9**, 168.

Hagen IJ, Billing AM, Rønning B *et al.* (2013) The easy road to genome-wide medium density SNP screening in a non-model species: development and application of a 10 K SNP-chip for the house sparrow (Passer domesticus). *Molecular Ecology Resources*, **13**, 429–439.

Harris RS (2007) *Improved pairwise alignment of genomic DNA*. PhD Thesis, Pennsylvania State University, University Park, Pennsylvania.

Hawkins RD, Hon GC, Ren B (2010) Next-generation genomics: an integrative approach. *Nature Reviews Genetics*, **11**, 476–486.

Hess JE, Campbell NR, Close DA, Docker MF, Narum SR (2012) Population genomics of Pacific lamprey: adaptive variation in a highly dispersive species. *Molecular Ecology*, **22**, 2898–2916.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, **11** (Suppl 1), 117–122.

Keller I, Wagner CE, Greuter L *et al.* (2012) Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. *Molecular Ecology*, **22**, 2848–2863.

Kofler R, Orozco-terWengel P, de Maio N *et al.* (2011a) PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One*, **6**, e15925.

Kofler R, Pandey RV, Schlötterer C (2011b) PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, **27**, 3435–3436.

Kolaczkowski B, Kern AD, Holloway AK, Begun DJ (2011) Genomic differentiation between temperate and tropical Australian populations of *Drosophila melanogaster*. *Genetics*, **187**, 245–260.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Miller M, Dunham J, Amores A *et al*. (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–248.

Peng Y, Leung HCM, Yiu SM, Chin FYL (2012) IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, **28**, 1420–1428.

Pérez-Enciso M, Ferretti L (2010) Massive parallel sequencing in animal genetics: wherefroms and wheretos. *Animal Genetics*, **41**, 561–569.

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.

Radwan J, Babik W (2012) The genomics of adaptation. *Proceedings of the Royal Society of London Series B. Biological Sciences*, **279**, 5024–5028.

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology open software suite. *Trends in Genetics*, **16**, 2–3.

Rockman MV (2012) The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution*, **66**, 1–17.

Rubin C-J, Zody MC, Eriksson J *et al*. (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature*, **464**, 587–591.

Schwartz S, Kent W, Smit A *et al*. (2003) Human–mouse alignments with BLASTZ. *Genome Research*, **13**, 103–107.

Takahashi T, Sota T, Hori M (2012) Genetic basis of male colour dimorphism in a Lake Tanganyika cichlid fish. *Molecular Ecology*, **22**, 3049–3060.

van Tassell C, Smith T, Matukumalli L *et al*. (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*, **5**, 247–252.

Travisano M, Shaw RG (2013) Lost in the map. *Evolution*, **67**, 305–314.

Turner TL, Stewart AD, Fields AT, Rice WR, Tarone AM (2011) Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*. *PLoS Genetics*, **7**, e1001336.

Wagner CE, Keller I, Wittwer S *et al*. (2013) Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology*, **22**, 787–798.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**, 57–63.

Wang N, Thomson M, Bodles WJA *et al*. (2012) Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. *Molecular Ecology*, **22**, 3098–3111.

Warren WC, Clayton DF, Ellegren H *et al*. (2010) The genome of a songbird. *Nature*, **464**, 757–762.

Willing EA, Hoffmann M, Klein JD, Weigel D, Dreyer C (2011) Paired-end RAD-seq for de-novo assembly and marker design without available reference. *Bioinformatics*, **27**, 2187–2193.

Zhu Y, Bergland AO, González J, Petrov DA (2012) Empirical validation of pooled whole genome population resequencing in Drosophila melanogaster. *PLoS One*, **7**, e41901.

Y.B., B.M. and C.T. planned the project and wrote the article. B.M. and C.T. obtained funding and organized sample collection. T.C. provided scripts for data analysis. Y.B., J.B., B.D., T.D., J.C., C.T. and B.M. provided samples and performed fieldwork. Y.B., E.L. and O.B. prepared libraries.

Data accessibility

Raw sequence information has been submitted as bam files to the European Nucleotide Archive (ENA) repository at the accession number ERP002555 (available at http://www.ebi.ac.uk/ena/data/view/ERP002555). All scripts used for contig reconstruction are available at https://github.com/tcezard/RADmapper. Files with allelic counts and afasta file with contigs have been deposited on DRYAD (http://dx.doi.org/10.5061/dryad.755b5).