Molecular Phylogenetics and Evolution 62 (2012) 397–406

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences

John E. McCormack [a,*], James M. Maley [a,b], Sarah M. Hird [a,b], Elizabeth P. Derryberry [a,e], Gary R. Graves [c,d], Robb T. Brumfield [a,b]

[a] Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, United States
[b] Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, United States
[c] Department of Vertebrate Zoology, MRC-116, National Museum of Natural History, Smithsonian Institution, PO Box 37012, Washington, DC 20013-7012, United States
[d] Center for Macroecology, Evolution and Climate, University of Copenhagen, DK-2100, Copenhagen Ø, Denmark
[e] Department of Ecology and Evolutionary Biology, Tulane University, New Orleans, LA 70118, United States

ARTICLE INFO

Article history:
Received 1 June 2011
Revised 20 September 2011
Accepted 15 October 2011
Available online 31 October 2011

Keywords:
Population genomics
Phylogenetics
Phylogeography
454 Pyrosequencing
Reduced representation library

ABSTRACT

Next generation sequencing (NGS) technologies are revolutionizing many biological disciplines but have been slow to take root in phylogeography. This is partly due to the difficulty of using NGS to sequence orthologous DNA fragments for many individuals at low cost. We explore cases of recent divergence in four phylogenetically diverse avian systems using a method for quick and cost-effective generation of primary DNA sequence data using pyrosequencing. NGS data were processed using an analytical pipeline that reduces many reads into two called alleles per locus per individual. Using single nucleotide polymorphisms (SNPs) mined from the loci, we detected population differentiation in each of the four bird systems, including: a case of ecological speciation in rails (*Rallus*); a rapid postglacial radiation in the genus *Junco*; recent *in situ* speciation among hummingbirds (*Trochilus*) in Jamaica; and subspecies of white-crowned sparrows (*Zonotrichia leucophrys*) along the Pacific coast. The number of recovered loci aligning closely to chromosomal locations on the zebra finch (*Taeniopygia guttata*) genome was highly correlated to the size of the chromosome, suggesting that loci are randomly distributed throughout the genome. Using eight loci found in *Zonotrichia* and *Junco* lineages, we were also able to generate a species tree of these sparrow sister genera, demonstrating the potential of this method for generating data amenable to coalescent-based analysis. We discuss improvements that should enhance the method's utility for primary data generation.

1. Introduction

The genetic study of recent divergence can be difficult when phenotypic or behavioral differentiation has outpaced random genetic differentiation such that few molecular markers reflect the demographic signal of divergence. Genetic differentiation can also proceed at different speeds throughout the genome (Wu, 2001), with some regions diverging especially quickly and others remaining homogeneous either due to gene flow or insufficient time for accumulation of neutral differences (Via and West, 2008; Nosil et al., 2009). In these cases, sequencing more of the genome is important not only to increase the probability of uncovering rare, divergent genomic regions, but also, from the perspective of phylogeography and population genetics, in order to sample a sufficient number of neutrally-evolving genes to detect a signal of the demographic history.

The field of phylogeography stands to benefit from new technologies for DNA sequencing that exponentially increase the amount of the genome that can be sequenced at one time (Rokas and Abbot, 2009; Holsinger, 2010; Lerner and Fleischer, 2010). Despite some recent progress (Emerson et al., 2010; Gompert et al., 2010), applications of next-generation sequencing (NGS) have lagged in this discipline for several reasons. First, phylogeographic studies typically involve non-model organisms that do not have many genetic resources (e.g., a genome or linkage map). Although advances in DNA sequencing are quickly reducing the gap between model and non-model organisms (Wheat, 2010), it is not yet cost-effective to sequence, let alone analyze, whole genomes for the hundreds to thousands of individuals that many phylogeographic studies entail. A more financially and computationally feasible alternative is to sequence a subset of the genome, i.e., a reduced representation genomic library (Barbazuk et al., 2005). To this end, a major challenge is in the development of cost-effective protocols for preparing samples prior to NGS such that they contain orthologous loci. Because many NGS platforms are only cost-effective when individuals can be combined

* Corresponding author. Present address: Moore Laboratory of Zoology, Occidental College, 1600 Campus Rd., Los Angeles, CA 90041, United States.
E-mail address: mccormack@oxy.edu (J.E. McCormack).

1055-7903/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2011.10.012

into a single run (Glenn, 2011), the best protocols include cost-effective ways to tag individuals, allowing pooling.

Such protocols are beginning to be developed and tested on a wide variety of non-model organisms. For example, restriction-site associated DNA (RAD) sequencing presents a way to reduce the genome via restriction-digest to a manageable number of overlapping genome fragments (Baird et al., 2008). These fragments can then be sequenced via NGS and mined across many individuals for single nucleotide polymorphisms (SNPs) that occur adjacent to common digest sites. This method has proven effective at generating data for marker development (Miller et al., 2007), genome scans (Hohenlohe et al., 2010), and distance-based metrics of population history (Emerson et al., 2010). However, the utility of SNPs in phylogeography is currently limited by the analytical tool kit, which contains many programs that estimate demographic parameters by modeling gene coalescence (Kuhner, 2009; Pinho and Hey, 2010). Presently, robust gene trees can be constructed only from molecular markers with sufficient information content per locus, although equivalent analytical methods may soon emerge that can estimate the same parameters from markers like SNPs (Bryant et al., submitted manuscript). For the time being, the ideal protocol from the perspective of phylogeography would be one that generates sequence data from many hundreds of loci, each containing high information content (i.e., hundreds of bases), for many individuals, allowing loci to be mined for SNPs and used in coalescent-based analyses that require gene trees.

Looking to the future, there are many promising routes for using NGS to generate DNA sequence data from unlinked loci. One promising method, called sequence capture or targeted resequencing (Garber, 2008; Gnirke et al., 2009), involves hybridizing genomic DNA to RNA probes and allows for efficient data collection without hundreds of separate amplifications (Mamanova et al., 2009). Sequence capture may ultimately provide the gold standard for DNA sequence data collection in the near future, but, for the moment, these protocols are only beginning to be applied to phylogeography.

Genome reduction through restriction digest (Altshuler et al., 2000) remains a cost-effective alternative, especially when DNA from multiple individuals can be pooled together. There are several variations on this basic idea in the literature (Whitelaw et al., 2003; Barbazuk et al., 2005; Van Tassell et al., 2008; Wiedmann et al., 2008; Gompert et al., 2010; Hyten et al., 2010; Williams et al., 2010), most resulting in the generation of SNPs, not full loci, and for a handful of individuals or population-level pools of individuals. Here, we demonstrate the utility for phylogeography of a wet lab method for the rapid and cost-effective generation of reduced representation libraries for NGS (in this case, for the 454 platform, although the basic idea is adaptable to any platform). We employ an analytical pipeline (Hird et al., 2011) to process raw NGS data into data suitable for phylogeography and population genomics. Our method for generating loci resembles that of Williams et al. (2010), but our focus is on recovery of whole loci as well as SNPs to maximize applicability to a wide array of analyses. We apply our method to recent divergences in four bird systems to assess whether this approach can uncover difficult-to-detect phylogeographic structure. Our choice of systems also allows us to pool results from multiple bird species to explore whether loci can be obtained from increasingly divergent groups, which will inform its usefulness for phylogenetics.

2. Methods

2.1. King Rails and Clapper Rails in southwestern Louisiana

These two species (King Rail, *Rallus elegans*, and Clapper Rail, *Rallus longirostris*) represent an interesting case of divergence coincident with a salinity gradient along the Atlantic and Gulf coasts of North America (Meanley, 1992; Eddleman and Conway, 1998). The two species are 0.8% divergent in mitochondrial DNA (mtDNA) and show evidence for hybridization across a steep salinity cline in Louisiana (J. Maley, unpublished data). For our study, we chose 10 individuals from a King Rail population and 10 individuals from a Clapper Rail population in Louisiana that were deemed to be "pure" based on geographical location, phenotype, and a lack of mtDNA mixing that characterizes hybrid populations (J. Maley, unpublished data). DNA was extracted from pectoral muscle from individuals collected in the field as part of a collections-based study of their hybrid zone.

2.2. Juncos in North and Middle America

Juncos on the North American continent consist of three recognized species, the Dark-eyed Junco (*Junco hyemalis*) of the United States, which itself consists of many populations with different plumage that originated in a rapid postglacial radiation (Milá et al., 2007); the Yellow-eyed Junco (*Junco phaeonotus*) of the Middle American highlands; and the Volcano Junco (*Junco vulcani*), a highland endemic of Costa Rica and Panama. For this study, we included individuals from a number of populations representing different Dark-eyed Junco plumage types ("Oregon Junco," *J. h. oreganus*, $N = 4$; "Pink-sided Junco," *J. h. mearnsi*, $N = 4$; "Red-backed Junco," *J. h. dorsalis*, $N = 3$), a population of Yellow-eyed Juncos from southeastern Arizona ($N = 3$), a suspected hybrid individual from New Mexico ($N = 1$), individuals from a geographically isolated population of Yellow-eyed Juncos from the southern tip of the Baja Peninsula ("Baja Junco," *J. p. bairdi*, $N = 3$), and two Volcano Juncos from Costa Rica. The rationale behind this sampling scheme was to maximize geographic and phenotypic differentiation from largely "pure" populations for the purpose of detecting genetic variation. The Yellow-eyed Junco on the Baja Peninsula and Volcano Junco are divergent from other juncos in mtDNA (Milá et al., 2007). Other Yellow-eyed Juncos and Dark-eyed Juncos on mainland North America, despite phenotypic differences, show negligible mtDNA divergence, but there are population-level frequency differences in amplified fragment length polymorphisms (AFLPs)(-Milá et al., 2007). DNA was extracted from pectoral muscle from juncos collected in the field.

2.3. Streamertail hummingbirds in Jamaica

The red-billed Streamertail (*Trochilus polytmus*) and black-billed Streamertail (*Trochilus scitulus*) represent a fascinating case of *in situ* speciation in one of the smallest oceanic island settings known (Gill et al., 1973). The major phenotypic difference lies in bill color, with the black-billed form confined to the extreme eastern tip of Jamaica, whereas the red-billed form occurs widely over the remainder of the island. All studies to date have failed to find spatially-structured genetic variation (e.g., Lance et al., 2009), suggesting that the divergence is extremely recent. We chose 10 individuals from the most geographically disparate populations of each of the two species. DNA was extracted from pectoral muscle from hummingbirds collected in the field.

2.4. White-crowned sparrow subspecies along the western coast of the United States

White-crowned sparrows (*Zonotrichia leucophrys*) consist of five subspecies whose breeding populations are widely distributed in North America (Chilton et al., 1995). Two of these subspecies meet and intergrade along the Pacific coast of northern California near Cape Mendocino (Grinnell, 1928; Banks, 1964). The subspecies to the south (*Zonotrichia leucophrys nuttalli*) is comprised primarily

of non-migratory birds whereas the subspecies to the north (*Zonotrichia leucophrys pugetensis*) tends to be migratory (Blanchard, 1941). There are clinal shifts in bill length, toe length and tarsus length between the two subspecies (Banks, 1964), which are thought to have diverged in Pleistocene glacial refugia, although phenotypic intermediates abound and no allozymic differences have been detected (Corbin, 1981; Corbin and Wilkie, 1988). We chose 10 individuals from each of the two subspecies from populations located away from the transition zone. DNA was extracted from a combination of blood and tissue samples. A small (20 μL) blood sample was drawn by brachial venipuncture and transferred onto EDTA-saturated filter paper, where it was allowed to dry and then stored in airtight containers on DriRite.

2.5. Genome reduction, sample preparation, and 454 sequencing

We prepared samples from 20 individuals from each of the four bird systems, with each set of 20 run on its own quarter plate of a 454 run. A basic workflow for the wet lab protocol is depicted in Fig. 1A. Similar protocols can be found in Gompert et al. (2010) and Williams et al. (2010), although a major difference in our goal was to infer genotypes for individuals directly from 454 data (as opposed to using population pools) and use these data as the primary data with no further genotyping. For each of the 20 samples in each pool, we quantified extracted DNA with a NanoDrop and standardized initial DNA concentrations at 100 ng/μL. Then, in a single step, 250 ng DNA was digested and adaptors (Vos et al., 1995) were ligated on the resulting sticky ends for 2 h at 37 °C in an 11 μL volume reaction containing 3.15 μL water, 2.5 μL DNA template, 1.1 μL T4 ligase buffer, 1.1 μL of 0.5 mM NaCl solution, 5 U EcoR1, 5 U MseI, 0.55 μl of 1 μg/μL BSA, 5 U T4 ligase (New England Biolabs), and 1.0 μL of 10 μM adaptor for both MseI and EcoR1. Each adaptor was made beforehand by heating equal volumes of two complementary pieces of DNA (Table 1) to 94 °C and allowing them to cool to room temperature. We then conducted a round of PCR in a 20 μL reaction containing 10 μL of a 10-fold dilution of the digest-ligation, 5.4 μL water, 2.0 μL of 25 mM $MgCl_2$, 2.0 μL of 10X buffer, 0.4 μL of 10 mM dNTPs, 0.06 μL of 100 μM concentration adaptor-specific primer (Table 1), and 0.08 μL of 5 U/μL Phusion high-fidelity Taq (Finnzymes, Woburn, MA). The PCR protocol was 2 min at 72 °C, followed by 15 cycles of 98 °C for 30 s, 56 °C for 30 s, and 72 °C for 2 min, followed by 72 °C for 10 min. For each individual, resulting PCR products were visualized on an agarose gel. A section of the resulting smear of DNA fragments (400–550 bp) was excised. During excision, DNA from each individual was separated by an empty well and care

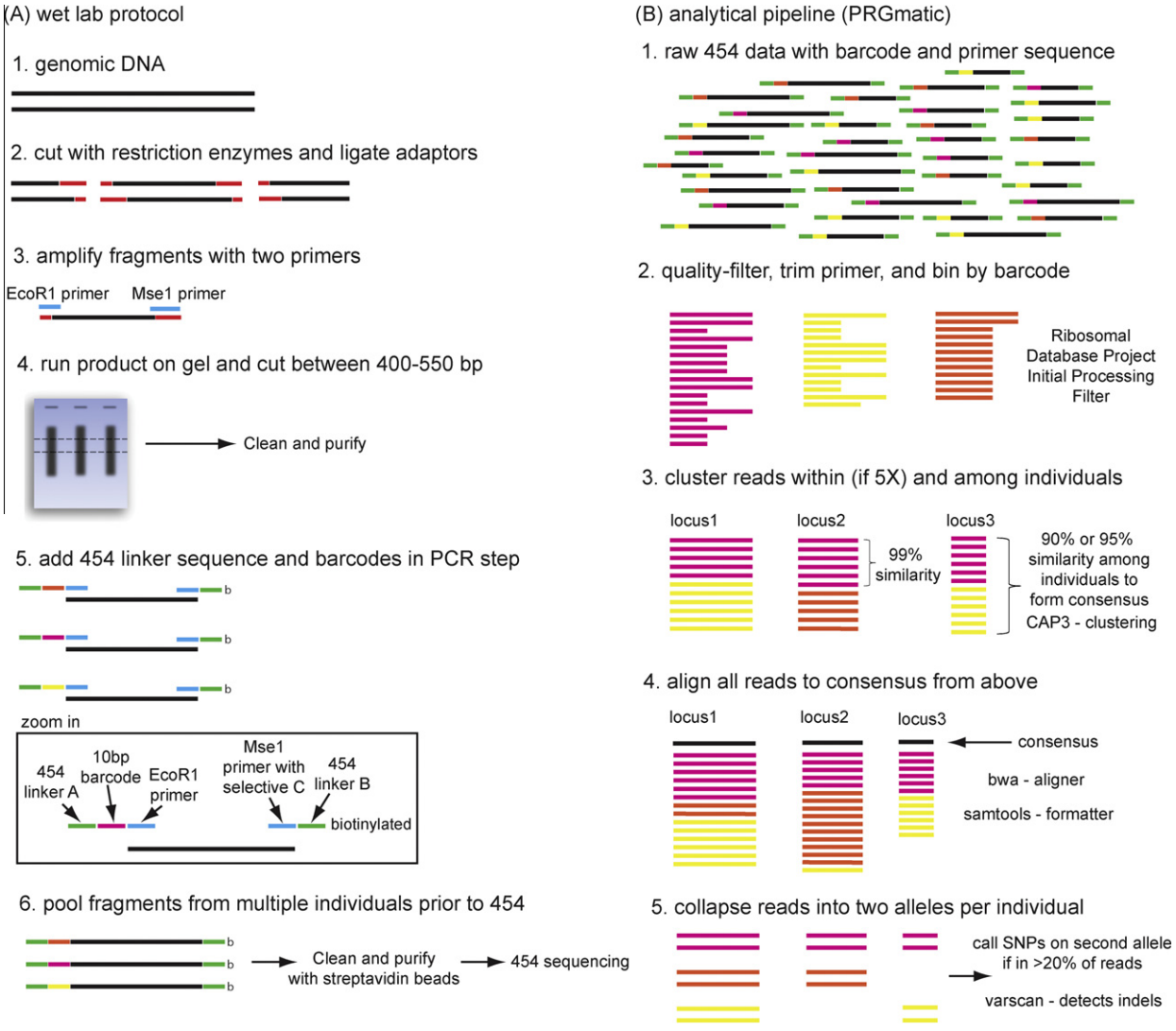

Fig. 1. Schematic of (A) wet lab method for generating loci for many individuals with NGS, and (B) analytical pipeline for processing raw NGS data into two called alleles per individual.

Table 1
Primers and adaptor sequences used in this study.

Name	Sequence (5′–3′)
EcoRI adaptor[a]	CTCGTAGACTGCGTACC
EcoRI adaptor[a]	AATTGGTACGCAGTCTCA
MseI adaptor[a]	GACGATGAGTCCTGAG
MseI adaptor[a]	TACTCAGGACTCAT
EcoRI primer[a]	GACTGCGTACCAATTC
MseI primer[a]	GATGAGTCCTGAGTAA
EcoRI fusion primer[b]	CGTATCGCCTCCCTCGCGCCATCAGXXXXXXXXXXGACTGCGTACCAATTC
MseI fusion primer[c]	b-CTATGCGCCTTGCCAGCCCGCTCAGGATGAGTCCTGAGTAA**C**

[a] From Vos et al. (1995).
[b] Different barcode sequence in X's for each individual in a sample pool.
[c] Biotinylated on the 5′ end. Selective base in bold.

was taken to avoid cross-contamination. Samples were column purified using a QIAquick gel extraction kit (Valencia, CA), eluted with 50 μL volume, and then subjected to another round of PCR, this time with longer "fusion primers" (Table 1) bearing the complementary adaptor priming sites in addition to overhanging DNA sequence containing the necessary binding sites for emulsion PCR and individual-identifying barcodes (20 different primers, each with a different barcode sequence). This PCR was 10 μL in volume and contained 2.5 μL of eluted PCR product, 4.56 μL water, 1.0 μL 25 mM Mg Cl_2, 1.0 μL 10X buffer, 0.2 μL 10 mM dNTPs, 0.06 μL of MseI reverse fusion primer (Table 1) at 100 μM, and 0.08 μL of 5 U/μL Phusion Taq. The MseI reverse fusion primer was biotinylated both to aid in removing non-targeted short fragments and to remove any EcoRI–EcoRI fragments remaining in the pool. To each reaction, we separately added 0.6 μL of EcoRI forward fusion primer with indexes (Table 1) at 10 μM concentration. For this PCR, we used a touchdown profile beginning with 94 °C for 2 min, then 10 cycles of 94 °C for 30 s, 65 °C for 30 s (reducing temperature by 0.7 °C in each cycle), 72 °C for 60 s, then 10 cycles of 94 °C for 30 s, 56 °C for 30 s, and 72 °C for 60 s, followed by 10 min at 72 °C.

For each of the 20 tagged samples in a set, PCR products were further purified using streptavidin beads (Dyanbeads, Invitrogen) and final DNA concentration was quantified using a combination of Picogreen (Molecular Probes, Eugene, OR) fluorescent dye assay, gel electrophoresis, and a LabChip assay on an Agilent (Palo Alto, CA) BioAnalyzer. For each bird system, 500 ng DNA from each of 20 barcoded samples was pooled and 3–5 μg was run on a separate quarter plate of a 454 Life Sciences Genome Sequencer Titanium machine at Research and Testing Laboratories (Lubbock, TX).

2.6. 454 Data processing and PRGmatic pipeline settings

Detailed information on the bioinformatics pipeline, PRGmatic, is contained in Hird et al. (2011), as well as a validation of the procedure using simulated data. The basic workflow of PRGmatic (Fig. 1B) is that raw NGS data (.fna and .qual files) are uploaded to the Ribosomal Database Project Initial Processing (Cole et al., 2009) for quality filtering (>20 phred and >100 bp fragments) and binning by barcode. PRGmatic takes the resulting .fasta and .qual files as input and begins by clustering reads together within individuals at high similarity (99%) to identify putative alleles. High coverage (default = 5) putative alleles are clustered across individuals into provisional loci, which are then concatenated into a provisional reference genome (PRG). All original reads are then aligned to the PRG using BWA (Li and Durbin, 2009) and SAMTOOLS (Li et al., 2009) with individuals at each locus being retained if they reach a certain depth of coverage (default = 6). PRGmatic then uses VarScan (Koboldt et al., 2009) and custom Perl scripts to call two alleles per individual using thresholds for the minimum number of reads to call a SNP at a particular base position (default = 3) and the minimum occurrence of a SNP variant needed to call an individual a heterozygote (default = 20%).

Prior to producing our final data sets, we decided on pipeline parameters by testing all parameters individually using data from the rails, to see what effect each parameter had on the number of loci detected. We found that the parameter with the single greatest influence on the number of loci comprising our final data set was the threshold read number required for individuals to have alleles called (Table S1). Lowering this value from the default of 6 reads to 4 reads resulted in many more loci that had $\geqslant 7$ and $\geqslant 15$ individuals (our thresholds for final data sets, see below) represented and did not jeopardize data quality (see SNP validation below). Individually, other parameters did not have a large effect on the total number of loci detected. However, when combined with a lowered threshold for individual reads, a lowered threshold for alleles to make the PRG (reads = 3 compared to default 5) resulted in many more loci detected. We therefore employed the following parameters for PRGmatic: threshold for allele reads to make the PRG = 3; loci identity for clustering = 90%; threshold read number for individuals to have alleles called = 4; threshold for a calling a SNP at a base position = 2; minimum percent of reads needed to call a heterozygote = 20%. For the junco and sparrow systems, where overall coverage was lower, we further relaxed the threshold for individual reads from 4 to 3. It should be noted that resulting loci do not represent the final data sets analyzed, but rather the coverage depth required for the pipeline. In other words, the thresholds implemented by PRGmatic and our own threshold for individual representation mean that the very lowest coverage alleles were not included in the final data sets.

In addition to analyses of each of the four bird data sets individually, we also combined and analyzed data from the junco and sparrow systems, as well as data from all four systems, together to evaluate if orthologous markers could be recovered from more phylogenetically distant taxa (i.e., above the species level).

2.7. Mining aligned loci for SNPs

After running PRGmatic, we had many aligned loci for each bird system, with two called alleles for each individual. We mined these loci for SNPs by importing them into Geneious (Biomatters, Auckland, NZ) and sorting by number of individuals, focusing our efforts on loci with the most individuals represented (i.e., the least missing data). Since loci with high coverage are more likely to be paralogous (Emerson et al., 2010), we assessed by eye whether sequences were likely to be paralogs by the amount of individual heterozygosity. Variation at putative paralogous loci was usually apparent because every individual appeared in a heterozygous state for divergent alleles, and there were often numerous SNPs (>6) per locus. These putative paralogous loci were removed. In the rail data, we also assessed deviations from Hardy–Weinberg

equilibrium. Results suggested that the qualitative screening was successful, as only one locus with high heterozygosity (observed >0.1 above expected) had made the final data set.

In the remaining loci, SNP positions were pasted into an Excel spreadsheet, retaining all SNPs that occurred in three or more sequences (threshold set in Geneious). The importance of tight linkage among SNPs was assessed with Structure (see below). For the rails, we validated SNP calls in 20 individuals at four loci using Sanger sequencing. The four loci were chosen because they contained informative SNPs and were thus likely to be useful for phylogeographic inference.

2.8. Divergence analysis

We ran STRUCTURE 2.2 (Pritchard et al., 2000) on coded SNP files, using an admixture model and correlated allele frequencies. We assessed values of population differentiation (K) between 1 and 5 since two populations were strongly suspected for all systems except the juncos, in which six populations were possible; in this case, we tested K from 1 to 10. We initially assessed the most likely K by the highest Ln P(D) score. In most cases we observed a clear peak in Ln P(D). If multiple K's had similar Ln P(D), we assessed whether assignment of the additional genetic cluster was informative, or whether it was assigned equally to our putative populations (Pritchard and Wen, 2004). We did not attempt to find the "true K" (Evanno et al., 2005) because our goal was not to discover the most appropriate number of clusters per se, but whether we could detect genetic structure that was suspected *a priori*.

Having tightly linked SNPs (i.e., SNPs from the same locus) is not expected to pose a problem for Structure provided many independent genomic regions are included (Pritchard and Wen, 2004). However, to assess whether tight linkage of some SNPs on the same locus were influencing our analyses, we conducted additional runs including map distances in the Structure files. We assigned map distances according to SNP position within a locus, introducing an arbitrary large number (10,000 bp) between loci to approximate widely separated genomic regions, an assumption that was validated by the BLAST results (see below). All runs were conducted for 1,000,000 iterations with a burn-in of 100,000, which yielded convergence as assessed by alpha scores. If alpha scores indicated problems with convergence or if multiple iterations at the same K gave different results, iterations were increased to 10,000,000 with a burn-in of 1,000,000. For visualization in figures, we chose the iteration with the highest posterior probability.

For each recovered locus that contained a SNP, we conducted a BLAST search against the zebra finch genome and recorded the chromosome number of the hit with the top "max score" (provided that value was >80). Birds show a high degree of shared synteny and chromosomal stability compared to most vertebrate groups (Ellegren, 2010), with a few notable exceptions, so locus position on the zebra finch is expected to roughly correspond to chromosomal location in species in our study. We conducted all searches against the zebra finch genome, as opposed to the chicken genome, which is more appropriate to our study species because all are members of Neoaves.

2.9. Multi-species alignments and phylogenetic analysis

To explore at which phylogenetic level loci could be recovered, we ran two PRGmatic alignments, one including reads from sparrows and juncos and one with reads from all four systems together. Recovered loci that included sparrows and juncos (which are in sister genera; Zink and Blackwell, 1996) were analyzed using the program *BEAST (Heled and Drummond, 2010), part of the BEAST v1.6 package (Drummond and Rambaut, 2007), which infers species trees from collections of gene trees. To balance the need to include as many loci as possible with resolution of as many terminal taxa as possible (*BEAST requires at least one individual from each terminal taxon), we considered seven terminal taxa ("species" in *BEAST terminology): white-crowned sparrows (both subspecies combined), "Oregon Junco," "Pink-sided Junco," "Red-backed Junco," and Yellow-eyed Junco, "Baja Junco," and Volcano Junco. We ran the analysis for 100,000,000 generations and assessed convergence by eye in Tracer using ESS values and the trace plot of likelihood. Of the 10,000 posterior trees, 1000 were discarded as burn-in, which was conservative as judged by eye.

3. Results

3.1. Descriptive NGS results

Our full plate of 454 sequencing produced just over 1,000,000 reads. Total reads for each quarter-plate (i.e., each bird system) were similar, ranging from 240,000 to 267,000 reads (Table 2). Quality-filtering and removal of reads <100 bp and reads without indexes left between 51% and 68% of the initial reads that averaged in length between 268 and 300 bp (after removing primer and index sequence). After running each system through the analytical pipeline, the total number of loci per system ranged from 1064 to 2281, although only a fraction of these loci were found in $\geqslant 7$ and $\geqslant 15$ individuals (Table 2). Loci length showed a high peak around 300 bp and a smaller peak around 100 bp (Fig. 2). In general, species that had more short loci relative to long loci had fewer total loci and fewer loci with high individual representation (e.g., sparrows; Fig. 2A). Loci with higher individual representation were longer (regression: $F_{1,6360} = 409.85$, $R^2 = 0.06$, $P < 0.001$). Total coverage per locus per individual was low when all loci were considered (Table 3), but increased substantially when including only loci with a certain threshold of individual representation (i.e., those that would comprise our analyzed data sets).

3.2. SNP validation

Sanger sequencing of 20 rails at four loci suggested high accuracy of the 454 data (Table 4). Error rates averaged 0.1%, lower than some others NGS studies using 454 sequencing (Niu et al., 2010), which might be attributed to the fact that called alleles went through a multi-level process of screening and error filtering in the analytical pipeline. Of the few errors, most involved calling individuals homozygous when they were actually heterozygotes. Others were likely PCR error. Error rates did not increase when the read threshold for individual representation was relaxed to four reads from the default setting of six reads (Table 4).

3.3. Clustering analysis

Structure analyses detected population genetic clustering in each of the four bird systems. In all cases, there was no difference if map distances were used. BLAST results of loci that contained SNPs used in our Structure analyses suggest broad genomic coverage (Fig. 3) in all systems. Total number of hits to a chromosome was tightly correlated with chromosome size ($R^2 = 0.85$, $P < 0.001$), suggesting loci were randomly distributed throughout the genome.

3.3.1. Rails

104 SNPs from loci with 15+ individuals recovered clear population genetic structure between King and Clapper Rails in Louisiana (Fig. 4A). The most-likely number of clusters was $K = 2$ (Table S2). Average assignment for the 10 King Rails and 10 Clapper

Table 2
Descriptive results from 454 sequencing run and analytical pipeline.

	Rails	Hummingbirds	Juncos	Sparrows
Total # reads	267,814	240,896	245,393	253,500
Total # reads >100 bp (% of total)	211,437 (79%)	180,545 (75%)	172,786 (70%)	161,887 (64%)
Total # reads >100 bp and QC (% of total)	187,827 (70%)	157,769 (65%)	147,761 (60%)	135,175 (53%)
Avg. read length after QC (st. dev.)	300 (71.6)	290 (78.8)	280 (71.6)	268 (82.9)
Total QC reads with barcode (% of total)	182,447 (68%)	150,882 (63%)	135,878 (55%)	128,262 (51%)
Avg. QC reads per barcode (range)	9122 (3959–16,723)	7544 (3319–15,011)	6794 (2375–10,283)	6413 (3974–13,643)
Total # aligned loci – four reads to include individual	2281	1891	1780	1064
Loci found in ⩾7 individuals	376	160	100	50
Loci found in ⩾15 individuals	67	19	8	19
Total # aligned loci – three reads to include individual	–	–	2483	1576
Loci found in ⩾7 individuals	–	–	189	76
Loci found in ⩾15 individuals	–	–	15	27
Average locus length in bp for all loci	298	287	275	238
Average locus length in bp for loci with ⩾7 individuals	320	322	308	284

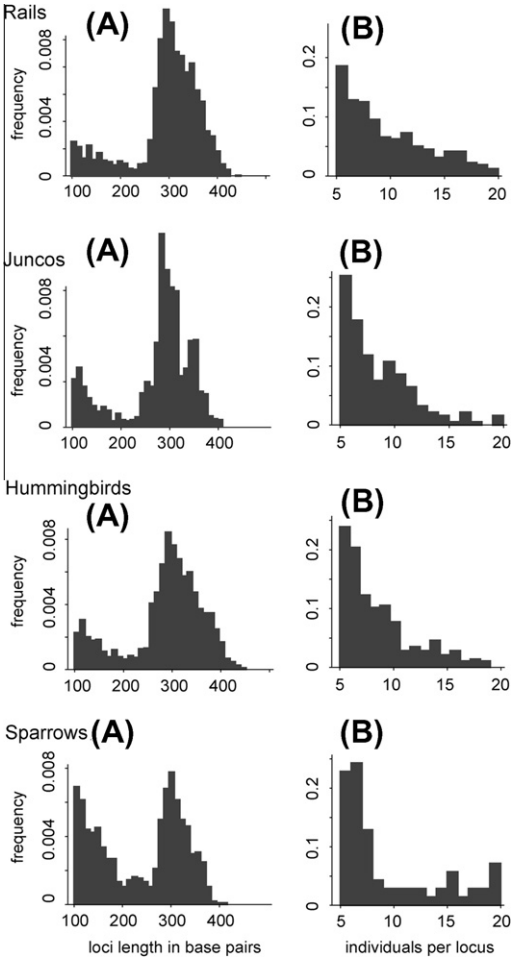

Fig. 2. (A) Distribution of loci lengths after trimming index and primer sequence, and (B) distribution of number of individuals represented at each locus (loci with only 1–4 individuals removed to allow for better visualization of bins with higher individual representation).

Rails to different genetic clusters was 93% and 91%. Average missing data for all individuals was 18%, which was largely driven by three individuals. Among the 17 other individuals, average missing data was only 10%.

3.3.2. Juncos

With the original parameter settings, only 9 SNPs were recovered from loci with 15+ individuals. Lowering the threshold of individual representation to 7+ individuals resulted in 46 SNPs with 46% missing data in the SNP matrix. Using these data, the two geographically isolated junco populations (Volcano Junco and Yellow-eyed Junco on the Baja Peninsula) were assigned with high probability to different genetic clusters from each other and from other juncos in Mexico and North America (Fig. 4B). Structure identified $K = 4$ genetic clusters as being most likely (Table S2), but the additional two clusters were divided among the remaining juncos with no clear geographic pattern (Fig. 4B).

Table 3
Coverage. Average reads per individual per locus (standard deviation).

System	Loci with 15+ individuals	Loci with 7+ individuals	All loci
Rails	9.1 (7.3)	5.2 (5.3)	1.6 (3.0)
Hummingbirds	8.7 (7.0)	4.7 (4.6)	1.1 (2.3)
Juncos	30.7 (50.9)	8.2 (20.9)	1.1 (5.3)
Sparrows	46.1 (92.7)	21.2 (62.5)	1.2 (11.6)

To attempt to uncover further genetic structure within North American juncos, we relaxed the analytical settings for individual representation to three reads, which resulted in more loci with more individuals (Table 2). In an analysis of 43 SNPs from loci with 15+ individuals (i.e., a similar number of SNPs as the prior analysis, but with only 19% missing data) Ln P(D) peaked at $K = 5$ (Table S2). When visualized at $K = 5$, both the Volcano and Baja Juncos were distinct, but again there was no clear differentiation among other juncos (Fig. 4B). Finally, an analysis of 247 SNPs from loci with 7+ individuals showed increasing Ln P(D) scores with a plateau after $K = 3$ (Table S2). Visualization at $K = 3$ again revealed Baja and Volcano juncos as distinct, but visualization at $K = 4$ and all higher Ks showed remaining genetic clusters assigned with no apparent geographic structure. No further resolution was provided when Volcano and Baja juncos were removed and the analysis re-run.

3.3.3. Hummingbirds

An analysis of 37 SNPs from loci with 15+ individuals (20% missing data) indicated $K = 1$ as most likely (Table S2). An analysis of 209 SNPs from loci with 7+ individuals (44% missing data) indicated $K = 3$ and $K = 4$ as both highly probable (Table S2). Visualization at $K = 3$ revealed high assignment of a few individuals to the same cluster with no geographic pattern (Fig. 4C); however, visualization at $K = 4$ showed clear differentiation between the two species (Fig. 4C). To assess whether artifacts arising from missing data might have affected our results, we also analyzed a data set of 111 SNPs from loci with 10+ individuals that had 33% missing data. Here, $K = 2$ was most likely (Table S2) and differentiation between species was even more clear (assignment to different clusters for black-billed and red-billed was 67% and 78%; Fig. 4C).

Table 4
Sanger sequencing validation for allele calls in 20 rails. Ind represents number of individuals with allele calls for a particular locus, # bp called represents a count of the total number of bp called for a locus across individuals (both allele calls included).

Locus	Read threshold	Ind	# bp called	# Called correctly	% Called correctly	# Informative SNPs miscalled	Comments
R139	4	9	3800	3800	100.0	0	
	6	2	718	718	100.0	0	
R472	4	9	2936	2932	99.9	1	1 Het called as homozygous in 1 individual
	6	6	1982	1982	100.0	0	
R1166	4	9	2736	2730	99.8	0	PCR error and not enough reads for a het
	6	4	1218	1216	99.8	0	PCR error
R1766	4	8	4096	4086	99.8	0	Uninformative hets miscalled
	6	5	2560	2550	99.6	0	Uninformative hets miscalled

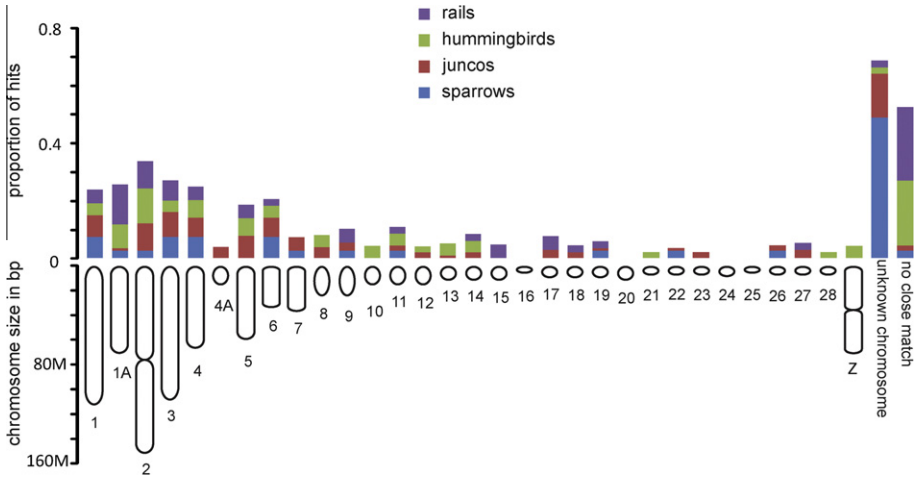

Fig. 3. Chromosomal location in the zebra finch of all loci that contained SNPs used in the Structure analysis for each species. Shown are proportional frequencies within species. There is a significant correlation between chromosome size and number of hits ($R^2 = 0.85$, $P < 0.001$, excluding the Z chromosome), suggesting a random genomic distribution of the loci. The high proportion of hits on unknown chromosomes for sparrows and juncos is likely due to repetitive DNA that could not be placed to a specific chromosome in the zebra finch. The high proportion of loci with no close match in rails and hummingbirds likely reflects phylogenetic distance from zebra finch. Data on chromosome size were taken from Backström et al. (2010).

3.3.4. Sparrows

With our original settings, only 10 SNPs were recovered from loci with 15+ individuals. An analysis of 31 SNPs from loci with 7+ individuals indicated no population structure ($K = 1$ was most likely). After relaxing the analytical pipeline settings to allow individuals with three reads to enter the data set, an analysis of 96 SNPs from loci with 7+ individuals (missing data = 37%) found evidence for genetic structure between the *pugetensis* and *nuttalli* subspecies (at $K = 2$, a population average of 83% and 86% were assigned to different clusters, respectively). Posterior probabilities identified $K = 3$ and $K = 4$ as most probable (Table S2; Fig. 4D).

3.4. Multi-species analysis

In an alignment of reads from the sparrows and juncos, we uncovered 30 loci that included both sparrow and junco individuals. Of these loci, eight included at least one individual from each of the seven taxa we considered as "species" for the analysis (see Section 2). The *BEAST analysis converged (all ESS values >200 and ESS of the likelihood = 709) and produced a species tree that largely conformed to suspected mtDNA relationships (Fig. 5), with perfect support for deep nodes including a sister relationship between the white-crowned sparrows and all juncos, and a sister relationship between the Volcano Junco and all other juncos. In the alignment of reads from all four systems, only two loci were found in three or more bird systems.

4. Discussion

4.1. Generating primary data sets with next-generation sequencing

Our results show that next-generation sequencing of reduced representation genomic libraries can produce primary data sets that reveal fine-scale, individual-based population genetic structure without the need for further genotyping or sequencing. A major attraction of NGS technology from the perspective of phylogeography is the possibility of generating primary data in addition to developing markers (SNPs or anonymous loci) that later require genotyping or sequencing by conventional methods. Only a few prior studies (e.g., Emerson, 2010) have demonstrated the potential to use NGS to generate primary data, and none, to our knowledge, using full loci in addition to SNPs. We would like to point out that it is not our recommendation to use ¼ plate of 454 sequencing for an entire project because clearly our results suffered from low coverage, which reduced the total number of loci that had many individuals represented. However, despite the challenge of low coverage and missing data, we were able to detect population structure in all four systems, which is promising for future phylogeographic work building on this and similar methods.

4.2. NGS reveals population structure of recent bird divergences

Even though the implementation of our method could be improved in terms of read depth (we provide some suggestions

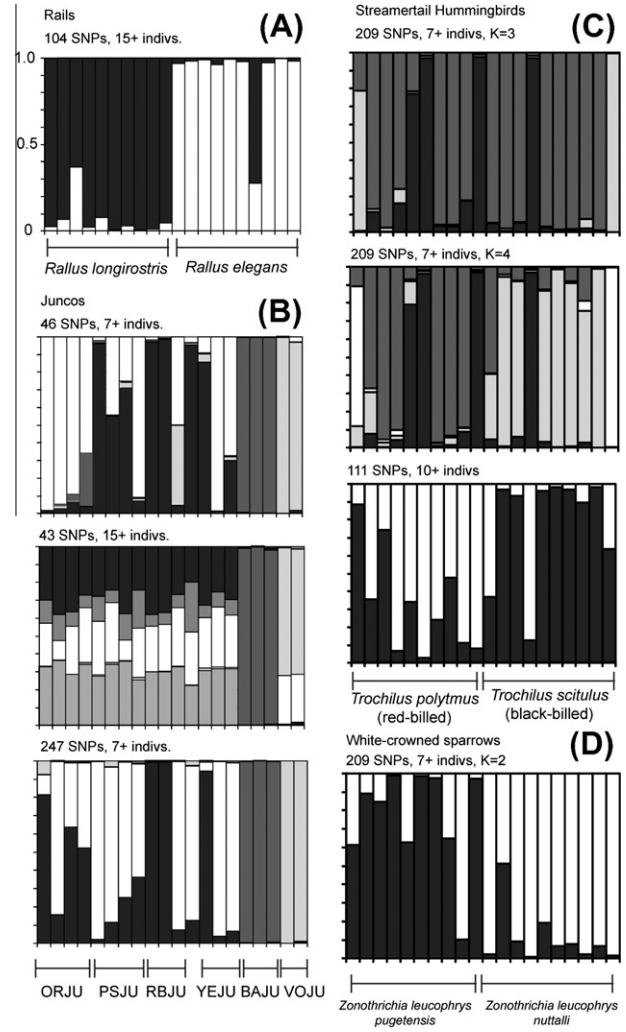

Fig. 4. Structure results for (A) rails, (B) juncos, (C) hummingbirds and (D) sparrows.

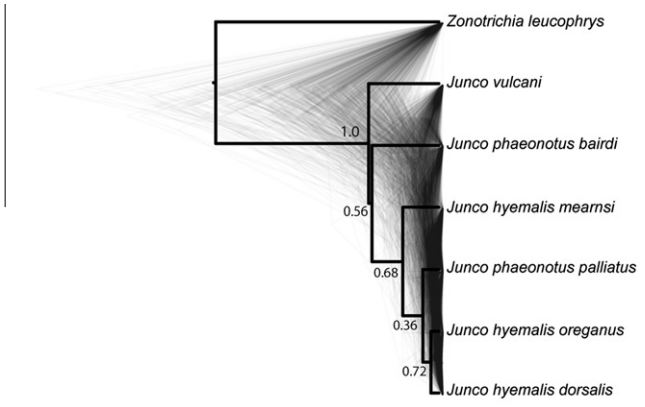

Fig. 5. Species tree of sparrows and juncos based on eight genes that included representatives of all terminal taxa.

below), we were still able to detect population structure for each of the four recent bird divergences we investigated, including two very difficult cases that had previously defied efforts to detect genetic structure with molecular markers (white-crowned sparrows and streamertail hummingbirds). The rails had the highest quality and quantity of data by most measures (e.g., most reads, most loci with the most individuals, longest loci; Table 2 and Fig. 2), which translated to data matrices with the least missing data. Even the most conservative analytical conditions resulted in >100 loci and SNPs that provided sufficient information to detect population-genetic structure between the two species (Fig. 2A). We are not certain whether the lack of perfect assignment of some individuals is reflective of missing data or a genuine biological signal of gene flow or incomplete gene sorting. Given that *R. longirostris* and *R. elegans* are roughly 0.8% divergent in mtDNA (i.e., they diverged about 400,000 years ago; J. Maley, submitted manuscript), our results suggest that this method should be ideal for phylogeographic studies of lineages that diverged during the Pleistocene.

The other extreme in terms of data quality and quantity was found in the white-crowned sparrow system, in which the length of detected loci had several high peaks (Fig. 2), which could be indicative of restriction digest sites within repetitive DNA (Hyten et al., 2010). Also, coverage was uneven (Table 3), suggesting that repetitive DNA might have used up many of the reads, exhausting coverage for the remaining loci. Nevertheless, allowing a relaxed threshold for the number of reads required for an individual to be included in the final data set, we were able to detect population structure between *Z. l. nuttalli* and *Z. l. pugetensis* (Fig. 4). This suggests that missing data may not pose a large problem for studies seeking to generate primary data with NGS.

In the juncos, the Baja and Volcano juncos were consistently detected as independent populations, which has been shown previously (Milá et al., 2007). The failure to detect a clear pattern of structure in the remaining juncos in mainland North America is perhaps odd given that previous results found evidence for population differentiation using a similar number of markers (Milá et al., 2007). However, the differences observed in Milá et al. (2007) emerged from population-level pools of assignment probability (i.e., population averages including many individuals); within a pool, individuals were heterogeneous and often mixed in their assignment, similar to our results. It could be that we simply did not have enough individuals per population for broad, population-level differences to emerge.

We address two potential problems with this method that could lead to spurious estimation of population structure; however, neither seems to be a plausible explanation for our results. First, missing data can skew estimates of population assignment in Structure if the pattern of missing data is systematically biased toward one population. In our case, there is no reason to suspect that missing data resulting from coverage problems would be biased more toward one population than another. Null alleles resulting from a mutation in the restriction site could cause systematic bias in missing data toward one population. However, as this form of missing data is still reflective of genetic differences, it is perhaps less troubling. Second, paralogous loci could cause systematic bias if one paralog was assigned to one population and a different paralog was assigned to another population. However, we cannot think of a scenario where such a bias would occur in practice. We screened for paralogs prior to our analysis using a qualitative approach based on coverage, number of expected SNPs per allele, number of divergent bases occurring at each SNP location, and heterozygosity. More quantitative methods for paralog screening are now available (Hohenlohe et al., 2011).

4.3. Improvements to the wet lab method for greater read depth and more loci

While the number of loci and SNPs in our study was sufficient for detecting a signal of population genetic structure, their overall numbers in the final data sets (about 100–200) were not particularly high given the capabilities of NGS technology (e.g., Emerson

et al., 2010; Hohenlohe et al., 2011). This was partially due to the conservative, multi-level screening carried out by the analytical pipeline in addition to the conditions we imposed for individual representation for loci to reach the final data set. For example, Sanger sequencing validation demonstrated not only that all SNPs called by the pipeline for four loci were real, but also that analytical calls of homozygotes versus heterozygotes were also highly accurate.

Improvements to the sample preparation method could likely boost the number of loci significantly. The number of reads for an individual was the major limiting factor for our final data sets (Table S1). Obviously, simply increasing the size of the run can increase total read number. In addition, a more even distribution of reads among individuals and fragments would also boost individual coverage. Using fresh tissue and high-quality DNA would likely produce longer high-quality reads. Although the evidence is circumstantial, our best results were from the rail system where fresh tissues were taken immediately post-collection and extracted. On the other hand, there was no difference in the amount of data recovered from sparrow individuals where DNA was extracted from tissue as opposed to dried blood (t-test: $P = 0.93$). This suggests that NGS studies can be conducted using DNA from a wide variety of sources. Ensuring there are no visible banding patterns in the gel-cut portion of the restriction digest, as was seen in the sparrows, will reduce repetitive DNA that consumes a disproportionate number of reads. If banding patterns are observed, either change restriction enzymes or cut from a different portion of the fragment size range. Cutting a smaller fragment size range would also reduce the number of loci and therefore boost coverage on the loci that are sequenced. A similar result could be obtained by adding more selective bases to the reverse primer.

4.4. Anchoring loci on the genome

Given the broad conservation of chromosomal structure and gene synteny in birds (Ellegren, 2010), our BLAST results against the zebra finch genome validate the hypothesis that the loci we uncovered were distributed randomly throughout the avian genome (Fig. 3). Proportion of hits to a chromosome was directly correlated to the size of the chromosome in the zebra finch. Furthering the idea that restriction digest sites for the sparrows fell in genomic regions that contained relatively high amounts of repetitive DNA, sparrow loci showed the highest proportion of BLAST hits to a virtual "chromosome unknown", which includes repetitive sequences that are difficult to place on a genomic map in the zebra finch (Balakrishnan et al., 2010). Hummingbirds and rails had the highest proportions of loci without a close match to the zebra finch genome, which likely reflects their phylogenetic distance. Given the large size of the Z chromosome, one puzzling aspect of the BLAST results was the lack of loci on the Z chromosome. Also puzzling was that only one locus (in the rails) had a close BLAST hit to mitochondrial DNA and even this fragment was likely a mitochondrial insert into the *Rallus* nuclear genome (J. Maley, unpublished data). Finally, we note that many of the loci we uncovered aligned closely to annotated genes in the zebra finch genome. Although a full description of these genes is beyond the scope of this manuscript, their presence in our data set is extremely promising from the perspective of genome scans.

4.5. Utility for phylogenetics and species-tree analysis

Species-tree analysis has made a large impact on the field of phylogenetics because it allows for more robust phylogenies that account for randomness in the coalescent process (Edwards, 2008). However, if divergence has been very recent or very rapid, the number of loci required for accurate phylogenetic inference can be daunting from an empirical standpoint, ranging from 20 (Maddison and Knowles, 2006) to hundreds, in some cases (Liu et al., 2009). Previous applications of NGS to phylogeography have focused on SNPs (Emerson et al., 2010; Gompert et al., 2010; Williams et al., 2010), which cannot be used in the coalescent-based framework in which most species-tree methods are grounded. With the caveat that we did not set out to generate a data set amenable to species-tree analysis, we were still able to find eight loci with most of the terminal taxa in sparrows and juncos. Despite the fact that the loci were relatively short (290–349 bp) and individual gene trees were discordant, the species tree supported the suspected relationships among *Zonotrichia* and within *Junco*, including the non-monophyly of *J. phaeonotus* and yellow eye color suggested by mtDNA data (Milá, unpublished data). These results provide evidence that loci derived from primary 454 data are amenable to coalescent-based analysis, which opens up a wide array of analytical possibilities in the fields of phylogenetics, population genetics, and phylogeography. Species trees provide a particularly compelling use of 454 data, since the generation of multilocus data sets by traditional means (e.g., anonymous loci or sequencing known nuclear genes) is time-consuming, requires phasing, and often suffers from ascertainment bias (Nielsen et al., 2004). Our method, in conjunction with the analytical pipeline (Hird et al., 2011), produced phased nuclear data immediately amenable to species-tree analysis at the genus-level where incomplete lineage sorting is prevalent. Refinements designed to achieve higher coverage of fewer loci (see above) are likely to result in many more loci than the eight uncovered here.

Finally, as expected, this method is not useful for generating data to assess deep-level phylogenetic relationships because our alignments of higher-level taxa (i.e., rails, hummingbirds, juncos, and sparrows) uncovered only two loci. This was expected given that homologous restriction fragments decrease with increasing phylogenetic distance (Althoff et al., 2007). We conclude that our method is most effectively applied for population genomics and phylogenomics among closely related species.

Acknowledgments

We thank B. Carstens, A. Zellmer, M. Koopman, S. Dowd, B. Milá, J. Pollinger, A. Freedman, Z. Gompert, Lydia Kai, B. Faircloth, T. Glenn, and the Brumfield Lab for help, discussion, and comments on the manuscript. The BioHPC at Cornell and UAF Life Science Informatics Portal provided computational time. UAF Life Science Informatics as a core research resource is supported by Grant Number RR016466 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH). This research was supported by NSF Grants DEB-0841729, DEB-0956069 and IBN-0508611, by Louisiana EPSCoR PFUND-167, and by the James Bond and Alexander Wetmore funds of the Smithsonian Institution. Research permits in Jamaica were issued by the National Environment Planning Agency, Kingston. Collecting permits in Louisiana were granted by United States Fish and Wildlife Service and the Louisiana Department of Wildlife and Fisheries for research on rails. Research on white-crowned sparrows was covered under Federal Fish and Wildlife Banding Permit 22712-G and Collecting Permit MB-813248, Washington State Scientific Collection Permit 04-110, and California Scientific Collecting Permit 801208-05. We thank J. Huner, L. Richard, Jr., and P. Smith, Jr. for access to private land. GRG thanks Brian Schmidt and Errol Francis, and EPD thanks K Oman, for field assistance. We thank C. Cicero at the Museum of Vertebrate Zoology, S. Birks at the Burke Museum, D. Dittmann and S. Cardiff at the Museum of Natural Science, and J. Klicka at the Barrick Museum for identifying samples and providing tissues.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2011.10.012.

References

Althoff, D.M., Gitzendanner, M.A., Segraves, K.A., 2007. The utility of amplified fragment length polymorphisms in phylogenetics: a comparison of homology within and between genomes. Syst. Biol. 56, 477–484.

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

Backström, N., Forstmeier, W., Schielzeth, H., Mellenius, H., Nam, K., Bolund, E., Webster, M.T., Öst, T., Schneider, M., Kempenaers, B., 2010. The recombination landscape of the zebra finch *Taeniopygia guttata* genome. Genome Res. 20, 485–495.

Baird, N., Etter, P., Atwood, T., Currey, M., Shiver, A., Lewis, Z., Selker, E., Cresko, W., Johnson, E., 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3, e3376.

Balakrishnan, C.N., Ekblom, R., Volker, M., Westerdahl, H., Godinez, R., Kotkiewicz, H., Burt, D.W., Graves, T., Griffin, D.K., Warren, W.C., 2010. Gene duplication and fragmentation in the zebra finch major histocompatibility complex. BMC Biol. 8, 29.

Banks, R.C., 1964. Geographic variation in the white-crowned sparrow (*Zonotrichia leucophrys*). Univ. Calif. Publ. Zool. 70, 1–123.

Barbazuk, W.B., Bedell, J.A., Rabinowicz, P.D., 2005. Reduced representation sequencing: a success in maize and a promise for other plant genomes. Bioessays 27, 839–848.

Blanchard, B.D., 1941. The white-crowned sparrows (*Zonotrichia leucophrys*) of the Pacific seaboard: environment and annual cycle. Univ. Calif. Publ. Zool. 46, 1–177.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N., RoyChoudhury, A., submitted for publication. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. <http://arxiv.org/abs/0910.4193>.

Chilton, G., Baker, M.C., Barrentine, C.D., Cunningham, M.A., 1995. White-crowned sparrow no. 183. In: Poole, A., Gill, F. (Eds.), The Birds of North America. The American Ornithologist's Union. Academy of Natural Sciences of Philadelphia.

Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., Marsh, T., Garrity, G.M., Tiedje, J.M., 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 37, D141–D145.

Corbin, K., 1981. Genic heterozygosity in the white-crowned sparrow: a potential index to boundaries between subspecies. Auk 98, 669–680.

Corbin, K., Wilkie, P., 1988. Genetic similarities between subspecies of the white-crowned sparrow. Condor 90, 637–647.

Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214.

Eddleman, W.R., Conway, C.J., 1998. Clapper Rail (*Rallus longirostris*). In: Poole, A., Gill, F. (Eds.). The Birds of North America, Inc., Philadelphia, PA.

Edwards, S., 2008. Is a new and general theory of molecular systematics emerging? Evolution 63, 1–19.

Ellegren, H., 2010. Evolutionary stasis: the stable chromosomes of birds. Trends Ecol. Evol. 25, 283–291.

Emerson, K., Merz, C., Catchen, J., Hohenlohe, P., Cresko, W., Bradshaw, W., Holzapfel, C., 2010. Resolving postglacial phylogeography using high-throughput sequencing. Proc. Natl. Acad. Sci. USA 107, 16196–16200.

Evanno, G., Regnaut, S., Goudet, J., 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. 14, 2611–2620.

Garber, K., 2008. Fixing the front end. Nat. Biotechnol. 26, 1101–1104.

Gill, F., Stokes, F., Stokes, C., 1973. Contact zones and hybridization in the Jamaican Hummingbird, *Trochilus polytmus* (L.). Condor 75, 170–176.

Glenn, T.C., 2011. Field guide to next-generation DNA sequencers. Mol. Ecol. Res. 11, 759–769.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. 27, 182–189.

Gompert, Z., Forister, M.L., Fordyce, J.A., Nice, C.C., Williamson, R.J., Buerkle, C.A., 2010. Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. Mol. Ecol. 19, 2455–2473.

Grinnell, J., 1928. Notes on the systematics of west American birds. III. Condor 30, 185–189.

Heled, J., Drummond, A.J., 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27, 570–580.

Hird, S.M., Brumfield, R.T., Carstens, B.C., 2011. PRGmatic: an efficient pipeline for collating genome-enriched second generation sequencing data using a "provisional-reference genome". Mol. Ecol. Res. doi:10.1111/j.1755-0998.2011.03005.x.

Hohenlohe, P., Bassham, S., Etter, P., Stiffler, N., Johnson, E., Cresko, W., 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. 6, e1000862.

Hohenlohe, P.A., Amish, S.J., Catchen, J.M., Allendorf, F.W., Luikart, G., 2011. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. Mol. Ecol. Res. 11, 117–122.

Holsinger, K.E., 2010. Next generation population genetics and phylogeography. Mol. Ecol. 19, 2361–2363.

Hyten, D., Song, Q., Fickus, E., Quigley, C., Lim, J., Choi, I., Hwang, E., Pastor-Corrales, M., Cregan, P., 2010. High-throughput SNP discovery and assay development in common bean. BMC Genomics 11, 475.

Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K., Ding, L., 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25, 3.

Kuhner, M., 2009. Coalescent genealogy samplers: windows into population history. Trends Ecol. Evol. 24, 86–93.

Lance, S.L., Hagen, C., Glenn, T.C., Brumfield, R.T., Stryjewski, K.F., Graves, G.R., 2009. Fifteen polymorphic microsatellite loci from Jamaican streamertail hummingbirds (*Trochilus*). Conserv. Genet. 10, 1195–1198.

Lerner, H., Fleischer, R., 2010. Prospects for the use of next-generation sequencing methods in ornithology. Auk 127, 4–15.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 7.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2.

Liu, L., Yu, L., Pearl, D.K., Edwards, S.V., 2009. Estimating species phylogenies using coalescence times among sequences. Syst. Biol. 58, 468–477.

Maddison, W., Knowles, L., 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55, 21–30.

Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J., Turner, D.J., 2009. Target-enrichment strategies for next-generation sequencing. Nat. Methods 7, 111–118.

Meanley, B., 1992. King rail. In: Poole, A., Stettenheim, P., Gill, F. (Eds.), The Birds of North America. The Academy of Natural Sciences, Philadelphia, PA.

Milá, B., McCormack, J., Castañeda, G., Wayne, R., Smith, T., 2007. Recent postglacial range expansion drives the rapid diversification of a songbird lineage in the genus *Junco*. Proc. Roy. Soc. B – Biol. Sci. 274, 2653–2660.

Miller, M.R., Dunham, J.P., Amores, A., Cresko, W.A., Johnson, E.A., 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 17, 240–248.

Nielsen, R., Hubisz, M.J., Clark, A.G., 2004. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics 168, 2373–2382.

Niu, B., Fu, L., Sun, S., Li, W., 2010. Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC Bioinf. 11, 187.

Nosil, P., Funk, D.J., Ortiz Barrientos, D., 2009. Divergent selection and heterogeneous genomic divergence. Mol. Ecol. 18, 375–402.

Pinho, C., Hey, J., 2010. Divergence with gene flow: models and data. Annu. Rev. Ecol. Evol. Syst. 41, 215–230.

Pritchard, J., Wen, W., 2004. Documentation for STRUCTURE Software: Version 2. Department of Human Genetics, University of Chicago, Chicago. <http://pritch.bsd.uchicago.edu>.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Rokas, A., Abbot, P., 2009. Harnessing genomics for evolutionary insights. Trends Ecol. Evol. 24, 192–200.

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

Via, S., West, J., 2008. The genetic mosaic suggests a new role for hitchhiking in ecological speciation. Mol. Ecol. 17, 4334–4345.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M., 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. 11, 4407–4414.

Wheat, C.W., 2010. Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. Genetica 138, 433–451.

Whitelaw, C.A., Barbazuk, W.B., Pertea, G., Chan, A.P., Cheung, F., Lee, Y., Zheng, L., van Heeringen, S., Karamycheva, S., Bennetzen, J.L., SanMiguel, P., Lakey, N., Bedell, J.A., Yuan, Y., Budiman, M.A., Resnick, A., Van Aken, S., Utterback, T., Riedmuller, S., Williams, M., Feldblyum, T., Schubert, K., Beachy, R., Fraser, C.M., Quackenbush, J., 2003. Enrichment of gene-coding sequences in maize by genome filtration. Science 302, 2118–2120.

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

Williams, L., Ma, X., Boyko, A., Bustamante, C., Oleksiak, M., 2010. SNP identification, verification, and utility for population genetics in a non-model genus. BMC Genet. 11, 32.

Wu, C.I., 2001. The genic view of the process of speciation. J. Evol. Biol. 14, 851–865.

Zink, R.M., Blackwell, R.C., 1996. Patterns of allozyme, mitochondrial DNA, and morphometric variation in four sparrow genera. Auk 59, 67.