MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2014) **14**, 1000–1010 doi: 10.1111/1755-0998.12249

Sequence capture using PCR-generated probes: a cost-effective method of targeted high-throughput sequencing for nonmodel organisms

JOSHUA V. PEÑALBA,* LYDIA L. SMITH,*† MARIA A. TONIONE,*‡ CHODON SASS,§ SARAH M. HYKIN,* PHILLIP L. SKIPWITH,* JIMMY A. MCGUIRE,*† RAURI C. K. BOWIE*† and CRAIG MORITZ†¶

*Museum of Vertebrate Zoology, University of California, 3101 Valley Life Sciences Building, Berkeley, CA 94720, USA, †Department of Integrative Biology, University of California, 3060 Valley Life Sciences Building, Berkeley, CA 94720, USA, ‡Department of Environmental Science Policy and Management, University of California, 130 Mulford Hall, Berkeley, CA 94720, USA, §UC and Jepson Herbarium, University of California, 411 Koshland Hall MC 3102, Berkeley, CA 94720, USA, ¶Research School of Biology, The Australian National University, Building 116, Acton, ACT 0200, Australia

Abstract

Recent advances in high-throughput sequencing library preparation and subgenomic enrichment methods have opened new avenues for population genetics and phylogenetics of nonmodel organisms. To multiplex large numbers of indexed samples while sequencing predominantly orthologous, targeted regions of the genome, we propose modifications to an existing, in-solution capture that utilizes PCR products as target probes to enrich library pools for the genomic subset of interest. The sequence capture using PCR-generated probes (SCPP) protocol requires no specialized equipment, is highly flexible and significantly reduces experimental costs for projects where a modest scale of genetic data is optimal (25–100 genomic loci). Our alterations enable application of this method across a wider phylogenetic range of taxa and result in higher capture efficiencies and coverage at each locus. Efficient and consistent capture over multiple SCPP experiments and at various phylogenetic distances is demonstrated, extending the utility of this method to both phylogeographic and phylogenomic studies.

Keywords: high-throughput sequencing, phylogenomics, phylogeography, sequence capture

Received 8 July 2013; revision received 11 February 2014; accepted 12 February 2014

Introduction

With the continually decreasing cost of data collection using high-throughput sequencing (HTS) techniques, there has been a movement to exploit this technology in the fields of phylogeography and phylogenetics. To address most questions in these fields, the data needed are orthologous loci from multiple individuals, typically from nonmodel organisms with deep and shallow timescales of divergence. Traditionally, this is performed by separately Sanger sequencing each individual for every locus of interest. A constraint typically seen with Sanger sequencing is that often some individuals repeatedly fail to PCR-amplify at particular loci, resulting in frustratingly patchy data matrices. Due to several novel and modified methods using HTS, researchers are now capable of sequencing significant numbers of orthologous loci for many individuals at far lower costs and in a fraction of the time compared to Sanger sequencing.

Obtaining homologous loci from the DNA of a larger-genome organism often requires reduction to a targeted genomic subset of interest. Some of the more common methods to reduce the genome include reduced representation libraries (Sánchez *et al.* 2009; Van Bers *et al.* 2010), restriction-site-associated DNA (RAD) sequencing (Hohenlohe *et al.* 2010; Peterson *et al.* 2012), amplicon sequencing (Chan *et al.* 2010; O'Neill *et al.* 2013) and transcriptome sequencing (Parchman *et al.* 2010; Smith *et al.* 2011). The advantages and disadvantages of these approaches and the bioinformatics associated with each were reviewed by Good (2011), McCormack *et al.* (2013b) and Lemmon and Lemmon (2013).

Another increasingly popular approach is capture by hybridization, a method in which specifically targeted markers from individuals belonging to multiple

Correspondence: Joshua Peñalba, Research School of Biology, Building 116, The Australian National University, Acton, ACT 0200, Australia. Fax: (+61 2) 6125 5573; E-mail: josh.penalba@gmail.com

populations or species are sequenced (reviewed in Mamanova *et al.* 2010). There are currently two approaches for target enrichment by hybridization: array-based captures which use commercially generated probes immobilized on a chip (Bi *et al.* 2012; Hancock-Hanser *et al.* 2013) and in-solution capture which can be implemented with either commercial or laboratory-generated probes (Mason *et al.* 2011; Faircloth *et al.* 2012; Lemmon *et al.* 2012; Li *et al.* 2013). Regardless of the method, hybrid capture requires a priori knowledge about the genomic sequence of the targets of interest for probe design. Genome reduction is achieved by hybridizing genomic DNA libraries to short nucleic acid probes, discarding fragments that did not hybridize and sequencing the fragments that were successfully captured. By consistently retaining homologous sequence-based loci across individuals, these capture methods are proving to be an effective means of generating massive multilocus data sets.

In this study, we demonstrate modifications to an existing in-solution capture protocol devised for mitogenome sequencing by Maricic *et al.* (2010). This method is part of a small suite of methods that use PCR-generated probes, rather than commercially generated probes, to target the orthologous regions during hybridization (Noonan *et al.* 2006; Horn 2012). Hereafter, we will refer to this method as 'sequence capture using PCR-generated probes' (SCPP – pronounced 'skip'). Here, we expand this method to include a large number of nuclear loci in addition to the mtDNA, and we demonstrate its effectiveness for nonmodel organisms. With the available primer resources from the literature (Backström *et al.* 2008; Kimball *et al.* 2009; Portik *et al.* 2012), the impressive accumulation of phylogenetically diverse genomes (Ellegren *et al.* 2012; Scally *et al.* 2012; Amemiya *et al.* 2013), and the ability to use high-throughput sequencing of genomes or transcriptomes to mine for markers (Parchman *et al.* 2010; Bi *et al.* 2012; Lemmon & Lemmon 2012), obtaining PCR primers for informative loci has become significantly easier. Additionally, this method holds special value for those laboratories transitioning from Sanger sequencing approaches that already have large numbers of well-optimized primer pairs and corresponding data for their study system.

The types of questions the SCPP method allows us to address are those that require multiple kilobase-scale sequence data from dozens to hundreds of individuals and dozens of loci. It will particularly benefit studies for which full-length loci with linked polymorphisms are more informative than isolated SNPs and a moderate number of loci are required for better resolution (Carling & Brumfield 2007; Edwards 2008; Brito & Edwards 2009). We propose the SCPP method as an alternative, cost-effective way to sequence a modest but sufficient number of loci for nonmodel organisms.

This experiment focuses on the utility of this method to capture orthologous nuclear markers from individuals across a wide range of phylogenetic divergences and assess systems with different levels of a priori information. To do this, we chose to test the capture efficiency of SCPP across three systems: *Enyalius*, complex of lizards found in the Atlantic Rainforest of Brazil; *Draco*, South-East Asian gliding lizards; and species from various deeply diverged families across the songbirds (Passeriformes).

Methods

Samples

Enyalius. To test the efficiency of SCPP at a relatively shallow phylogenetic depth (average *ND4* p-distance ~9.0%), we carried out two projects: a small pilot study to identify loci with consistent coverage and capture success and a larger applied study. The smaller study consisted of 10 individuals, from the following species: *Enyalius pictus*, *E. catenatus*, *E. erythroceneus*, *E. bibroni* and *E. perditus*. For the larger project, we prepared libraries from 60 individuals from the species above and one additional species (*E. bilineatus*).

Draco. The *Draco* project aimed to capture species across a greater phylogenetic depth (average *ND2* p-distance ~16.33%). We selected six *Draco* samples from the species *Draco boschmai* (2), *D. maximus* (1), *D. sumatranus* (1), *D. timorensis* (1) and *D. volans* (1).

Passeriformes. The passerine project's aim was to test the efficiency of the SCPP method across deep phylogenetic distances (average *ND2* p-distance ~22.04%). We chose one sample from each of 13 selected families from the order Passeriformes within the three major clades. From the suboscines, we chose *Smithornis capensis* (Eurylaimidae). From the oscines, we chose representatives from the two major clades. The samples from the 'core Corvoidea' clade included *Cyanocitta stelleri* (Corvidae) and *Vireo cassinii* (Vireonidae). The samples from the 'Passerida' clade included *Turdus migratorius* (Turdidae), *Regulus calendula* (Regulidae), *Apalis flavigularis* (Cisticolidae), *Eremophila alpestris* (Alaudidae), *Chamaea fasciata* (Sylviidae), *Sitta carolinensis* (Sittidae), *Carduelis psaltria* (Fringilidae), *Nectarinia olivacea* (Nectarinidae), *Tachycineta bicolor* (Hirundinidae) and *Bombycilla cedrorum* (Bombycillidae). Finally, we chose one out-group from the order Psittaciformes: *Amazona amazonica* (Psittacidae) Table 1.

Table 1 Raw and filtered read numbers for all samples. The columns to the right display the percentage of reads passing filter that mapped to the targets. The counts for any species with more than one sample were averaged between all samples

Species	Sample	Raw reads	Filtered reads (%)	% Mapped (nuclear)	% Mapped (mitochondrial)	% Mapped (total)	Average coverage (nu)	Average coverage (mt)
Smithornis capensis	MLW-B10	296107	104351 (35.24)	20.88	16.23	37.11	43.06	1088.37
Cyanocitta stelleri	MVZ183639	358373	123605 (34.49)	32.95	20.14	53.09	81.47	1685.11
Vireo cassinii	MVZ170215	348951	121999 (34.96)	26.76	22.31	49.07	64.55	1773.07
Turdus migratorius	MVZ179251	258778	96778 (37.39)	34.44	32.18	66.62	68.20	1702.92
Regulus calendula	MVZ183153	524594	174693 (33.33)	33.04	31.91	64.95	108.95	3053.52
Apalis flavigularis	RCKB 1304	141060	62135 (44.04)	76.34	0.19	76.53	101.73	10.37
Eremophila alpestris	MVZ171841	233246	94670 (40.59)	42.86	19.12	61.98	86.06	1157.10
Chamaea fasciata	MVZ178263	995129	249726 (25.09)	14.74	40.96	55.71	74.04	4807.95
Sitta carolinensis	MVZ177836	374620	138750 (37.03)	43.74	23.97	67.71	118.07	1758.68
Carduelis psaltria	MVZ176672	245849	97702 (39.74)	46.79	26.79	73.58	90.29	1453.26
Nectarinia olivacea	MLW-B6	1112429	281667 (25.31)	18.26	35.69	53.95	104.96	3019.32
Tachycineta bicolor	MVZ182221	1028628	245910 (23.90)	16.03	56.35	72.38	76.73	5340.55
Bombycilla cedrorum	MVZ180052	401764	135123 (33.63)	29.75	39.15	68.90	82.61	3384.43
Amazona amazonica	MVZ181954	176092	63980 (36.33)	22.34	42.14	64.48	30.86	1451.74
Draco boschmai	JAM11647	557261	121991 (21.89)	6.38	11.57	17.96	43.74	1191.35
Draco boschmai	JAM11744	499933	110205 (22.04)	3.95	5.26	9.21	22.54	717.22
Draco timorensis	JAM12801	625569	134059 (21.42)	4.00	16.63	20.63	24.28	1274.33
Draco volans	JAM2079	445707	109778 (24.63)	11.18	4.29	15.47	57.53	454.12
Draco sumatranus	JAM4038	568687	143592 (24.63)	6.16	25.19	31.34	42.95	2345.20
Draco maximus	RMBR1002	224936	56193 (24.98)	6.87	3.56	10.42	20.49	170.75
Enyalius bibroni	MTR22506	2806296	693358 (24.71)	1.75	54.84	56.59	100.54	4597.79
Enyalius bilineatus	JC771	3286056	763048 (23.22)	0.81	66.31	67.12	74.15	5445.32
Enyalius catenatus	Various (29)	2200180	625491 (28.42)	1.90	53.62	55.52	132.73	4560.71
Enyalius erythroceneus	Various (4)	2671357	649791 (24.32)	1.74	58.84	60.58	134.59	4893.03
Enyalius perditus	Various (4)	1650479	474408 (28.74)	1.67	37.62	39.29	87.07	3864.86
Enyalius pictus	Various (21)	2399936	570314 (23.76)	1.42	59.85	61.27	84.15	4282.82

Probe selection

The marker sets selected for each project were an independent but diverse array of loci, both in terms of location in the genome (autosomal, sex-linked, and mitochondrial loci) and function (exons, introns, anonymous loci). Amplicon sizes ranged from 300 to 1500 base pairs. The secondary *Enyalius* project included a total of 19 gene regions: 17 autosomal loci and two mitochondrial genes. The *Draco* project targeted 51 gene regions: four known autosomal loci, 44 anonymous nuclear loci and three mitochondrial genes. The passerine project targeted a total of 81 gene regions: 71 autosomal loci, 6 Z-linked loci and 4 mitochondrial genes. The total length of the targeted loci was approximately 7.5 kilobases for *Enyalius*, 36 kilobases for *Draco* and 52 kilobases for passerines.

For the pilot *Enyalius* project, we prepared 22 gene regions as bait and tested for utility in the larger project. Eighteen of these were anonymous loci, discovered from a previous 454 sequencing run on *E. catenatus* (Gardner *et al.* 2011). Additional nuclear (*KIAA1549* and *EXPH5*) and mitochondrial (*ND4* and *16S*) markers were amplified using published primers (Arèvalo *et al.* 1994; Palumbi 1996; Portik *et al.* 2012). The resulting data from this pilot study determined which loci were to be used for the larger secondary study: 15 of the 18 anonymous loci tested were retained for the larger study along with the nuclear and mitochondrial markers. We regarded markers that displayed unusually high coverage and variation as potential multicopy markers and were omitted during the larger study. Similar to the *Enyalius* project, most of the loci used for the *Draco* project were anonymous loci from two separate reduced representation Illumina sequencing runs targeting different sets of *Draco* species. However, these loci were not tested in advance for multicopy loci. The loci used in the passerine project were all from published primers (Backström *et al.* 2008; Kimball *et al.* 2009; Wang *et al.* 2012). Detailed information on each locus can be found in Tables S1 and S2 (Supporting information).

For the *Enyalius* and *Draco* projects, we generated probes from species within the same genus as the samples to be captured. For *Enyalius*, three species (*E. pictus*, *E. catenatus* and *E. erythroceneus*) were PCR-amplified, and the resulting amplicons were used as capture

probes. For *Draco*, the loci were amplified from the same species as the samples prepared for libraries. As the passerine project spanned the entire order Passeriformes, the probes were generated from five species within the Pycnonotidae (*Andropadus virens*, *A. fusciceps*, *A. tephrolaemus*, *A. milajensis* and *Phyllastrephus alfredi*), a family deeply nested within the passerine tree. Multiple species were chosen to increase the likelihood of successful amplification. Attempts were made to amplify each species at each locus, and PCR products were pooled for all species for which PCR amplification was successful.

DNA extraction and library preparation

Genomic DNA (gDNA) was extracted from liver or muscle tissue for library preparation. Two methods of extractions were performed: for the Passeriformes and *Draco* projects, the Qiagen DNeasy tissue extraction protocol was used, whereas for the *Enyalius* project, a high-salt precipitation protocol was employed (Miller *et al.* 1988). For all projects, the concentrations of the extractions were measured using a Qubit fluorometer, and a total of 120 μL of 10 μg/μL gDNA was sheared using a Bioruptor sonicator (Diagenode) prior to library preparation. The samples were fragmented to < 500 bp by being subjected to 3–5 rounds of sonication, each of which included two periods of 3.5-min sonication on high with a 0.5-min rest period between each sonication period. The Illumina TruSeq kit was used for library preparation for the pilot *Enyalius*, *Draco* and passerine projects and, to reduce costs, the secondary *Enyalius* project used the in-house protocol described by Meyer & Kircher (2010).

Modifications

Amplicon bait generation. The first modification made to the Maricic *et al.* (2010) protocol (Fig. 1) was in the manner that we generated the amplicon bait. Similar to Noonan *et al.* (2006), the amplicon bait was generated for multiple nuclear and mitochondrial loci using standard PCR conditions and without subsequent sonication. After amplification, 30–50 μL of PCR product was pooled from different individuals and run on an agarose gel, and the corresponding band was excised and purified together using a Qiagen QIAquick PCR Purification kit. Finally, all bait loci for a given project were pooled together in equimolar concentrations based on quantification with the Qubit fluorometer (Life Technologies) to a total of 1.3 μg. Two aliquots of bait were prepared to allow for sequential enrichment hybridizations (see below).

COT-1 blocker. Human Cot-1 DNA (Invitrogen) was added as an additional component of the capture reaction in step 22 of the Maricic *et al.* (2010) protocol. Cot-1 DNA is an enriched collection of 50-300 bp highly repetitive genomic elements used to block nonspecific hybridization in sequence captures (Hodges *et al.* 2007; Bi *et al.* 2012). A total of 5.2 μL of 1 mg/mL human Cot-1 DNA was added to the hybridization reaction, and the volume of the pooled libraries was reduced to 12.8 μL to keep a consistent total volume of 52 μL. Although the volume of the pooled libraries was reduced, the total mass of the libraries was kept consistent as 2 μg.

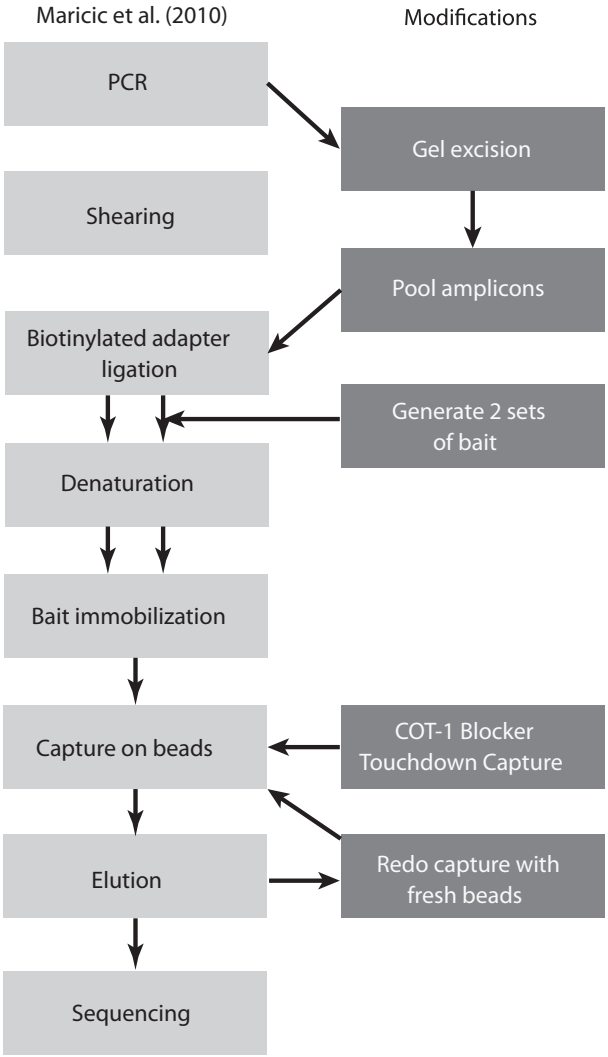

Fig. 1 Protocol modifications. The laboratory pipeline to the *left* depicts the protocol as presented by Maricic *et al.* (2010). The modifications implemented on this study are depicted to the right of this pipeline.

Hybridization parameters. Most of the modifications to the Maricic *et al.* (2010) protocol were implemented during the hybridization steps to increase capture efficiency of targeted material and to minimize the retention of

nontarget material. Independently, Li *et al.* (2013) applied identical modifications using commercial RNA bait. The first modification was a touch-down hybridization as used in Mason *et al.* (2011) coupled with a longer overall hybridization time. The hybridization started at 65 °C and was decreased by 2.5 °C every 24 h over 72 h and ended at 60 °C. This was performed to gradually decrease stringency and increase the likelihood of capturing fragments with higher sequence divergence from the amplicon bait. The second modification involved repeating the capture using the eluate obtained from the first hybridization. As assessed by quantitative PCR (qPCR), early trials found that using freshly generated bait for the second capture retained more of the targeted loci compared with reusing the bait and beads from the first capture. Using freshly generated beads is similar to what was done by Horn (2012). A summary of these modifications is depicted in Fig. 1, and a detailed protocol of the methods can be found in Appendix S2 (Supporting information).

Enrichment evaluation and sequencing

Quantitative PCR was used to evaluate the enriched library pool prior to sequencing. For each project, internal primers were designed from both targeted markers and nontargeted markers to be used as positive and negative controls of capture, respectively. Primers were designed to amplify 100- to 150-bp fragments within the total length of the locus. Three nuclear loci and one mitochondrial gene were selected for positive controls; one nuclear locus and one mitochondrial gene were used for negative controls. Both original and target-enriched libraries were used as template for amplification to detect shifts in amplification plots due to enrichment. After evaluation, the samples were sequenced using the Illumina platform. The passerine project was sequenced in 90% of one MiSeq lane (150-bp paired-end reads), the *Draco* project was 65% of one MiSeq lane (150-bp paired-end reads), and the *Enyalius* project was sequenced using one HiSeq lane (100-bp paired-end reads).

Bioinformatics

Raw reads were rigorously filtered prior to assembly and analysis. First, low-quality reads, PCR duplicate reads and optical duplicate reads were removed. Second, a rigorous adapter trimming process was implemented using Trimmomatic (v.0.20), Cutadapt (v.1.2.1) and Bowtie2 (Martin 2011; Langmead & Salzberg 2012; Lohse *et al.* 2012). Third, overlapping paired-end reads were merged into single long reads using FLASH and COPEread (Magoč & Salzberg 2011; Liu *et al.* 2012). Lastly, the reads were filtered for contamination. The cleaned reads for each library were assembled de novo using ABySS (v.1.3.4) (Simpson *et al.* 2009). Contigs from all assemblies were combined for a second round of assembly. Identical contigs were clustered together using CD-HIT-EST (v.4.5.4) and BLAT (v.34) to make the file smaller for assembly (Kent 2002; Li & Godzik 2006). Finally, the clusters were assembled using CAP3 to generate longer contigs (Huang & Madan 1999).

The final step was to identify which of the final assembled contigs corresponded to the targeted loci to serve as a reference for mapping. A representative sequence was collected from Sanger sequencing or from NCBI GenBank for each locus, and the final assembled contigs were matched using BLAST or a longer method shown in Fig. 2. Each library was aligned to its corresponding references using Novoalign (http://www.novocraft.com). We called haplotypes using GATK's (http://www.broadinstitute.org/gatk/) read-backed phasing algorithm. A combination of Perl and Python scripts were implemented to run these programs, and the pipeline is available in http://github/MVZSEQ/SCPP. The Perl scripts are modifications from the pipeline from Singhal (2013). A more detailed description of the bioinformatics can be found in Appendix S1 (Supporting information).

Results

Raw and filtered data

The raw and filtered read information is summarized in Table 1. The passerine project produced a total of 6 495 620 reads. The *Draco* project resulted in 3 474 538 reads. The *Enyalius* project resulted in 137 583 599 reads. After filtering, 34.36% (Passerine), 23.37% (*Draco*) and 27.62% (*Enyalius*) of the raw reads were recovered. The majority of the filtered data were of duplicate reads, encompassing approximately 75% of all reads. Contaminant and low complexity reads were negligible for all samples.

Capture performance

Capture performance of the SCPP method was measured in three ways: (i) capture specificity, (ii) capture sensitivity and (iii) capture uniformity. Capture specificity measures the percentage of postfiltered reads that map to targets. Detailed information on the capture specificity of each taxon is described in Table 1. Average capture specificity varied between projects: Passeriformes – 61.86%, *Draco* – 17.51% and *Enyalius* – 57.00% (Fig. 3).

Here, capture sensitivity is defined as the percentage of targeted loci that are covered by at least one read. Coverage for each locus and each individual is depicted in the Tables S3–S5 (Supporting information). The

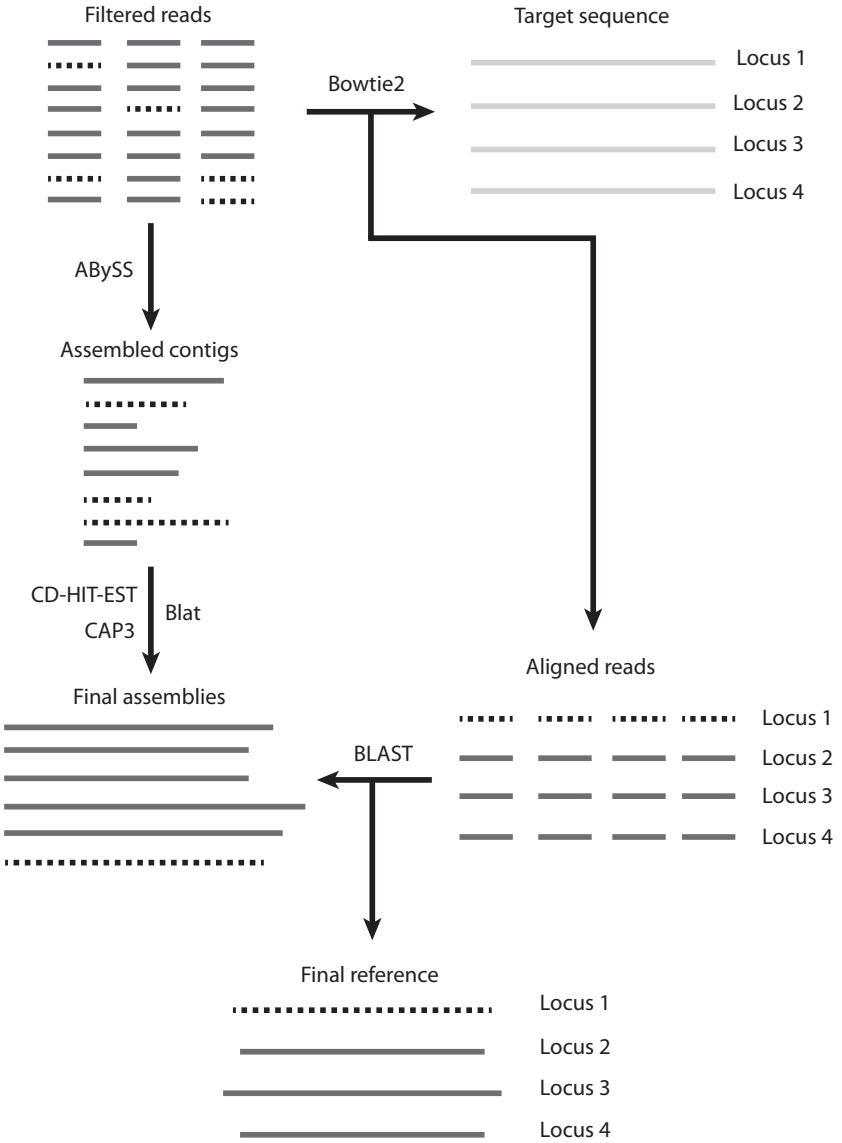

Fig. 2 Generating references. Reference sequence was generated using this pipeline for the passerine project where a reference sequence from a congeneric taxon was not available. The dashed bars indicate the path of a set of reads that correspond to a particular locus.

average coverage for each locus is an underestimate of the coverage spanning the targets as the average includes the flanking regions. Averaged coverages for nuclear and mitochondrial loci can be found in Table 1. On average, mitochondrial coverage is 35 times higher than nuclear coverage. Considering that mitochondrial targets only comprise 2.4–5.9% of the total target length, the excess coverage of the mitochondrial loci can be attributed to the much higher copy number of the mitochondrial genome compared with the nuclear genome per cell. The only sample in which this is not observed is *Apalis flavigularis* from the passerine project. There is no clear reason as to why this sample is an outlier.

We recovered 99% of the data matrix for the passerines and 100% for *Enyalius*. Where coverage dropped below 20X, this typically represented low efficiency across all individuals for a particular locus, or across a particular individual for most loci (Tables S3 and S4, Supporting information); in both cases, these loci/individuals can be repeated or dropped to yield a complete data matrix. The more heterogeneous results for *Draco* (85%; Table S5, Supporting information) reflect the overall lower specificity, as discussed later.

Finally, capture uniformity assesses the evenness of coverage throughout the length of a target. Capture uniformity averaged among all nuclear loci, and all samples within a project is depicted in Fig. 4. For the passerine project, the in-target region was located to assess the amount of flanking region captured. On average, approximately 20% of each side of the contig was flanking regions of the target. Average nuclear

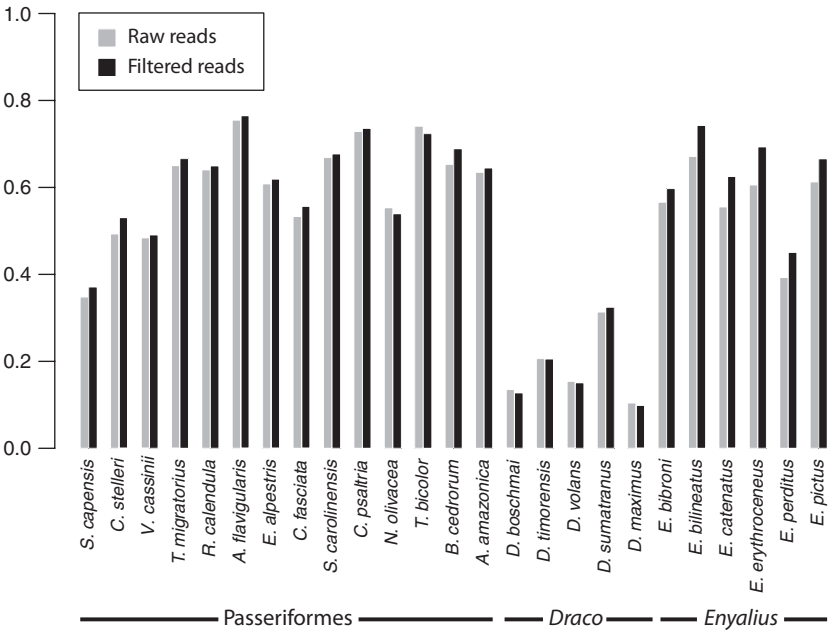

Fig. 3 Capture specificity. The bars depict the proportion of the raw and filtered reads that map to the targets.

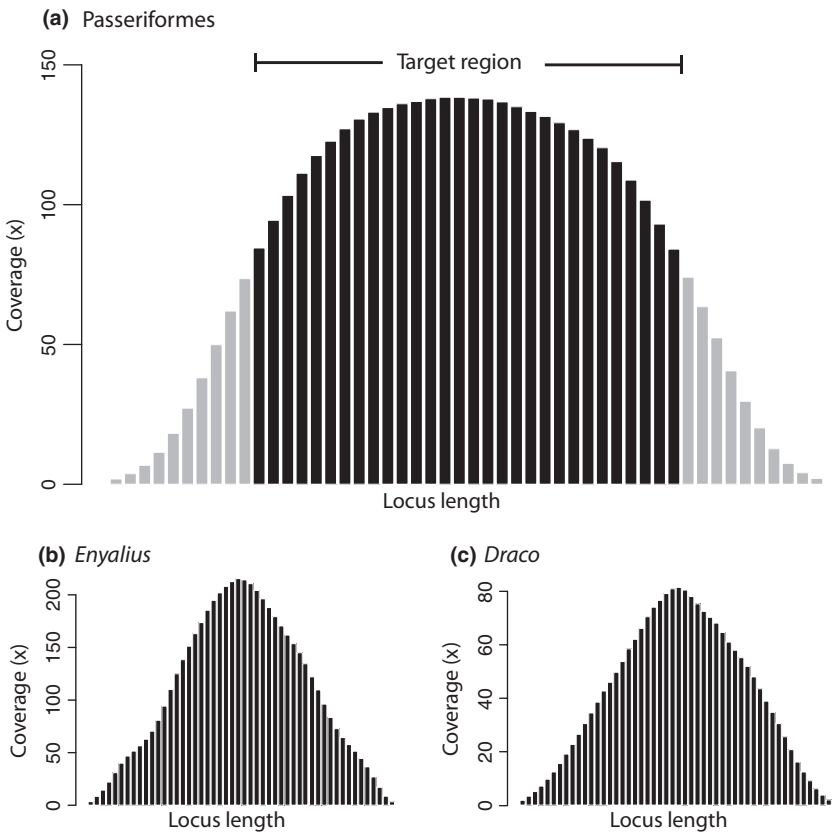

Fig. 4 Capture uniformity. Results represent all nuclear loci for all samples combined. Each bar depicts 2% of the length of the locus captured. The average base coverage within each bin is depicted on the *y*-axis. The darker bars in the middle represent the length of the PCR-generated probe, and the lighter bars depict the flanking region that extends beyond the primer binding site (target and flanking regions are not depicted in panels b and c).

coverage within the target locus was 91X compared with the average nuclear coverage of the entire contig of 77X. For the passerine project, average coverage of the targets without the flanking region can be found in the supplementary material (Table S6, Supporting information). Coverage within the target was largely uniform, although a drop-off on the flanking regions was observed.

Divergence and coverage

The passerine project captured samples with a minimum of 19.94% (Hirundinidae) and a maximum of 27.14% (Eurylaimidae) mitochondrial divergence (*ND2* p-distances) from the taxa used to generate the capture probes. Within the Passerida clade, species more closely related to bait species did not necessarily hybridize better than species at a greater phylogenetic distance (Fig. 5). Actual assessment is difficult as the differences observed here may reflect pipetting and assessment errors in making library pools in addition to true capture sensitivity variation. However, outside of Passerida, there is a clear pattern in which the amount of coverage decreases with the growing evolutionary distance. A plot comparing sequence divergence with coverage can be found in the Fig. S1 (Supporting information). In the data matrix that uses a 20X average coverage cut-off, 95.5% of the matrix is full for the oscine clade, 93.7% of the matrix is complete when the suboscine clade is included, and 90.5% of the matrix is complete when the out-group is included.

Discussion

This study demonstrates that the SCPP method can successfully and efficiently capture nuclear loci in addition to mitochondrial targets and at moderate phylogenetic distance. Our capture efficiency for either raw or filtered reads was higher than the 16% efficiency observed by Maricic *et al.* (2010), in some cases (*Enyalius*, passerines), significantly more so. In all projects, there is a high level of PCR duplicates; this is likely due to the moderate genomic target size. Moderate target sizes such as these may not require the large number of reads we allotted to them for the trial runs, leading to an oversequencing of reads. Future projects can take advantage of this space by pooling more individuals to sequence.

The advantages of this method over Sanger sequencing extend beyond the low cost of sequencing by greatly reducing the labour required for optimization, PCR amplification and sequencing of individual loci compared with Sanger sequencing. With the SCPP method, there is a high likelihood of obtaining sequence data from all the individuals and all of the loci in just a few experiments. As the PCR-generated bait probe is much longer than the standard length PCR primer, the sensitivity for capture is less affected by mutations or indels than a typical PCR. This is particularly important in projects that include species with deeper divergences, where mutations in primer binding sites are more likely.

Although the method has generally worked well, the *Draco* project seems to have performed at a lower quality in comparison with the other experiments. The low capture specificity of the *Draco* project can likely be attributed to the use of previously uncharacterized anonymous loci. Reducing mapping stringencies result in a two- to threefold increase in mapped reads. This can be due to multicopy loci in the array of targets leading to nonspecific binding. Additionally, primers that have not been tested could have amplified and captured a different locus than the desired target. Researchers should be cautious when identifying orthologs for analysis. Current methods cannot guarantee the targeting of orthologous loci, and further tests to screen out paralogous sequences will be crucial for robust downstream analyses and inference.

One of the advantages of the SCPP method is how well it worked with whole clades of organisms. This

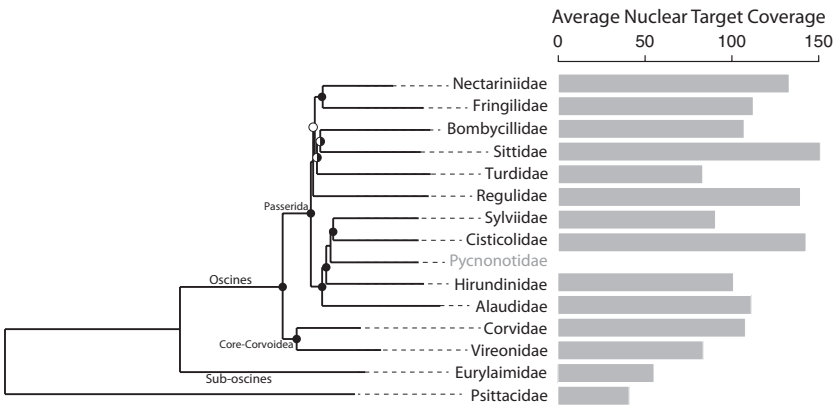

Fig. 5 Divergence and coverage. This plot depicts the average coverage of all nuclear loci for each target in the passerine project. Bait sequences were generated from multiple species in Pycnonotidae. Phylogenetic methods can be found in the supplementary material. The topology represents the multilocus ML tree except for Pycnonotidae whose location on the tree is based on Barker *et al.* 2004. The left half of the circle on the nodes represents maximum likelihood with a black fill depicting bootstrap values ≥70, and right half represents Bayesian inference with a black fill depicting Bayesian posterior probabilities ≥0.95.

makes the method easily applicable to many phylogenetic projects, expanding the application beyond that of Maricic *et al.* (2010) which applied the method to human populations. The only drop in coverage was seen when including the out-group in the passerine project, which is likely caused by hybridization inefficiency due to mutations. The out-group to the passerines was still captured with sufficient average coverage to allow for genotyping, but four loci were not captured and coverage is patchy in others. The plot comparing sequence divergence between the probes and target DNA shows a reduction in coverage at >10% sequence divergence to the probes (Fig. S1, Supporting information). However, the data still demonstrate the ability of the SCPP method to efficiently capture nuclear data for target taxa that have moderate divergence from the bait species.

Although this method has many benefits, certain modifications can potentially improve capture efficiency. First, as longer PCR probes tend to lead to higher average coverage (Fig. S2, Supporting information), shearing could improve capture consistency between different loci by having a narrow probe size range. Another possible modification pertains to the proportion of mitochondrial to nuclear PCR-generated bait. Disproportionately more reads from mitochondrial loci were obtained due to the high ratio of mitochondrial to nuclear genomes in our DNA extractions. As mitochondrial genome bycatch was negligible (Fig. S3, Supporting information), performing separate captures for nuclear and mitochondrial targets may be the best way to compensate for the variations in genome copy number. To maximize efficiency, we encourage the use of taxon specific Cot-1 DNA and the use of a small pilot project to screen against multicopy loci. We are also optimistic that SCPP can also be used to capture historical or damaged DNA. Further experimentation and modification of the method is highly encouraged.

In comparison with alternative target-enrichment probe sets, the SCPP method provides a way to enrich for a moderate target size (25–100 loci) of interest. For the many projects in which this moderate target size range is sufficient to answer the research questions, users are able to pool more individuals per capture and per sequencing lane and still be able to get sufficient coverage for calling haplotypes (Knowles 2010; Rannala & Yang 2013; Lanier *et al.* 2014). Research questions that would require larger genomic regions would likely benefit from available probe sets such as the ultraconserved elements or anchored-tagged sequencing approaches (Faircloth *et al.* 2012; Lemmon *et al.* 2012; McCormack *et al.* 2013a). Alternatively, one can design probe sets from existing or newly generated genomic data (Bi *et al.* 2012; Hedtke *et al.* 2013; Li *et al.* 2013). The trade-off with these larger target sizes is that they would require more space per lane to get sufficient coverage, which could reduce the number of samples multiplexed per lane. Careful evaluation of the research question will determine the size and type of region a researcher would need to target.

In comparison with commercial target-enrichment methods, SCPP provides a lower per capture cost and greater per capture flexibility. Cost comparison between probe generation of the in-house SCPP method with commercial methods and a detailed cost analysis of SCPP can be seen in the Tables S7–S9 (Supporting information). Although the target set for commercial kits are flexible during the design phase, each capture in a particular kit is limited to the targets selected prior to synthesis. While this will not be a problem for researchers interested in capturing the same set of targets for all projects, SCPP allows each capture to be adjusted to any combination of loci based on the researcher's discretion and need for that particular experiment.

We believe that the SCPP method will prove particularly useful for many researchers working in the realms of phylogenetics and phylogeography because it allows for efficient generation of data sets that are sufficiently large to address research questions. Additionally, SCPP promises utility for previously problematic samples and markers, leaving far fewer holes in the data set than Sanger sequencing methods as well as being able to combine past Sanger data sets to SCPP data sets. By maximizing efficiency, minimizing cost and allowing customizable genomic targets of each discrete capture, this subgenomic enrichment method is another useful tool for bringing phylogenetic and population genetic studies of nonmodel organisms into the era of high-throughput genomic sequencing.

Acknowledgements

We would like to thank Sonal Singhal and Ke Bi for assistance with the bioinformatic pipeline and continued support and troubleshooting of the scripts. We would also like to thank Roberta Damasceno and Miguel Rodrigues for providing the samples for the *Enyalius* project and the Museum of Vertebrate Zoology for some of the vouchers used for the passerine project. We also gratefully acknowledge protocol and study design assistance obtained in correspondence with Martin Kircher, Tomislav Maricic, Victor Mason, Matthias Meyer, Matthew Morgan, William Murphy and Dan Vanderpool. Finally, we would like to thank the reviewers for helpful comments for this manuscript. This research was supported by grants from the National Science Foundation and the Museum of Vertebrate Zoology at UC Berkeley.

References

Amemiya CT, Alföldi J, Lee AP *et al.* (2013) The African coelacanth genome provides insights into tetrapod evolution. *Nature*, **496**, 311–316.

Arèvalo E, Davis SK, Sites JW (1994) Mitochondrial DNA sequence divergence and phylogenetic relationships among eight chromosome races of the Sceloporus grammicus complex (Phrynosomatidae) in central Mexico. *Systematic Biology*, **43**, 387–418.

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.

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, USA*, **101**, 11040–11045.

Bi K, Vanderpool D, Singhal S *et al.* (2012) Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, **13**, 403.

Brito PH, Edwards SV (2009) Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica*, **135**, 439–455.

Carling MD, Brumfield RT (2007) Gene sampling strategies for multilocus population estimates of genetic diversity (theta). *PLoS ONE*, **2**, e160.

Chan Y-C, Roos C, Inoue-Murayama M *et al.* (2010) Mitochondrial genome sequences effectively reveal the phylogeny of Hylobates gibbons. *PLoS ONE*, **5**, e14419.

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

Ellegren H, Smeds L, Burri R *et al.* (2012) The genomic landscape of species divergence in Ficedula flycatchers. *Nature*, **491**, 756–760.

Faircloth BC, McCormack JE, Crawford NG *et al.* (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, **61**, 717–726.

Gardner MG, Fitch AJ, Bertozzi T, Lowe AJ (2011) Rise of the machines–recommendations for ecologists when using next generation sequencing for microsatellite development. *Molecular Ecology Resources*, **11**, 1093–1101.

Good JM (2011) Reduced representation methods for subgenomic enrichment and next-generation sequencing (V Orgogozo, M V. Rockman, Eds,). *Methods in Molecular Biology*, **772**, 85–103.

Hancock-Hanser BL, Frey A, Leslie MS *et al.* (2013) Targeted multiplex next-generation sequencing: advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics. *Molecular Ecology Resources*, **13**, 254–268.

Hedtke SM, Morgan MJ, Cannatella DC, Hillis DM (2013) Targeted enrichment: maximizing orthologous gene comparisons across deep evolutionary time. *PLoS ONE*, **8**, e67908.

Hodges E, Xuan Z, Balija V *et al.* (2007) Genome-wide in situ exon capture for selective resequencing. *Nature Genetics*, **39**, 1522–1527.

Hohenlohe PA, Bassham S, Etter PD, *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Horn S (2012) Target enrichment via DNA hybridization capture. *Methods in Molecular Biology*, **840**, 177–188.

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. *Genome Research*, **9**, 868–877.

Kent WJ (2002) BLAT—the blast-like alignment tool. *Genome Research*, **12**, 656–664.

Kimball RT, Braun EL, Barker FK *et al.* (2009) A well-tested set of primers to amplify regions spread across the avian genome. *Molecular Phylogenetics and Evolution*, **50**, 654–660.

Knowles LL (2010) Sampling strategies for species tree estimation. In: *Estimating Species Trees: Practical and Theoretical Aspects* (ed. Knowles LL, Kubatko LS), pp. 163–173 (Chapter 10). John Wiley and Sons, Hoboken, New Jersey.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

Lanier HC, Huang H, Knowles LL (2014) How low can you go? The effects of mutation rate on the accuracy of species-tree estimation. *Molecular Phylogenetics and Evolution*, **70**, 112–119.

Lemmon AR, Lemmon EM (2012) High-throughput identification of informative nuclear loci for shallow-scale phylogenetics and phylogeography. *Systematic Biology*, **61**, 745–761.

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

Lemmon AR, Emme SA, Lemmon EM (2012) Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, **61**, 727–744.

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, **22**, 1658–1659.

Li C, Hofreiter M, Straube N, Corrigan S, Naylor GJP (2013) Capturing protein-coding genes across highly divergent species. *BioTechniques*, **54**, 321–326.

Liu B, Yuan J, Yiu S-M *et al.* (2012) COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. *Bioinformatics (Oxford, England)*, **28**, 2870–2874.

Lohse M, Bolger AM, Nagel A *et al.* (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, **40**, W622–W627.

Magoč T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics (Oxford, England)*, **27**, 2957–2963.

Mamanova L, Coffey AJ, Scott CE *et al.* (2010) Target-enrichment strategies for next- generation sequencing. *Nature Methods*, **7**, 111–118.

Maricic T, Whitten M, Pääbo S (2010) Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS ONE*, **5**, e14004.

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, **17**, 10.

Mason VC, Li G, Helgen KM, Murphy WJ (2011) Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. *Genome Research*, **21**, 1695–1704.

McCormack JE, Harvey MG, Faircloth BC *et al.* (2013a) A phylogeny of birds based on over 1,500 Loci collected by target enrichment and high-throughput sequencing. *PLoS ONE*, **8**, e54848.

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

Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, **2010**, 1–10.

Miller SA, Dykes DD, Polesky HF (1988) A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Research*, **16**, 1215.

Noonan JP, Coop G, Kudaravalli S *et al.* (2006) Sequencing and analysis of Neanderthal genomic DNA. *Science*, **314**, 1113–1118.

O'Neill EM, Schwartz R, Bullock CT *et al.* (2013) Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (Ambystoma tigrinum) species complex. *Molecular Ecology*, **22**, 111–129.

Palumbi SR (1996) Nucleic acids II: the polymerase chain reaction. *Molecular Systematics*, **2**, 205–247.

Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics*, **11**, 180.

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.

Portik DM, Wood PL Jr, Grismer JL *et al.* (2012) Identification of 104 rapidly-evolving nuclear protein-coding markers for amplification across scaled reptiles using genomic resources. *Conservation Genetics Resources*, **4**, 1–10.

Rannala B, Yang Z (2013) Improved reversible jump algorithms for Bayesian species delimitation. *Genetics*, **194**, 245–253.

Sánchez CC, Smith TPL, Wiedmann RT *et al.* (2009) Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics*, **10**, 559.

Scally A, Dutheil JY, Hillier LW *et al.* (2012) Insights into hominid evolution from the gorilla genome sequence. *Nature*, **483**, 169–175.

Simpson JT, Wong K, Jackman SD *et al.* (2009) ABySS: a parallel assembler for short read sequence data. *Genome Research*, **19**, 1117–1123.

Singhal S (2013) De novo transcriptomic analyses for non-model organisms: an evaluation of methods across a multi-species data set. *Molecular Ecology Resources*, **13**, 403–416.

Smith SA, Wilson NG, Goetz FE *et al.* (2011) Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*, **480**, 364–367.

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.

Wang N, Braun EL, Kimball RT (2012) Testing hypotheses about the sister group of the passeriformes using an independent 30-locus data set. *Molecular Biology and Evolution*, **29**, 737–750.

J.V.P. performed the research, analysed the data and wrote the manuscript; L.L.S. designed and performed the research; M.A.T. designed and performed the research and analysed the data; C.S. performed the research and wrote scripts to analyse the data; S.M.H. performed the research and analysed the data; P.L.S. performed the research and conducted the phylogenetic analysis; J.A.M. assisted with *Draco* study design; R.C.K.B. assisted with Passeriformes study design; and C.M. designed the research.

Data Accessibility

SRA Accession SRP039333 contains the raw reads for each library. Bioinformatic Pipeline: https://github.com/MVZSEQ/SCPP. DRYAD entry doi:10.5061/dryad.rk204 includes

1 *_CleanReads.zip – contains the cleaned reads and information on the filtered reads for each sample.
2 AbyssAssemblies.zip – contains the contig output of the ABySS assemblies.
3 FinalAssemblies.zip – contains the final assemblies after pooling of the ABySS assemblies, clustering and long-read assembly.
4 References.zip – contains the recovered references from the final assemblies for each locus for each sample.
5 *_Alignments.zip – contains novoalign alignments for each sample.
6 Haplotypes.zip – contains the haplotypes for each sample.

Supporting Information

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

Figure S1 Comparison of coverage and sequence divergence.

Figure S2 Comparison of coverage and probe length.

Figure S3 Mitochondrial Capture Specificity.

Table S1 Locus information for passerine project.

Table S2 Locus information for *Draco* project.

Table S3 Average coverage per locus for the passerine project (with flanking regions).

Table S4 Average coverage per locus for the *Enyalius* project (with flanking regions).

Table S5 Average coverage per locus for the *Draco* project (with flanking regions).

Table S6 Average coverage per locus for the passerine project (without flanking regions).

Table S7 Cost comparisons of SCPP and commercial probe kits.

Table S8 SCPP Reagent Costs – Probe Generation & Hybridization.

Table S9 SCPP Reagent Costs – Oligos.

Table S10 Capture Assessment – qPCR.

Appendix S1 Bioinformatics of read processing and phylogenetic methods for passerine tree.

Appendix S2 Detailed protocol for SCPP.