**Supplementary material and methods for:**

Complete genomes of two extinct New Zealand avian passerines show sensitivity to climate fluctuations but no evidence for genomic erosion prior to extinction

Nicolas Dussex1,2, Johanna von Seth1,3, Michael Knapp2, Olga Kardyalski2, Bruce C. Robertson4, Love Dalén1

1 Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Box 50007, Stockholm 10405, Sweden

2 Department of Anatomy, University of Otago, PO Box 913, Dunedin 9016, New Zealand

3 Department of Zoology, Stockholm University, Stockholm 10691, Sweden

4 Department of Zoology, University of Otago, PO Box 56, Dunedin 9016, New Zealand

**Materials and Methods**

*Sample collection, DNA extraction, library preparation and sequencing*

We extracted DNA from historical toepads for one Huia (*Heteralocha acutirostris*) and one South Island kōkako (*Callaeas cinereus*) collected in 1886 and 1849, respectively (Table S1) using a DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany). These samples were selected based on high endogenous DNA content (i.e. 75.9 and 85.9%) which was estimated with a screening sequencing step of four specimen of each species and where sequencing reads were mapped to the North Island kōkako genome (see below).

We then built double stranded Illumina libraries according to Meyer & Kircher [1]. We used 20 μl of DNA extract in a 40 μl blunt-end repair reaction with the following final concentration: 1× buffer Tango, 100 μM of each dNTP, 1 mM ATP, 25 U T4 polynucleotide kinase (Thermo Scientific) and 3U USER enzyme (New England Biolabs). Treatment with USER enzyme was performed to excise uracil residues resulting from post-mortem damage [2,3]. Samples were incubated for 3 h at 37°C, followed by the addition of 1 μl T4 DNA polymerase (Thermo Scientific) and incubation at 25°C for 15 min and 12°C for 5 min. The samples were then cleaned using MinElute spin columns following the manufacturer's protocol and eluted in 20ul EB Buffer. Next, we performed an adapter ligation step where DNA fragments within each library were ligated to a combination of incomplete, partially double-stranded P5- and P7-adapters (10 μM each). This reaction was performed in a 40 μl reaction volume using 20 μl of blunted DNA from the clean-up step and 1 μl P5-P7 adapter mix per sample with a final concentration of 1× T4 DNA ligase buffer, 5% PEG-4000, 5U T4 DNA ligase (Thermo Scientific). Samples were incubated for 30 minutes at room temperature and cleaned using MinElute spin columns as described above. Next, we performed an adapter fill-in reaction in 40 μl final volume using 20 μl adapter ligated DNA with a final concentration of 1× Thermopol Reaction Buffer, 250 μM of each dNTP, 8U *Bst* Polymerase, Long Fragments. The libraries were incubated at 37°C for 20 minutes, heat-inactivated at 80°C for 20 minutes. These libraries were then used as stock for indexing PCR amplification.

In order to increase library complexity six indexing PCR amplification were performed for each library using six different P7 indexing primers [1]. Amplifications were performed in 25 μl volumes with 3μl of adapter-ligated library as template, with the following final concentrations: 1x AccuPrime reaction mix, 0.3μM IS4 amplification primer, 0.3μM P7 indexing primer, 7 U AccuPrime Pfx (Thermo Scientific) and the following cycling protocol: 95°C for 2 min, 12 cycles at 95°C for 30 s, 55°C for 30 s and 72°C for 1 min and a final extension at 72°C for 5 minutes.

Purification and size selection of libraries was then performed using Agencourt AMPure XP beads (Beckman Coulter, Brea, CA, USA), first using 0.5X bead:DNA ratio and secondly 1.8X to remove long and short (i.e. adapter dimers) fragments, respectively. Library concentration was measured with a high-sensitivity DNAchip on a Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA). Finally, multiplexed libraries (i.e. 6 indexed libraries) were pooled into a single pool in equimolar concentrations and sequenced each of them on an Illumina HiSeqX platform with a 2 × 150 bp setup in the High Output mode at the SciLifeLab sequencing facility in Stockholm.

Extractions and library preparation were conducted in a separate state-of-the-art historical DNA lab and appropriate precautions were taken to minimize the risk of contamination in historical samples [4].

*Bioinformatics*

*Reference genome*: I order to map genomic data for the two historical specimen, we used a *de novo* genome for the North Island kōkako (*Callaeas wilsoni;* https://b10k.genomics.cn/) consisting of 154,271 scaffolds with an N50 size of 303,331 kb, an average scaffold length of 6,893 Mb and a total scaffold length of 1.06 Gb. We identified Z-linked scaffolds and a mtDNA scaffold from the assembled genome by blasting all scaffolds against the Z-chromosomes of zebra finch v. 3.2.4 (*Taeniopygia guttata*; GenBank: GCA\_000151805.2) and the mitogenome of *C. cinereus* (GenBank: KU158191.1) using BLAST+ 2.5.0 [5]. The BLAST+ parameters were set as: -evalue=1e-10; -word\_size=15; -max\_target\_seqs=10. We excluded sex chromosome-linked scaffolds (n=10; total 1.9Mb) with 1000bp and 80% identity with the zebra finch Z-chromosomes from the assembled genome for all downstream analyses. Finally, we identified repetitive elements in the genome assembly using RepeatMasker (http://repeatmasker.org) [6] applying the repeat element library of the aves database. We also identified CpG sites using a custom script.

*Mapping of historical data:* Raw historical sequence data was demultiplexed using bcl2Fastq v2.17.1 with default settings (Illumina Inc.). SeqPrep 1.1 [7] was then used to trim adapters and merge paired-end reads, using default settings but with a minor modification in the source code, allowing us to choose the best quality scores of bases in the merged region instead of aggregating the scores following Palkopoulou *et al.* [8]. As recommended for historical and ancient DNA short reads, we merged sequencing reads and mapped them against the reference genome using the BWA 0.7.13 aln algorithm [9] and slightly modified default settings with deactivated seeding (-l 16,500), allowing more substitutions (-n 0.01) and allowing up to two gaps (-o 2). The BWA samse command was then used to generate alignments in SAM format. The resulting reads were then processed in SAMtools and converted to BAM format, as well as coordinate sorted and indexed. We removed duplicates from the alignments using a custom python script to avoid inflation of length distribution for loci with deep coverage [8]. Next, we used Picard v. 1.141 (http://broadinstitute.github.io/picard) to assign read group information including library, lane and sample identity to each bam file. Reads were then re-aligned around indels using GATK 3.4.0 [10]. Only reads/alignments with mapping quality 30 were kept for subsequent analysis.

We then masked repeats and CpG sites ﻿to limit possible biases from DNA damage using the bed files created above and BEDtools [11]. Finally, we estimated the depth of genome coverage for each sample was estimated using Qualimap 2.2.1 [12]. Average genome coverage was of 10 (min=3, max=20) and 14 (min=5, max=29) for Huia and South Island kōkako, respectively (Fig. S1, Table S1). Finally, we estimated the proportion of missing data as the proportion of sites with a root-mean-squared mapping quality of reads covering the site below 30. Missing data was estimated at 5.5% and 4.1% for Huia and South Island kōkako, respectively.

*Variant calling*:We called variants in historical huia South Island kōkako genomes as well as in the North Island kōkako using bcftools mpileup (v. 1.3)  [13] and bcftools (v. 1.3) using a minimum depth of coverage of 1/3X of the average coverage, base quality 30 and removed SNPs within 5bp of indels. As for bam files, we masked CpG sites and repeats using BEDtools [11]. For all downstream analyses, excluded the 10 sex chromosome-linked scaffolds (see above). Overall, we obtained 20,067,979 SNPs. We then filtered this dataset for HWE and removed SNPs no scored in both samples (--geno 0) and had 16,879,201 SNPs left. Raw fastq reads are deposited at the European Nucleotide Archive (ENA), accession number (pending).

*Data analysis*

We first used the Pairwise Sequentially Markovian Coalescent (PSMC 0.6.5) [14] model to infer changes in the effective population sizes (*N*e) of huia and South Island kōkako over time. This approach infers the distribution of the time to the most recent common ancestor (TMRCA) between the two alleles across all chromosomes using the density of heterozygous sites across the diploid genome of a single individual. Regions of low heterozygosity reflect recent coalescent events while regions of high heterozygosity reflect more ancient coalescent events. The rate of coalescent events in each segment is then informative about changes in *N*e through time since the rate of coalescence is inversely proportional to *N*e. We generated consensus sequences for all autosomes of historical and a subset of the modern genomes using SAMtools mpileup (v. 1.3) [13] command and the ‘vcf2fq’ command from vcfutils.pl. We used filters for base quality, mapping quality and root-mean-squared mapping quality below 30, and depth below 1/3 and higher than 2-times the average coverage estimated for each specimen. In order to infer the distribution of the time to the most recent common ancestor (TMRCA) between the two copies of each chromosome from each individual across all autosomes, we set N (the number of iterations)=25, t (Tmax)=15 and p (atomic time interval)=64 (4+25\*2+4+6,for each of which parameters are estimated with 28 free interval parameters). We scaled population parameters assuming a generation time of 6 years [15,16] and a substitution rate of 1.38×10−8 substitutions/site/generation, inferred from a rate of 2.3×10−9 substitutions/site/year as estimated form bird pedigree data [17]. As a comparison and in order to account for variation in substitution rates among passerines, we also used a rate of 1.92×10−8 substitutions/site/generation inferred from a passerine-specific rate of 3.2×10−9 substitutions/site/year from Zhang *et al.* [18] (Fig. S2).

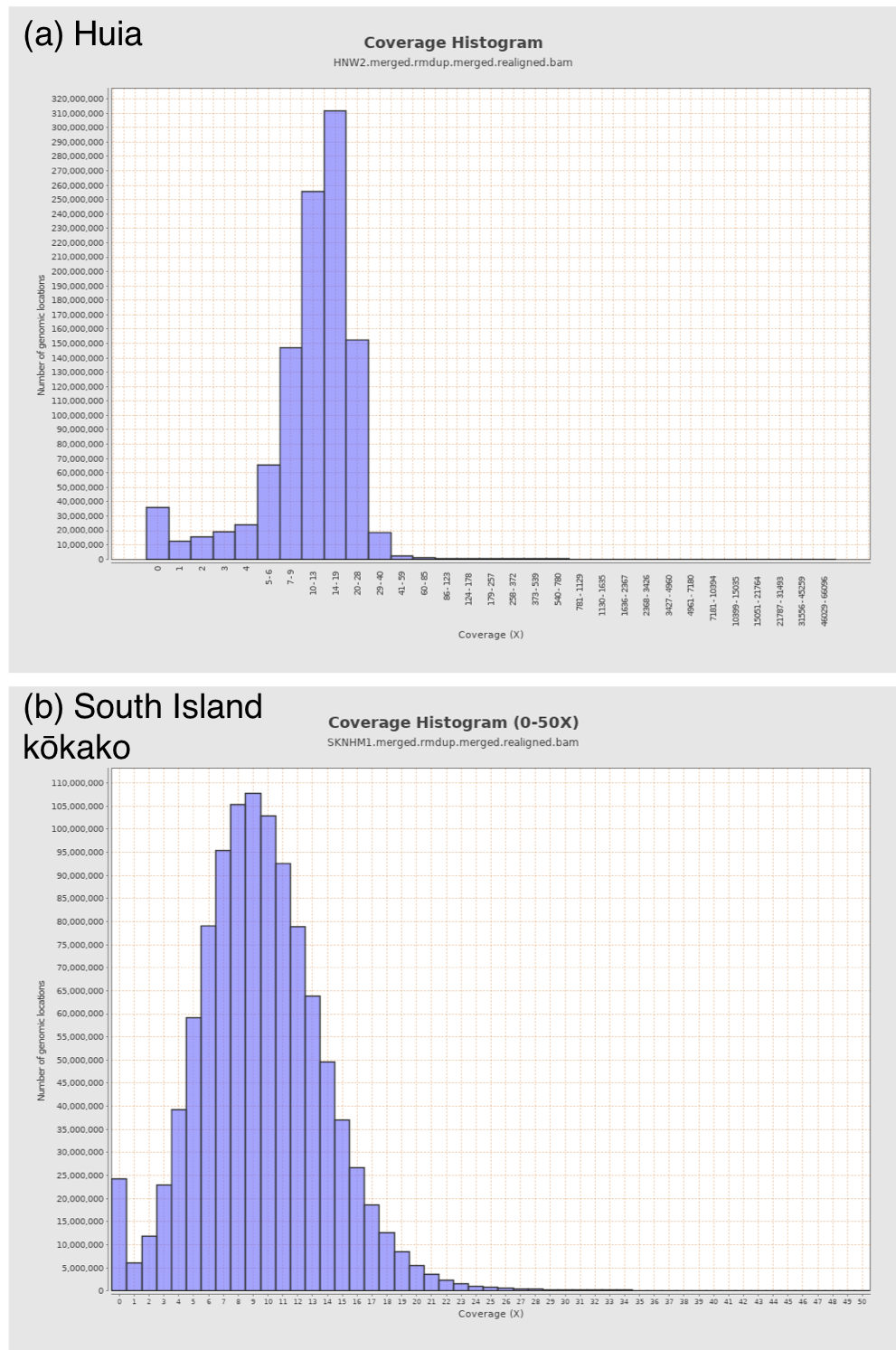
Because ﻿PSMC has high false-negative rates at low coverage genomes (< 18X), which can lead to biases in the estimate of timing and magnitude of population events [19], we performed a uniform False Negative Rate (uFNR) correction to the PSMC for the two genomes by using the option -M of the script and using uFNR of 0.2 and 0.4 as described in the PSMC manual. ﻿While the PSMC without uFNR correction showed a reduced timing and lower *N*e estimates for the two species, the overall the demographic trajectories were similar across tests with or without correction (Fig. 2, S3).

Secondly, we estimatedoverall autosomal heterozygosity. We used mlRho v.2.7 [20] to estimate the population mutation rate (θ), which approximates expected heterozygosity under the infinite sites model. This maximum likelihood approach provides unbiased estimates of average within-individual heterozygosity at high coverage [20,21]. We filtered out bases with quality below 30, reads with mapping quality below 30 and positions with root-mean-squared mapping quality below 30. We also filtered out sites with depth lower than 1/3X and higher than 2X the average coverage for each specimen to avoid false heterozygous calls due to variation in coverage across the genome [8].

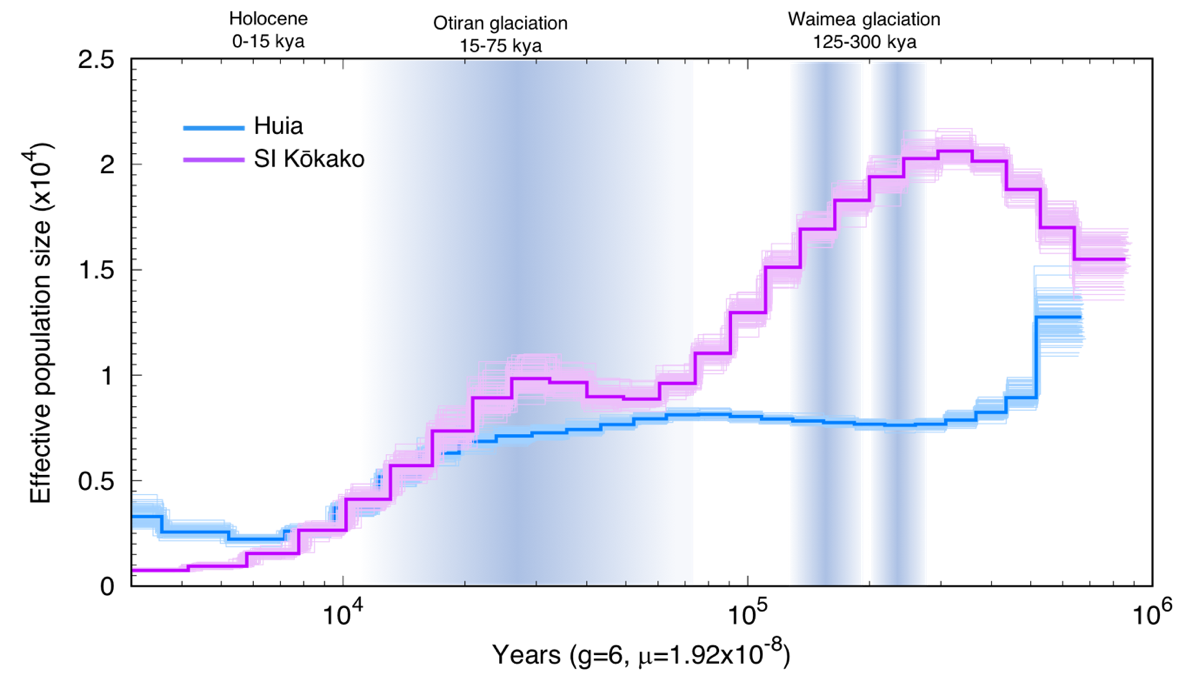
Third, we estimated inbreeding by identifying runs of homozygosity (ROH) and individual inbreeding coefficients (FROH) using the sliding-window approach implemented in PLINK [22]. We converted the filtered (i.e. 16,879,201 SNPs) multi-individual vcf file comprising the two historical genomes into a ped file and identified ROHs in the 31 largest (> 1Mb) autosomal scaffolds, representing 88.6% of the total genome. We used a sliding window size of 100 SNPs (--*homozyg-window-snp* *100*). A window was then defined as homozygous if there were not more than 5 missing sites (*homozyg-window-missing 5*) and not more than 1 heterozygous site per window (*homozyg-window-het 1*). If at least 5% of all windows that included a given SNP were defined as homozygous, the SNP was defined as being in a homozygous segment of a chromosome (*homozyg-window-threshold 0.05*). This threshold was chosen to ensure that the edges of a ROH are properly delimited. A homozygous segment was defined as a ROH if all of the following conditions were met: the segment included ≥25 SNPs (*homozyg-snp 25*) and covered ≥100 kb (*homozyg-kb* 100). Furthermore, the minimum SNP density was one SNP per 50 kb (*homozyg-density 50*) and the maximum distance between two neighbouring SNPs was ≤1,000 kb (*homozyg-gap 1,000*). For the number of heterozygous sites within ROHs, we set the value at 750 (*homozyg-het 750*) in order to prevent sequencing errors to cut ROHs.

**Table S1**. Specimen origin with reads and coverage statistics



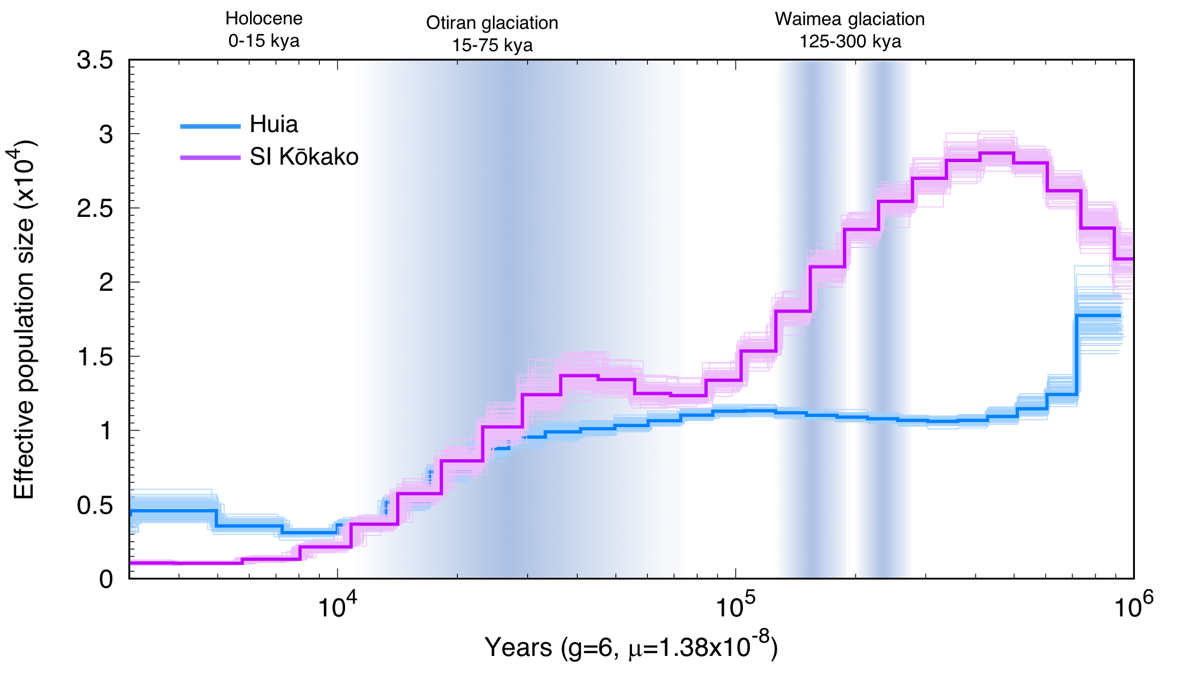


**Figure S1.** Distribution of coverage across the genome for (a) Huia (*Heteralocha acutirostris*) and (b) South Island kōkako (*Callaeas cinereus*).

****

**(b)**

**(a)**



**Figure S2.** PSMC for huia and South Island kōkako assuming a generation time of 6 years [15,16] without ﻿Uniform False Negative Rate (uFNR) correction. The x axis corresponds to time before present in years on a log scale, assuming a substitution rate of **(a)** 1.9210-8 substitution/site/generation inferred from a passerine-specific rate [18] and, **(b)** assuming a substitution rate of 1.3810-8 substitution/site/generation inferred from [17]. Thin lines depict one hundred bootstrap replicates for each individual bird of the same colour. The y axis corresponds to the effective population size.

**(b)**

****

**(a)**

**(b)**

**(c)**

**(d)**

****

**Figure S3.** PSMC for (a,b) huia and (c,d) South Island kōkako showing the comparison of uniform False Negative Rate (uFNR) correction, two different substitution rates and assuming a generation time of 6 years [15,16]. Panels (a) and (c) show the PSMC based on a substitution rate of 1.92 10-8 substitution/site/generation inferred from a passerine-specific rate [18]. Panels (b) and (d) show the PSMC based on a substitution rate of 1.38 10-8 substitution/site/generation based on pedigree estimates inferred from [17]. For each figure, the red, green and blue lines depict no correction, a 20% and a 40% uFNR correction, respectively.

**References**

1. Meyer M, Kircher M. 2010 Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **5**. (doi:10.1101/pdb.prot5448)

2. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S. 2010 Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.* **38**, 1–12. (doi:10.1093/nar/gkp1163)

3. Kircher M, Sawyer S, Meyer M. 2012 Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* **40**. (doi:10.1093/nar/gkr771)

4. Knapp M, Clarke AC, Horsburgh KA, Matisoo-Smith EA. 2012 Setting the stage - Building and working in an ancient DNA laboratory. *Ann. Anat.* **194**, 3–6. (doi:10.1016/j.aanat.2011.03.008)

5. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009 BLAST + : architecture and applications. *BMC Bioinformatics* **10**, 421. (doi:10.1186/1471-2105-10-421)

6. Smit AFA, Hubley R, Green P. 2010 RepeatMasker Open-3.0. 1996-2010. *Inst. Syst. Biol.*

7. John JS. 2011 SeqPrep 1.1. *https://github.com/jstjohn/ SeqPrep.*

8. Palkopoulou E *et al.* 2015 Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Curr. Biol.* **25**, 1395–1400. (doi:10.1016/j.cub.2015.04.007)

9. Li H, Durbin R. 2010 Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595. (doi:10.1093/bioinformatics/btp698)

10. McKenna A *et al.* 2010 The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303. (doi:10.1101/gr.107524.110.20)

11. Quinlan AR, Hall IM. 2010 BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842. (doi:10.1093/bioinformatics/btq033)

12. Okonechnikov K, Conesa A, García-Alcalde F. 2015 Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294. (doi:10.1093/bioinformatics/btv566)

13. 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**, 2078–2079. (doi:10.1093/bioinformatics/btp352)

14. Li H, Durbin R. 2011 Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496. (doi:10.1038/nature10231)

15. Double M, Murphy S. 2000 Genetic variation within and among populations of North Island kokako. *New Zeal. Dept Conserv.*

16. Armstrong DP, Davidson RS, Perrott JK, Roygard J, Buchanan L. 2005 Density-dependent population growth in a reintroduced population of North Island saddlebacks. *J. Anim. Ecol.* **74**, 160–170. (doi:10.1111/j.1365-2656.2004.00908.x)

17. Smeds L, Qvarnström A, Ellegren H. 2016 Direct estimate of the rate of germline mutation in a bird. *Genome Res.* **26**, 1211–1218. (doi:10.1101/gr.204669.116)

18. Zhang G *et al.* 2014 Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1321.

19. Nadachowska-Brzyska K, Burri R, Smeds L, Ellegren H. 2016 PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. *Mol. Ecol.* **25**, 1058–1072. (doi:10.1111/mec.13540)

20. Haubold B, Pfaffelhuber P, Lynch M. 2010 MlRho - A program for estimating the population mutation and recombination rates from shotgunsequenced diploid genomes. *Mol. Ecol.* **19**, 277–284. (doi:10.1111/j.1365-294X.2009.04482.x)

21. Lynch M. 2008 Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. *Mol. Biol. Evol.* **25**, 2409–2419. (doi:10.1093/molbev/msn185)

22. Purcell S *et al.* 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575. (doi:10.1086/519795)