Barred owl (*Strix varia*) population genomic summary statistics

Zachary R. Hanna1, Jeffrey D. Wall1

1Institute for Human Genetics, University of California, San Francisco, San Francisco, California, United States of America

# Methods

We utilized high-coverage genomic sequence data from twelve eastern and thirteen western *Strix varia* as well as from four spotted owls (*S. occidentalis*) [1,2]. We downloaded the raw genomic sequences from the NCBI Sequence Read Archive (SRA) accessions SRR4011595-SRR4011597, SRR4011614-SRR4011617, SRR6048826-SRR6048853 (TableSampInfo). We used Trimmomatic version 0.36 [3] to remove adapter sequences as well as low quality leading and trailing bases. We used BWA-MEM version 0.7.17-r1188 [4] to align the processed sequences to our new reference *Strix occidentalis caurina* nuclear genome assembly together with the Hanna et al. [5] *S. o. caurina* mitochondrial genome sequence. We merged the alignments, sorted the alignments, and marked duplicate sequences with Picard version 2.17.6 (http://broadinstitute.github.io/picard).

## Variant calling and filtering

We called variants for each sample using the Genome Analysis Toolkit (GATK) version 3.8.0 HaplotypeCaller tool and then combined the variants across samples using the GATK GenotypeGVCFs tool [6–9]. Using the GATK VariantFiltration and SelectVariants tools, we implemented a hard filtering of the variants to generate a set of high quality single nucleotide polymorphisms (SNPs) and indels by following an adapted version of the GATK guidelines for hard filtering of variants. We utilized these filtered sets SNPs and indels for base quality score recalibration (BQSR). We first recalibrated the alignment files for each sample using the GATK BaseRecalibrator and PrintReads tools and then performed a new round of individual sample variant calling with the GATK HaplotypeCaller tool followed by combining variants across samples with the GATK GenotypeGVCFs tool. We extracted SNPs from the resulting variant call format (VCF) file using the GATK SelectVariants tool.

We used GNU AWK (GAWK) version 4.2.0 [10] to exclude SNPs that fell on the mitochondrial genome and then removed SNPs in repetitive or low complexity regions using BEDTools version 2.25.0 [11] along with the file of repetitive regions that we generated following repeat annotation of the new nuclear genome. We excluded low quality SNPs using the GATK VariantFiltration tool and extracted only biallelic SNPs by using the GATK SelectVariants tool to omit multiallelic sites. We calculated the mean and standard deviation of the site-level coverage of the remaining SNPs and then removed sites with excessive coverage with the vcf\_filter\_highDP.sh script from genetics-tools version 1.0.1 [12]. We determined the mean and standard deviation of the sequence coverage depth for each individual across the final set of SNPs using the DP\_sample\_calc.sh script from genetics-tools version 1.0.1 [12]. For use in downstream analyses, we compressed the final set of SNPs using the bzip tool from HTSlib version 1.8 [13] and then indexed the file using the Tabix tool from HTSlib version 1.8 [13,14].

# References

1. Hanna ZR, Henderson JB, Wall JD, Emerling CA, Fuchs J, Runckel C, et al. Northern Spotted Owl (*Strix occidentalis caurina*) Genome: Divergence with the Barred Owl (*Strix varia*) and Characterization of Light-Associated Genes. Genome Biol Evol. 2017;9: 2522–2545. doi:10.1093/gbe/evx158

2. Hanna ZR, Dumbacher JP, Bowie RCK, Wall JD. Whole-genome sequences suggest long term declines of spotted owl (*Strix occidentalis*) (Aves: Strigiformes: Strigidae) populations in California. bioRxiv. 2018; 343434. doi:10.1101/343434

3. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120. doi:10.1093/bioinformatics/btu170

4. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv:1303.3997 Q-Bio. [Accessed 2016 Feb 16]. 2013; Available: http://arxiv.org/abs/1303.3997

5. Hanna ZR, Henderson JB, Sellas AB, Fuchs J, Bowie RCK, Dumbacher JP. Complete mitochondrial genome sequences of the northern spotted owl (*Strix occidentalis caurina*) and the barred owl (*Strix varia*; Aves: Strigiformes: Strigidae) confirm the presence of a duplicated control region. PeerJ. 2017;5: e3901. doi:10.7717/peerj.3901

6. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43: 491–498. doi:10.1038/ng.806

7. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–1303. doi:10.1101/gr.107524.110

8. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinforma. 2013;11: 11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43

9. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Auwera GAV der, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2017; 201178. doi:10.1101/201178

10. Free Software Foundation. GNU Awk [Internet]. 2017. Available: https://www.gnu.org/software/gawk

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

12. Hanna ZR. genetics-tools. Version 1.0.1. Zenodo. 2018; doi:10.5281/zenodo.1257508

13. Davies R, Randall JC, McCarthy SA, Bonfield J, Pollard MO, Marshall J, et al. HTSlib [Internet]. 2018. Available: https://github.com/samtools/htslib

14. Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics. 2011;27: 718–719. doi:10.1093/bioinformatics/btq671