File S1
for:

Whole-genome analysis of introgression between the spotted owl and barred owl (*Strix occidentalis* and *Strix varia*, respectively; Aves: Strigidae) in western North America

Zachary R. Hanna[*,†,‡,§,**], John P. Dumbacher[§,**], Rauri C.K. Bowie[†,‡], James B. Henderson[**], Jeffrey D. Wall[*,**]

[*]Institute for Human Genetics, University of California San Francisco, San Francisco, California, United States of America.
[†]Museum of Vertebrate Zoology, University of California, Berkeley, Berkeley, California, United States of America.
[‡]Department of Integrative Biology, University of California, Berkeley, Berkeley, California, United States of America.
[§]Department of Ornithology & Mammalogy, California Academy of Sciences, San Francisco, California, United States of America.
[**]Center for Comparative Genomics, California Academy of Sciences, San Francisco, California, United States of America.
Corresponding author: Zachary.Hanna@ucsf.edu (ZRH)

1. Supplementary details of materials and methods

1.1. Mapping

1.1.1. We created a map of the samples in QGIS version 2.18.2 (Quantum GIS Development Team 2017) using the high resolution (21,600 x 10,800 pixels), 1:10 million-scale Gray Earth with Shaded Relief, Hypsography, Ocean Bottom, and Drainages version 2.1.0 raster file from Natural Earth (http://www.naturalearthdata.com; accessed 2017 Oct 1) as the base map layer. We overlaid this with the 1:50 million-scale Admin 1 - States, Provinces boundaries version 3.0.0 vector file from Natural Earth and then plotted the coordinates of our samples.

1.2. Sequence data

1.2.1. We prepared libraries and sequenced the samples in two indexed pools, which we will refer to as "Sample Set 1" and "Sample Set 2" (see Table S1 for the samples included in each sample set).

1.2.2. **Sample Set 1 - *Strix varia* sample CAS:ORN:95964.** We extracted genomic DNA from CAS:ORN:95964 using a DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany). We used 50 ng genomic DNA to prepare a whole-genome library using a Nextera DNA Sample Preparation Kit (Illumina, San Diego, California). After tagmentation, we cleaned the reaction with a DNA Clean & Concentrator -5 kit (Zymo Research, Irvine, California). We amplified the reaction with five cycles of PCR using a KAPA Library Amplification kit (KAPA Biosystems, Wilmington, Massachusetts) and then cleaned the reaction with a DNA Clean & Concentrator -5 kit (Zymo Research, Irvine, California). We used Dye-Free, 1.5% agarose, 250-1,500 base pair (bp) cassette on a BluePippin (Sage Science, Beverly, Massachusetts) to select library fragments in the size range of 534-634 bp, which, after subtracting the 134 bp of adapters, corresponded to selecting an average insert size of 450 bp. We next performed a real-time PCR (rtPCR) using a KAPA Real-Time Library Amplification Kit (KAPA Biosystems, Wilmington, Massachusetts) on a CFX96 Touch Real-Time PCR Detection System (Bio-Rad, Hercules, California) to further amplify the library with nine cycles PCR. We then cleaned the PCR products with a DNA Clean & Concentrator -5 kit (Zymo Research, Irvine, California). We assessed the library fragment size distribution with a 2100 BioAnalyzer (Agilent Technologies, Santa Clara, California) and the concentration of double-stranded DNA material with a Qubit 2.0 Flurometer (Invitrogen, Carlsbad, California). Due to the presence of small peaks in the BioAnalyzer trace, we further cleaned the library using 0.6X Agencourt AMPure XP (Beckman Coulter, Brea, California) magnetic beads. We then reassessed the concentration of double-stranded DNA material with a Qubit 2.0 Flurometer (Invitrogen, Carlsbad, California) and the library fragment size distribution with a 2100 BioAnalyzer (Agilent Technologies, Santa Clara, California), which revealed that the average fragment size of the pool was 583 nucleotides (nt).

1.2.3. **Sample Set 1 - sixteen additional samples.** For each sample, we used 10 ng genomic DNA to prepare a whole-genome library using a Nextera DNA Sample Preparation Kit (Illumina, San Diego, California). After tagmentation, we cleaned the reaction with a DNA Clean & Concentrator -5 kit (Zymo Research, Irvine, California). We amplified the reaction and added Illumina indexed adapters with

five cycles of PCR using a KAPA Library Amplification kit (KAPA Biosystems, Wilmington, Massachusetts) and then cleaned the reaction with a DNA Clean & Concentrator -5 kit (Zymo Research, Irvine, California). We assessed the concentration of double-stranded DNA material with a Qubit 2.0 Flurometer (Invitrogen, Carlsbad, California), combined the library into an equimolar sixteen-sample pool, and then concentrated the pool with a DNA Clean & Concentrator -5 kit (Zymo Research, Irvine, California). We used a BluePippin (Sage Science, Beverly, Massachusetts) to select library fragments in the size range of 550-750 nt, which, after subtracting the 134 nt of adapters, corresponded to selecting an average insert size of 516 nt. We then performed a real-time PCR (rtPCR) using a KAPA Real-Time Library Amplification Kit (KAPA Biosystems, Wilmington, Massachusetts) on a CFX96 Touch Real-Time PCR Detection System (Bio-Rad, Hercules, California) to amplify the pool with nine cycles of PCR. We then cleaned the PCR products with a DNA Clean & Concentrator -5 kit (Zymo Research, Irvine, California). We assessed the concentration of double-stranded DNA material with a Qubit 2.0 Flurometer (Invitrogen, Carlsbad, California) and the library fragment size distribution with a 2100 BioAnalyzer (Agilent Technologies, Santa Clara, California), which revealed that the average fragment size of the pool was 607 nt.

1.2.4. **Sample Set 1 - final pool.** We pooled the CAS:ORN:95964 library in an equimolar ratio with the equimolar pool of the sixteen other libraries. We then sequenced the final pool across both lanes of a two-lane flow cell with a HiSeq PE Rapid Cluster Kit and a 200 cycle HiSeq Rapid SBS Kit v1 on a HiSeq 2500 (Illumina, San Diego, California). The raw sequences are available from the NCBI Sequence Read Archive (SRA) in the run accessions indicated in Table S1.

1.2.5. **Sample Set 2.** For each sample, we used 10 ng genomic DNA to prepare a whole-genome library using a Nextera DNA Sample Preparation Kit (Illumina, San Diego, California). After tagmentation, we cleaned the reaction with a DNA Clean & Concentrator -5 kit (Zymo Research, Irvine, California). We amplified the reaction and added Illumina indexed adapters with five cycles of PCR using a KAPA Library Amplification kit (KAPA Biosystems, Wilmington, Massachusetts) and then cleaned the reaction with a DNA Clean & Concentrator -5 kit (Zymo Research, Irvine, California). We assessed the concentration of double-stranded DNA material with a Qubit 2.0 Flurometer (Invitrogen, Carlsbad, California), combined the indexed library into a thirty-six-sample pool, and then concentrated the pool with a DNA Clean & Concentrator -5 kit (Zymo Research, Irvine, California). We used a BluePippin (Sage Science, Beverly, Massachusetts) to select fragments in the size range of 500-700 nt, which, after subtracting the 134 nt of adapters, corresponded to selecting an average insert size of 466 nt. We cleaned the BluePippin products with 0.6X Agencourt AMPure XP (Beckman Coulter, Brea, California) magnetic beads and then performed a real-time PCR (rtPCR) using a KAPA Real-Time Library Amplification Kit (KAPA Biosystems, Wilmington, Massachusetts) on a CFX96 Touch Real-Time PCR Detection System (Bio-Rad, Hercules, California) to amplify the pool with eight cycles of PCR. We then cleaned the PCR products with a DNA Clean & Concentrator -5 kit (Zymo Research, Irvine, California). We assessed the concentration of double-stranded DNA material with a Qubit 2.0 Flurometer (Invitrogen, Carlsbad, California) and the fragment size distribution with a 2100

BioAnalyzer (Agilent Technologies, Santa Clara, California), which indicated that the average fragment size of the pool was 579 nt. We sequenced the pool on two successive runs of 150 nt paired-end sequencing using a two-lane flow cell with a HiSeq PE Rapid Cluster Kit and a 300 cycle HiSeq Rapid SBS Kit v1 on a HiSeq 2500 (Illumina, San Diego, California) in rapid mode. We obtained sequencing data from each of the two flow cell lanes on the first run. We obtained data from a portion of one of the two flow cell lanes on the second run. The raw sequences from the Sample Set 2 samples are available from the NCBI SRA in the run accessions indicated in Table S1.

1.3. Sequence data processing

1.3.1. We performed adapter and quality trimming of the low-coverage sequence data using Trimmomatic version 0.32 (Bolger *et al*. 2014) with the following options: "ILLUMINACLIP:<fasta of Illumina adapter sequences>:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:28 MINLEN:36".

1.4. Alignment and filtering

1.4.1. For all samples, in order to align trimmed paired and unpaired reads to "StrOccCau_1.0_nuc_masked" (Hanna *et al*. 2017c, 2017b) we used bwa mem version 0.7.12-r1044 (Li 2013) with default options other than parameters "bwa mem -M". We separately aligned paired-end and unpaired reads. For alignment of the paired-end reads, we set the insert size to be equal to the size estimate of the final library given by the 2100 BioAnalyzer (Agilent Technologies, Santa Clara, California) minus the length of the adapters (insert sizes for CAS:ORN:98821 and CNHM<USA-OH>:ORNITH:B41533 obtained from Hanna et al. (2017c, 2017b), 446 nt insert size used for CAS:ORN:95964, 481 nt insert size for the rest of Sample Set 1, and 466 nt insert size for Sample Set 2). Additionally, for the alignment of the paired-end reads we set the parameter "-w", the maximum insert size, equal to 1000.

1.4.2. We used the Picard version 1.104 function MergeSamFiles (http://broadinstitute.github.io/picard) with default settings to merge the paired-end and unpaired sequence alignments and then used the Picard version 1.104 function SortSam (http://broadinstitute.github.io/picard) with default settings to sort the alignments. We used the Picard version 1.104 function MarkDuplicates (http://broadinstitute.github.io/picard) with default settings to mark duplicate sequences (both PCR and optical).

1.4.3. We used the Genome Analysis Toolkit (GATK) version 3.4-46 PrintReads tool (McKenna *et al*. 2010; DePristo *et al*. 2011; Van der Auwera *et al*. 2013) with the parameters "--read_filter BadCigar --read_filter BadMate --read_filter UnmappedRead --read_filter NotPrimaryAlignment --read_filter FailsVendorQualityCheck --read_filter DuplicateRead --read_filter MappingQualityUnavailable" to filter the bam files to only retain the high quality alignments. Alignments with these flags are ignored by the GATK SNP discovery tools, but since reads with these flags can still contribute to the DepthPerAlleleBySample (depth of coverage of each allele per sample) field in the variant call format files output by the GATK tools (GATK Dev Team 2017) and we

intended to use this field downstream for purposes where this extra coverage may be misleading, we performed this filtering of the bam files.

1.5. SNP calling

1.5.1. We used the GATK version 3.4-46 UnifiedGenotyper tool (McKenna *et al*. 2010; DePristo *et al*. 2011; Van der Auwera *et al*. 2013) to call SNPs using all of the filtered bam files as simultaneous inputs and employing default options other than setting "--output_mode EMIT_ALL_SITES".

1.5.2. There is a sample included in the variant call format (vcf) file output by UnifiedGenotyper for which we do not report any results. We initially included this sample that was provided by another research group as a potential hybrid to test. Our results suggested that the sample was not what it was originally purported to be. It was clear that it was a *Strix occidentalis* sample, but, as we did not know the geographic origin or date of collection of the sample, we decided to drop it from our analyses. We provide this information as explanation for its presence in the vcf file.

1.6. Filtering and spotted owl ancestry analyses

1.6.1. We used vcf_qual_filter.sh from SPOW-BDOW-introgression-scripts version 1.1.1 (Hanna *et al*. 2017a) to retain only biallelic sites sites where CAS:ORN:98821 (the source of the StrOccCau_1.0_nuc_masked reference genome) was homozygous for the reference allele and CNHMB41533, the *S. varia* reference sample, was homozygous for the alternative allele. We used this script also to exclude indels. Additionally, we only retained sites where the Phred-scaled probability that a polymorphism exists was >50 and the Phred-scaled genotype quality was >=30 for both CAS:ORN:98821 and CNHMB41533. We required that CNHMB41533 had zero reads that supported the CAS:ORN:98821 allele at each retained variant site. We also required that CAS:ORN:98821 had zero reads that supported the CNHMB41533 allele and >=10 reads in support of the CAS:ORN:98821 allele at each retained variant site.

1.6.2. We used dp_cov_script.sh from SPOW-BDOW-introgression-scripts version 1.1.1 (Hanna *et al*. 2017a) to calculate the mean and standard deviation of the total coverage at the remaining sites. We then used vcf_dp_filter.sh from SPOW-BDOW-introgression-scripts version 1.1.1 to remove those with coverage in excess of the mean + 5σ (we only kept sites with coverage <301 X).

1.6.3. We used AD_pct.sh from SPOW-BDOW-introgression-scripts version 1.1.1 to calculate the spotted owl ancestry for each sample at each of the final variant sites. We then used compute_ad_mean_stdev.sh from SPOW-BDOW-introgression-scripts version 1.1.1 to calculate the mean and standard deviation (σ) of the spotted owl ancestry across all variant sites.

1.6.4. We tested for significant difference in the spotted owl ancestry of the individuals in four sets of populations using Welch's *t*-test (Welch 1947) and applied a Bonferroni adjustment (Dunn 1961) to the p-value cut-off to correct for multiple comparisons. We performed Welch's *t*-test using the Welch_ttest.py script from SPOW-BDOW-introgression-scripts version 1.1.1. We first tested for significant difference in the spotted owl ancestry of the Siskiyou barred owls versus the rest of the western barred owls. Based on the result of this test, we then grouped the Siskiyou barred owls with the other western barred owls into a population including all western barred owls. Similarly, we tested for significant difference between the

pre and post-contact spotted owl samples before grouping all of the spotted owls together into a combined spotted owl population. We then tested for significant difference between all western barred owls and the eastern barred owls. Finally, we grouped all of the barred owl samples together and tested for significant difference in spotted owl ancestry between the barred and spotted owls.

1.6.5. We used AD_pct_ex.sh from SPOW-BDOW-introgression-scripts version 1.1.1 to calculate the spotted owl ancestry for each sample at each variant site and also return the number of reads for the sample the site. We then used ext_fmt_sliding_window_reads.sh from SPOW-BDOW-introgression-scripts version 1.1.1 with default parameters to conduct a sliding window analysis with adjacent windows each 50,000 nt in length and calculate the average spotted owl ancestry in each window.

1.6.6. We used outlier_window_detection.py from SPOW-BDOW-introgression-scripts version 1.1.1 to detect and graph outlier windows. For each sample, we only considered windows where that sample had data for at least ten sites in that window. Outlier windows were those that had an average spotted owl ancestry >= 0.4 in samples with an average genome-wide ancestry close to 0. In samples with an average genome-wide ancestry close to 1, outliers were windows with an average spotted owl ancestry <= 0.6. The outlier_window_detection.py script also merged adjacent outlier windows.

1.6.7. In order to create the input file for calculation of π and F_{ST} statistics, we filtered the raw variant file using vcf_qual_filter_pi.sh from SPOW-BDOW-introgression-scripts version 1.1.1 to retain only biallelic sites sites where CAS:ORN:98821 (the source of the StrOccCau_1.0_nuc_masked reference genome) was not homozygous for the alternate allele. We used this script also to exclude indels, only retain sites where the Phred-scaled probability that a polymorphism exists was >50, and to filter out high coverage sites so as to only retain sites with coverage <301 X. We then used the script ad_pi_no_coords.sh from SPOW-BDOW-introgression-scripts version 1.1.1 to output the number of reads in the AD field that supported either the reference or alternative allele at each site. We then used this file as input to the script countFstPi from SPOW-BDOW-introgression-scripts version 1.1.1 to calculate π_{Within}, $\pi_{Between}$, and F_{ST} for various population comparisons. We used the "Category" column from Table S4 for population groupings. We included the reference genome source sample (CAS:ORN:98821) in the "Spotted Owl (pre-contact)" population. In order to arrive at our final π_{Within} and $\pi_{Between}$ values, we divided the output of the script by the number of A, C, G, and T characters in the reference genome sequence.

References

Bolger, A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire *et al*., 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491–498.

Dunn, O. J., 1961 Multiple Comparisons among Means. Journal of the American Statistical Association 56: 52–64.

GATK Dev Team, 2017 GATK Tool Documentation. [Accessed 2017 Oct 3]. Available from: https://software.broadinstitute.org/gatk/documentation/.

Hanna, Z. R., J. B. Henderson, and J. D. Wall, 2017a SPOW-BDOW-introgression-scripts. Version 1.1.1. Zenodo.

Hanna, Z. R., J. B. Henderson, J. D. Wall, C. A. Emerling, J. Fuchs *et al*., 2017b Northern Spotted Owl (*Strix occidentalis caurina*) Genome: Divergence with the Barred Owl (*Strix varia*) and Characterization of Light-Associated Genes. Genome Biol Evol 9: 2522–2545.

Hanna, Z. R., J. B. Henderson, J. D. Wall, C. A. Emerling, J. Fuchs *et al*., 2017c Supplemental dataset for Northern Spotted Owl (*Strix occidentalis caurina*) genome assembly version 1.0. Zenodo.

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv:1303.3997 Q-Bio. [Accessed 2016 Feb 16].

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al*., 2010 The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297–1303.

Quantum GIS Development Team, 2017 Quantum GIS Geographic Information System. Open Source Geospatial Foundation Project. [Accessed 2017 Sep 16]. Available from: http://qgis.org.

Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel *et al*., 2013 From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 11: 11.10.1-11.10.33.

Welch, B. L., 1947 The Generalization of "Student's" Problem When Several Different Population Variances Are Involved. Biometrika 34: 28–35.