Supporting Information

Genome divergence and diversification within a geographic mosaic of coevolution

Thomas L. Parchman, C. Alex Buerkle, Victor Soria-Carrasco, and Craig W. Benkman

Methods

DNA sequencing, assembly, and variant calling

We sequenced DNA from 219 red crossbills representing nine morphologically and vocally differentiated ecotypes (call types), as well as 12 white-winged crossbills (*L. l. leucoptera*; Table S1, Fig. S1). Blood samples were taken from live-caught birds, or tissue samples from vouchered specimens at the University of California at Berkeley Museum of Vertebrate of Zoology. Red crossbills were assigned to call type based on analysis of recorded contact calls (Groth, 1993; Benkman *et al.,* 2009; Irwin, 2010). We used samples of multiple populations from geographically separate regions within red crossbill ecotypes 2, 3, 4, 5, and 7 (Table S1, Fig. S1). DNA was extracted from blood or tissue samples using Qiagen DNeasy 96 Blood and Tissue Kits (Qiagen Inc., Valencia, CA, USA), and quantified with a NanoDrop spectrophotometer (Thermo Sci. Inc., Waltham, MA, USA).

We generated reduced representation genomic libraries for each individual using a genotyping by sequencing protocol that we have used previously for a diversity of taxa (Gompert 2012; Nosil *et al.,* 2012; Parchman *et al.,* 2012, 2013). We digested the genomic DNA of each individual with two restriction endonucleases (*EcoRI* and *MseI*; NEB Inc., Ipswich, MA, USA), and then ligated double-stranded adaptor oligonucleotides to the digested fragments. Uniquely barcoded *EcoRI* adaptors (containing 8, 9, or 10 bp barcodes) were ligated to fragments from each bird, and the adaptor-ligated fragments from all individuals were subsequently pooled. We

then used 288 replicate PCR reactions to amplify the pooled, adaptor-ligated fragments using a high fidelity proofreading polymerase (Iproof polymerase, BioRad Inc., Hercules, CA, USA) and Illumina-based PCR primers. To reduce the proportion of the genome targeted for sequencing, we ran libraries for 3.5 hours at 90V on a 2% agarose gel and excised gel regions containing fragments between approximately 350 and 450 bp in length. We purified these fragments using Qiaquick Gel Extraction Kits (Qiagen Inc., Valencia, CA, USA). A detailed description of this method and associated oligos is deposited at the Dryad digital repository (http://dx.doi.org/10.5061/dryad.65d97). For quality control, fragment size distributions were quantified with a BioAnalyzer (Agilent Inc., Santa Clara, CA, USA), and libraries were evaluated for sequencability using qPCR. Sequencing of the libraries was performed by the National Center for Genome Research (Santa Fe, NM, USA) using three lanes of single-end 100-base sequencing on the Illumina HiSeq 2000.

Using custom Perl scripts, we removed potential contaminant DNA (PhiX, *E. coli*) and low quality reads from the raw sequence data. Our sequencing library protocol resulted in reads that began with individual barcodes followed by the *EcoRI* restriction site. We trimmed barcodes and restriction-cut-site-associated bases from sequences, and matched barcodes from each read to the correct individual using a custom Perl script that allows for correction of sequencing errors in the barcode sequences. As a high quality *Loxia* genome is unavailable, we used a two-step assembly procedure to assemble reads into homologous genetic regions. We first used SeqMan NGen 3.0.4 (DNASTAR) to perform a *de novo* assembly for a subset of 30 million randomly sampled sequences. This assembly used a minimum match percentage of 92%, a gap penalty of 50, a match size of 50 bp, and a mismatch penalty of 15 (full details of assembly parameterization are available from TLP upon request). This assembly placed 24,352,918 reads

into 403,678 contigs. After removing low-quality contigs and those with consensus sequences <82 or >88 bases in length, we generated a partial artificial reference genome from the consensus sequences. This reference genome contained 349,865 contig consensus sequences from the original *de novo* assembly. We next assembled the full collection of sequencing reads for each individual onto the reference genome using the 'aln' and 'samse' algorithms in `bwa` 0.7.12 (Burrows-Wheeler Aligner; Li & Durbin, 2009).

We used `samtools` 1.19 and `bcftools` 1.19 (Li *et al*., 2009) to identify variant sites in the assembled sequence data and to call bi-allelic single nucleotide variants and estimate genotype likelihoods. We first identified single nucleotide polymorphisms (SNPs) in the assembled red crossbill reads for use in population genetic analyses. We disregarded insertions and deletions, and only called variants when 98% of the individuals had at least one read at that locus. Genotype likelihoods were calculated with `bcftools`, stored in variant call format, and converted to a convenient composite genotype likelihood format for downstream analyses. We removed variants for which more than one alternate allele was observed, and randomly selected a single variant from contigs containing multiple variants to increase the independence of loci. In addition, we discarded variants where the observed allele counts from apparent heterozygous individuals were very unlikely ($P < 0.001$) given a binomial distribution with $P = 0.5$. Finally, we excluded low frequency variants and limited analyses to loci having a SNP and minor allele frequencies >0.03.

After removing barcodes from the raw reads, and discarding reads with contaminants, we retained 321,627,388 reads representing 219 individuals. Initial *de novo* assembly placed 24,352,918 reads into 403,678 contigs; the 349,865 highest quality contigs from this assembly were used as an artificial reference genome. Assembling the reads for each individual onto this

reference created alignments containing a total of 253,736,350 reads, with an average of 1,071,000 reads aligned per individual. After using `samtools` and `bcftools` to call variant sites, discarding loci with minor allele frequency <0.03, and randomly sampling a single SNP per contig, we retained a final set of 18,385 SNPs (mean coverage per individual per locus of 7.2x) for population genetic analyses across the red crossbill complex.

References

Benkman CW, Smith JW, Keenan PC, Parchman TL, Santisteban L (2009) A new species of the red crossbill (Fringillidae: *Loxia*) from Idaho. *Condor*, **111**, 169-176.

Gompert Z, Lucas LK, Nice CC, Fordyce JA, Forister ML, Buerkle CA (2012) Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution*, **66**, 2167-2181.

Groth JG (1993) *Evolutionary differentiation in morphology, vocalizations, and allozymes among nomadic sibling species in the North American red crossbill (Loxia curvirostra) complex,* vol. 127. University of California Publications in Zoology.

Irwin K (2010) A new and cryptic call type of the red crossbill. *Western Birds*, **41**, 10-25.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760.

Li H, Handsaker B, Wysoker A, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.

Nosil P, Gompert Z, Farkas TE, et al. (2012) Genomic consequences of multiple speciation processes in a stick insect. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 5058-5065.

Parchman TL, Gompert Z, Braun MJ, et al. (2013) The genomic consequences of adaptive divergence and reproductive isolation between species of manakins. *Molecular Ecology*, **22**, 3304-3317.

Parchman TL, Gompert Z, Mudge J, Schilkey F, Benkman CW, Buerkle CA (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, **21**, 2991-3005.

Supporting Tables and Figures:

Table S1. The number of individuals sampled of *Loxia curvirostra* ecotypes, including geographic locations of separate samples within each ecotype, as well as white-winged crossbills *Loxia leucoptera*.

Taxa/species/location and abbreviation	Sample Size	Source
Loxia curvirostra		
Type 1	21	
Brush Mt., VA (1_A)	6	MVZ
Poverty Hollow, VA (1_B)	10	MVZ
Roan Mt., NC (1_C)	5	MVZ
Type 2	74	
Bears Paw Mts., MT (2_D)	4	CWB/PE
Black Hills, SD (2_E)	7	PE
Grand Marais, MI (2_F)	7	MVZ
Little Rocky Mts., MT (2_G)	8	CWB
Sagehen Creek, CA (2_H)	28	MVZ
Sandia Mts., NM (2_I)	12	CWB
Tennessee Pass, CO (2_J)	8	MVZ
Type 3	18	
Berkeley, CA (3_K)	15	MVZ
Sea Lion Point, OR (3_L)	3	MVZ
Type 4	16	
Sea Lion Point, OR (4_M)	9	MVZ
Thompson Plateau, BC (4_N)	7	MVZ
Type 5	22	
San Juan Mts., CO (5_O)	4	MVZ
Santa Catalina Mts., AZ (5_P)	4	MVZ
Tennessee Pass, CO (5_Q)	14	MVZ
Type 6	13	
Chiricahua Mts., AZ (6)	13	MVZ
Type 7	11	
Odell Creek, OR (7_R)	4	MVZ
Thompson Plateau, BC (7_S)	7	MVZ
Type 9/South Hills crossbill	37	
South Hills, ID (9)	37	CWB
Type 10	7	
Eureka, CA (10)	7	MVZ
Loxia leucoptera	12	
Alaska, Colorado	5	MVZ

MVZ: Museum of Vertebrate Zoology, UC Berkeley; CWB: Craig W. Benkman field caught; PE: Pim Edelaar field caught.

Table S2. Pairwise estimates of Nei's *D* (upper triangle) and F_{ST} (lower triangle) among *Loxia curvirostra* ecotypes (call types).

	Type 1	**Type 2**	**Type 3**	**Type 4**	**Type 5**	**Type 6**	**Type 7**	**Type 9**	**Type 10**
Type 1	–	0.0059	0.0167	0.0098	0.0096	0.0139	0.0122	0.0134	0.0170
Type 2	0.0099	–	0.0068	0.0068	0.0058	0.0101	0.0088	0.0098	0.0141
Type 3	0.0151	0.0111	–	0.0167	0.0109	0.0150	0.0109	0.0145	0.0167
Type 4	0.0169	0.0115	0.0162	–	0.0104	0.0150	0.0132	0.0149	0.0171
Type 5	0.0168	0.0096	0.0178	0.0178	–	0.0137	0.0119	0.0124	0.0177
Type 6	0.0239	0.0182	0.0250	0.0258	0.0230	–	0.0167	0.0170	0.0216
Type 7	0.0203	0.0162	0.0209	0.0242	0.0192	0.0285	–	0.0156	0.0201
Type 9	0.0219	0.0155	0.0231	0.0227	0.0190	0.0259	0.0241	–	0.0215
Type 10	0.0257	0.0257	0.0250	0.0272	0.0278	0.0330	0.0349	0.0331	–

Table S3. Deviance Information Criterion (DIC) estimates for entropy models run for $k = 2$ through $k = 9$. Lower estimates of DIC reflect better model fit.

k	DIC
2	8,606,675
3	8,651,678
4	8,502,511
5	8,514,392
6	8,573,614
7	13,860,352
8	117,161,960
9	151,557,301

Figure S1. The map illustrates sampling localities for red crossbill (*Loxia curvirostra* complex) ecotypes (call types). Individual points refer to geographically separate collection locations and correspond with Table S1. Circles of the same number and color represent different geographical samples from a given ecotype.