Supplementary Materials for

The genomic landscape underlying phenotypic integrity in the face of gene flow in crows

J. W. Poelstra, N. Vijay, C. M. Bossu, H. Lantz, B. Ryll, I. Müller, V. Baglione, P. Unneberg, M. Wikelski, M. G. Grabherr, J. B. W. Wolf*

*Corresponding author. E-mail: jochen.wolf@ebc.uu.se

Published 20 June 2014, *Science* **344**, 1410 (2014)
DOI: 10.1126/science.1253226

This PDF file includes:

Author contributions

JW, MW, VB and JP conducted field work; JP, JW and IM conducted the common garden experiment; NV, HL, MG and JW were responsible for genome assembly and annotation; JP and JW generated the RNA-seq data and conducted differential gene expression analyses with help from NV; BR performed all histological experiments; NV, CB, MG and PU ran the population genetic analyses; JW, NV, JP, CB, HL and MG wrote the Supplementary Information with input from the other authors; JW conceived and designed the study, supervised the project and wrote the manuscript together with JP, NV, CB and input from all other authors.

Supplementary Text

1. Genome assembly

Genome sequencing

DNA was extracted from blood of one male hooded crow individual (*Corvus (corone) cornix*) sampled in Uppsala, Sweden (59°51′N, 17°38′E) in 2010 using standard laboratory procedures, including proteinase-K digestion, phenol-chloroform extraction and RNA removal with RNase-A (Fermentas). Since birds have female heterogameity (ZW), the choice of a male specimen (ZZ) means that equal coverage can be obtained for autosomes and the large Z chromosome, and that the female-specific W chromosome will not form part of the assembly. Libraries were prepared with the Illumina TruSeq DNA Sample Prep Kit v2 according to the manufacturer's instructions. *De novo* sequencing was performed on an Illumina HiSeq 2000 platform using a combination of short paired-end libraries and longer mate-pair libraries suitable for use with the *ALLPATHS-LG* genome assembler (26). Paired-end libraries with a targeted insert size of 180bp each were constructed and sequenced from both ends at the SNP & SEQ Technology Platform at Uppsala University, Sweden. Mate-pair libraries with insert sizes of approximately two, five and twenty kb were generated and sequenced at the Beijing Genomics Institute, China. In total, 229 Gb of raw sequence data were generated (for details see **table S1**).

Genome assembly

Genome assembly was performed with *ALLPATHS-LG* version-41687 using the default parameters for a ploidy value of two. *ALLPATHS-LG* was run on 32 nodes, 512 GB RAM with disk usage of up to 3TB using the computational infrastructure available at the UPPMAX computing facility at Uppsala University (http://www.uppmax.uu.se/). Total run time was ~ 18

days (403.39 hours). The *ALLPATHS-LG* pipeline consists of 57 separate modules which are run consecutively to obtain a genome assembly starting from the raw sequencing data. The module "FIX_LOCAL" had very large I/O requirements and repeatedly crashed. Hence, this module had to be skipped by setting "FIX_LOCAL=False" option when running *ALLPATHS-LG*. All the other modules completed without any error messages.

ALLPATHS-LG works on raw data without prior adapter removal and trimming, and performs its own read correction steps. The sequencing error rate per base was estimated at 0.0056 (Q = 22.5), and 2.2% of the raw reads were removed. Information on insert-size for each library along with the expected standard deviation is required as an input parameter, but cannot be provided with sufficient accuracy by sequencing centres in the absence of a reference genome. Therefore, we first generated a rough draft genome assembly with *SOAPdenovo* 1.05 (27) using default parameters (k = 45), on which we subsequently mapped all raw reads using *bwa* 0.7.4 (28). From the resulting sequence alignment data file we derived the insert-size distributions for each library. Two mate-pair libraries (libraries 7 and 12 in **table S1**) showed large deviations from their expected insert-size due to paired-end contamination, and were excluded to prevent misassemblies. A series of test assemblies further showed that data reduction of paired-end coverage by one third improved the assembly considerably. The data-reduced assembly was more contiguous, had less circular scaffolds, and correctly retrieved the mitochondrial genome which had been placed into the nuclear-mitochondrial rich scaffold in the full data assembly. The final assembly was based on 184 Gb of raw sequencing data.

Genome size and genome coverage

Data from DNA flow cytometry suggests haploid genome sizes of 1.46 pg (29) and 1.23 pg (30) for carrion/hooded crows. The latter estimate is very similar to chicken (1.25 pg) and is supported by genome size information from other species in the same genus including northern raven (C. *corax,* 1.21 pg), western jackdaw (*C. monedula,* 1.25 pg) and the American crow (*C. brachyrhynchos,* 1.25 pg, 1.27 pg) (30–32). A mass-based genome size of 1.23 pg corresponds to 1.21 Gb, which is also similar to the 1.22 Gb golden path length of the zebra finch draft assembly. We obtained a k-mer frequency based estimate of genome size with a mean estimate of 1.26 Gb (33), which is in good agreement with the C-value based information.

The final genome assembly generated by *ALLPATHS-LG* consisted of 1,298 contigs with an N50 of 16.3 Mb, had a contig length of 1.01 Gb, and a total scaffold length of 1.04 Gb. 95% of the

final assembly was contained within the 100 largest scaffolds. This corresponds to an average of 2.5 scaffolds per chromosome (hooded crow karyotype: 2n=80 (34)). The initial draft assembly generated by the *SOAPdenovo* assembler (see above) had a genome size of 1.16 Gb. The discrepancy in assembly sizes is partly due to *ALLPATHS-LG* removing all contigs shorter than 1 kb, which made up 19Mb in the *SOAPdenovo* assembly. Overall, given the high raw-read sequencing coverage of 184 Gb (152x assuming a genome size of 1.21 Gb) it seems reasonable to assume that most of the genome has been sequenced despite somewhat lower coverage in regions with extreme GC content. The 'missing sequence' can be attributed to the conservative approach of *ALLPATHS-LG* of only using longer contigs and excluding repeat content or regions that are otherwise difficult to assemble (see below). Further properties of the final assembly can be found in **table S2**.

Assembly validation and improvement

Duplicate scaffold removal

Although *ALLPATHS-LG*'s "CleanAssembly" step is designed to remove duplicate scaffolds, the final assembly contained 14 scaffolds that were exact duplicates with a total duplicate length of 30,365 bp. Another 28 scaffolds were found to be exact duplicates of reverse complemented scaffolds with a total length of 33,734 bp. For each of these pairs, we kept only one unique copy for the final version of the genome.

Contamination

The multitude of steps involved in sampling, extraction and library preparation provides many potential sources of contaminant DNA, which can be misassembled along with the genuine DNA from crow. *ALLPATHS-LG* attempts to reduce the contribution of contaminant sequences to the final assembly by removing low frequency k-mers, and by discarding scaffolds shorter than 1 kb. Indeed, the final assembly appeared to be largely free from contamination. Using the *Satsuma* aligner (35), we found that most scaffolds in the assembly readily aligned to chicken (1,139 of 1,298) and zebra finch (1,248 of 1,298). Scaffolds that failed to align had top hits from bird nucleotide sequences or from the lizard genome (*Anolis carolinensis)* in *BLASTN* searches against NCBI with an e-value cut-off of 10^{-5}. To identify contaminant sequences that could have been incorporated into the scaffolds, each of the contigs was also aligned to the NCBI nucleotide database with *BLASTN* and the search was repeated by cutting up the assembly into 10 kb chunks to ensure that no contaminant sequences were present within contigs. No putative contaminant sequences were found in the assembly using this approach.

Repeat content

Repetitive sequences are difficult to assemble and are prone to be excluded or collapsed in genome assemblies, especially in assemblies from short-read shotgun sequence data. *ALLPATHS-LG'*s own k-mer spectrum analysis estimated that 13.3% (0.12 Gb) of the genome was repetitive (copy number > 1). Repeats were identified in the assembly using *RepeatMasker* version open-3.2.9 (36) with the settings -gccalc and -species "aves". A total of 60 Mb (7.76%) of the assembly was repeat masked using these settings (compared to 9.08% and 7.93% in chicken and zebra finch, respectively). Most of the identified repeat elements were retro-elements which made up 4.87% of the assembly (**table S9**). All repeat regions were excluded for subsequent population genetic analysis.

Circular scaffolds

ALLPATHS-LG identifies circular scaffolds using the "Tag circular scaffolds" module. We expected only one circular scaffold corresponding to the mitochondrial genome. The mitochondrial scaffold generated by *ALLPATHS-LG* matched the assembly obtained by local reassembly (using SOAPdenovo) of reads mapping to the rook (*C. frugilegus)* mitochondrial genome with the *stampy* read mapper (37), allowing for 15% divergence. In addition to the mitochondrial genome, a total of sixty-nine scaffolds were tagged as circular (median length: 2819 bp, max length: 17 Mb). While the overall repeat content of the circular scaffolds was similar to the genome-wide average, the ends of most of these scaffolds were enriched with repeat content. Since all of these circular scaffolds aligned to bird genomes in *BLASTN* contamination checks, they were retained in the final assembly.

Gaps

The final assembly obtained from *ALLPATHS-LG* contained 39,425 stretches of N's each corresponding to an average gap size of 938 bp (range 1 to 19,410 bp). *GapCloser version 1.10* from the *SOAP* package (http://soap.genomics.org.cn/about.html#resource2) was utilized to close gaps in the scaffolds using all the paired end libraries. Of the 39,425 gaps found in the *ALLPATHS-LG* assembly, 11,632 were closed and the number of N's was reduced from 36,971,193 (3.52%) to 27,496,850 (2.62%) resulting in a more contiguous assembly.

Local re-assembly of targeted regions

Genes embedded in repetitive regions are often not included in short-read assemblies (38, 39). In

this study, we focused on genes relevant to melanin-based feather pigmentation. To ensure that all key genes of interest were included in the assembly, the initial annotation (see below) was screened for the presence of orthologous genes implicated in the melanogenesis pathway (see "Gene expression analysis" section). Of these genes, 13 appeared to be absent from the assembly. To retrieve the missing genes, their approximate position in the crow assembly was identified based on the position of orthologous genes in chicken and zebra finch genomes. All raw reads were then mapped to this target region in zebra finch or chicken, containing the gene of interest, with the *bwa* read mapper (28). Successfully mapped reads were then re-assembled *de novo* by the *SOAPdenovo* assembler using different k-mer values. The best resulting contigs were then used to bridge gaps in the targeted scaffold of the crow assembly. Of the 13 missing genes, only the *MC1R* gene could be satisfactorily assembled by this local reassembly method. However, nine of the remaining twelve genes were found as transcripts (full CDS as compared to zebra finch) in an RNA-seq based *de novo* transcriptome assembly using *Trinity* (40). Of these, most critical to our context were *POMC* and *PMEL*. The *de novo* assembled transcripts were not included in the final genome assembly, but were used as a backbone for the differential expression analyses.

Synteny based chromosome building

The crow genome assembly was aligned to chicken and zebra finch genomes using the *Satsuma* aligner with default settings (35). As is typical for bird species (41), synteny was conserved with respect to both chicken and zebra finch. Judging from alignments in 50 kb blocks, the 139 largest scaffolds (covering 98% of the assembly) showed no large-scale inter-chromosomal rearrangements. With the exception of a few smaller regions, entire scaffolds aligned to a single zebra finch / chicken chromosome (**fig. S2**) and allowed synteny-based assignment of scaffolds to the respective chromosomes. Chromosomal assignment did not depend on the choice of reference genome except for the known karyotypic differences between zebra finch and chicken (42). Most of the smaller micro-chromosomes (chromosomes 11-28 in zebra finch) were covered entirely by only two to three scaffolds (mean: 2.5, max: 6, min: 1). Macrochromosomes (chromosomes 1-5 in zebra finch) were covered by 10 to 15 scaffolds (mean: 10.6, max: 15, min: 7) and intermediate chromosomes (chromosomes 6-10 in zebra finch) by three to four scaffolds (mean: 3.4, max: 5, min: 2) each. Alignments for the remaining scaffolds harbouring repetitive sequences could not be assigned to a chromosome with certainty, and were grouped in the 'Unknown' category where no synteny for at least a 50kb syntenic block could be found. As expected, chromosome 16, which is known to harbour the highly duplicated region of the major

histocompatibility complex in chicken, was found to have fewer regions that could be aligned with corresponding scaffolds in the crow genome.

Despite overall highly conserved synteny, several intra-chromosomal rearrangements were detected (see **fig. S2C** for an example). When the crow assembly was aligned to the chicken and zebra finch genomes, alignment-based scaffold orientation depended on which reference genome was used. Future construction of a linkage map will be necessary to distinguish between assembly artifacts and fine-scale rearrangements. For the purpose of this study, scaffolds were thus merely assigned to chromosome groups, and population genetic analyses (see below) were conducted at the scaffold level. In cases where regions of interest lay at the end of potentially contiguous scaffolds, scaffold orientation was primarily based on the positions on the zebra finch chromosome using the *Satsuma* alignments followed by verification with *progressiveMauve* (43), and chicken-based scaffold orientation is also reported and discussed.

2. Genome annotation

Mitochondrial genome

The mitochondrial genome was correctly identified as a circular scaffold by *ALLPATHS-LG*. We rotated the original scaffold so that its ends did not span any genes. The MITOS Webserver (44) was used to identify 2 rRNA, 22 tRNA and 13 protein coding genes. Gene order and content were found to be the same as found in the previously published mitochondrial genome of the rook (*Corvus frugilegus*). RNA-seq data was used to further validate the genes identified *in silico*. Relative expression of the different protein coding genes was found to correspond well to previously published estimates for carrion/hooded crows based on 454 sequencing technology (45).

Synteny inferred gene models

For annotation of nuclear genes, the crow genome was aligned to the zebra finch genome (Ensembl, version taeGut3.2.4) using *SatsumaSynteny* (35) with default settings. Annotations were then syntenically lifted over from the zebra finch Ensembl annotation version 68 onto the crow genome (M. Grabherr, custom pipeline. The source code is freely available under the GNU Lesser General Public Licence from http://github.com/nedaz/kraken.). In addition, *Scipio* 1.4 (46) was used to align zebra finch proteins (Ensembl version 68) to the crow genome with an 80% identity cut-off and otherwise default settings. The alignments were then filtered to exclude introns longer than 120kb (maximum zebra finch intron length) using “gffread” in the *Cufflinks*

2.0.2 package (47, 48). The intron-size cutoff was chosen based on the distribution of intron sizes in chicken, where 99.9% of all introns are 120 kb or smaller.

RNA-seq based annotation

RNA-seq data was obtained from 19 individuals from forebrain, liver, gonads, skin, cerebellum, hypothalamus & pituitary, eye, heart, and spleen (see "Gene expression analysis" section for details). A total of 120 libraries (**table S3**) were sequenced, generating 2,067 million paired-end reads resulting in a total of 413 x 10^9 bp of sequence. Assuming a maximum transcriptome size of 343 Mb (as inferred from mapping to the genome), this corresponds to approximately 1,205x total transcriptome coverage. RNA-seq reads from each individual and tissue were separately mapped to the crow genome assembly using *Tophat* version 2 (49). The multiple read correction flag "u" was used to ensure better initial estimation of reads mapping to multiple locations in the genome. The resulting 120 alignments were merged by tissue and taxon using *picard tools* (version 1.46) and subsequently assembled with *cufflinks* (version 2) (47, 48). The minimum isoform fraction flag "F" was set to a value of 0.4 and the pre-mRNA fraction flag "j" was set to a value of 0.8 after investigating various combinations of these parameters. *Cuffmerge 2.0.2* (47) was then used to merge the 19 resulting transcript files (see **table S3**) into a single file using default settings. All transcripts with unknown strand information were assumed to be on the positive (+) strand. Finally, single exon transcripts that overlapped an intron but not any other exon were suspected to be remnants of pre-mRNA. A version of the annotation where 12,986 of such transcripts were removed was created using a combination of the *IntersectBed* tool of *Bedtools 1.4* (50) and custom scripts.

Total number of genes and gene name assignment

The RNA-seq based annotation identified 75,676 unique transcripts from 42,100 unique genes. The number of transcripts per gene varied from 1 to 21, with an average of 1.79 and a median of 1 transcripts per gene (29,641 genes had only one transcript, and 12,459 genes had more than one). 20,794 genes had continuous open reading frames of more than 100 amino acids as reported by *Transdecoder* (B. Haas; http://transdecoder.sourceforge.net).

In a next step, we assigned orthologous genes to the gene IDs that were identified *de novo* on the basis of RNA-seq data. Based on the *Scipio* zebra finch protein alignments, we transferred zebra finch names to gene IDs using *Cuffcompare 2.0.2.* (48) with default settings. We also made use of gene name information from our custom liftover pipeline and *BLAST* routines (*TBLASTX*,

BLASTX and *BLASTN)*. Combining results from *BLAST* (46,974 crow gene IDs linked to 13,255 unique Ensembl IDs), *Scipio* (14,022 genes linked to 12,302 unique Ensembl IDs), and liftover (15,590 genes linked to 11,448 unique Ensembl IDs), a total number of 23,456 crow genes (out of 42,100) were linked to 15,566 unique zebra finch Ensembl ID. 10,386 crow genes were linked to more than one zebra finch reference gene. We ranked our confidence in any zebra finch Ensembl ID reference as follows, from highest to lowest rank: references identified by (1) all three methods (*BLAST*, *Scipio*, liftover), (2) *Scipio* and *BLAST*, (3) *BLAST* only, (4) *Scipio* and liftover, (5) *BLAST* and liftover, (6) *Scipio* only, (7) liftover only. For the purpose of functional annotation, we defined our primary reference gene as the ID with the highest rank.

3. Population re-sequencing

Population sampling

A total of 60 crow nestlings were obtained from two populations of each taxon during May-June 2010 (see map in **Fig. 1**). Fifteen carrion crows were obtained from Germany (Radolfzell/Konstanz in Baden-Württemberg: 47°40'-47°45N', 9°0'-9°12'E) and 15 from Spain (La Sorriba in the province of León: 42°35'N, 5°29'W). In addition to the genome individual, 15 hooded crows were obtained from the Uppsala area in Uppland, Sweden (59°52'N,17°38'E). Fifteen Polish nestlings from the Warsaw area in Mazowieckie, Poland (52°14'N, 8°55'E) were collected courtesy of Dr. Andrzej Kruszewicz from the animal rehabilitation centre of Warsaw Zoo, which receives nestling crows that are found unattended by parents. In the three remaining populations, we collected birds from their nests, individually marked them with leg-rings, and took blood from the brachial vein. From the blood, we extracted DNA using QuickExtract™ (Epicentre, Illumina) and molecularly sexed all nestlings (51). We restricted our sample to one individual per nest to obtain only unrelated individuals. In one case of two German individuals, a sibling pair could not be avoided. We further chose only male individuals. This guarantees equal sequencing coverage for autosomes and the Z-chromosome and avoids systematic biases arising due to coverage differences for population genetic analyses. One female individual needed to be included in the German population sample to reach our target sample size of fifteen.

Permissions for sampling of wild crows were granted by *Junta de Castilla León* (Ref: CML/mjg, Expte: EP/LE/410/2010) in Spain, by *Regierungspräsidium Freiburg* (Aktenzeichen: 55-8852.15) in Germany, and by *Jordbruksverket* (Dnr 30-1326/10) in Sweden.

Whole genome re-sequencing

In addition to 30 carrion crow and 30 hooded crow individuals sampled from the four European populations described above, we generated whole genome sequence data from several outgroup species. We included five individuals of the American crow, two individuals of rook and two individuals of western jackdaw, in order to confidently estimate the ancestral state of variants segregating in carrion and hooded crows (see below). American crow samples were provided courtesy of American Museum of Natural History, New York (AMNH# DOT7632, DOT10136, DOT13715, DOT13858, DOT13859), and rook and jackdaw samples were provided by hunters in Belfast, Ireland and Uppsala, Sweden, respectively.

For each individual we extracted genomic DNA with a standard phenol-chloroform protocol, and subsequently prepared one paired-end library per individual (average insert size: 400 bp, standard deviation: 28 bp) with Illumina TruSeq DNA Sample Prep Kit v2 according to the manufacturer's instructions. All 69 libraries were sequenced on an Illumina HiSeq2000 machine to an average raw read sequence coverage of 12.2x (range 7.1x – 28.6x) assuming a genome size of 1.21 Gb. Individually barcoded libraries were split into five aliquots each and distributed across lanes and flow cells in a random blocked design. This approach avoids that potential biases in the sequencing setup systematically influencing our variant calling, genotyping, and population genetic analyses. More detailed information on sequencing depth is given in **table S4**.

Variant discovery and genotyping

In a first step, raw reads were mapped to the hooded crow reference assembly using *bwa 0.7.4* (28) with default settings. On average, 93% of the reads were mapped, resulting in a final average sequencing coverage of 11.2x per individual (range: 6.8x - 26.6 x). The *picard* software was subsequently used to assign readgroup information containing library, lane, and sample identity. To enhance the alignments in regions of insertion-deletion polymorphisms, we additionally performed local realignment using *GATK* (52). Duplicate read-pairs were marked at the library level with *picard*, removing 6 to 15% of the mapped reads depending on the quality of the library.

Variant discovery analyses were conducted at the population level for each of the four populations and the American crow samples separately. Base quality score recalibration (BQSR) is known to improve variant discovery and genotype calls, but requires knowledge of true variant sites. As we do not have previous knowledge of such sites for crows, we used the following iterative approach to identify a reliable set of high quality SNPs. An initial round of variant

calling was performed on the original, uncalibrated sequence alignment files using three variant discovery software programs: *UnifiedGenotyper* in *GATK* version 2.3.6 (52), *samtools* version 0.1.18 (53) and *FreeBayes* version 0.9.8 (54). Default software settings were used for each program. The intersect of variant sites between all three methods was extracted and used as the set of known variants for the first round of BQSR. BQSR and variant calling was conducted a second time using *GATK* exclusively. A variant set filtered for the highest quality variants detected in the initial BQSR implementation (approximately 10-15%) was used as the set of known variants for the second round of base quality score recalibration. Convergence was rapid; approximately 99.5% of the detected variants were shared between the first and second round of base quality score recalibration.

Variant quality score recalibration (VQSR), a post-discovery error modeling algorithm implemented in *GATK* (52), can further improve variant calling. VQSR uses known variant sites to estimate the probability that each variant is a true genetic variant or a machine artifact. In the absence of previous knowledge of "true" variants, we utilized 10-15% variants with the highest quality to generate an error model. Finally, a catalog of all variable sites within and between all hooded and carrion populations was merged and subsequently genotyped using *GATK*. Genotyping was performed for each population separately in order to avoid biases against minor allele frequencies varying between populations. In a final filtering strategy, variants detected in fewer than seven individuals per population and variants within repeat-masked regions were removed, resulting in a total of 8,438,802 SNPs.

Validation of variant calls and genotypes

As a first validation of the variant calling procedure, we compared the number of segregating sites identified by *GATK* to the expected number of segregating sites as inferred from intronic amplicon sequences obtained with Sanger technology (55). We limited the considerations to autosomally-linked scaffolds, as nucleotide diversity is expected to differ between sex chromosomes and autosomes. Autosomal Sanger data was available for 97 intronic loci (totalling 54.2 kb) sequenced for 23 individuals from Germany and 20 individuals from Poland (55). We here generated an additional dataset using a subset of 37 intronic loci (totalling 20.5 kb) for 12 individuals each from Spain and Sweden.

Based on whole genome re-sequencing coverage, we first estimated, for each individual, the

average (binomial) probability that both chromosomes of the diploid genome had been sampled (i.e. sequenced). To conservatively account for the possibility of sequencing error, a chromosome had to be sampled at least twice. These estimates were then used to assess the most likely number of chromosomes sampled in a given population. This number was slightly below the maximum number of 30 chromosomes per population (Germany / Spain / Poland / Sweden: n = 25 / 23 / 23 / 25). Using estimates of Watterson's theta from the Sanger data, we then calculated the expected number of segregating sites for the total assembly size of 974 Mb of autosomally-linked scaffolds. The estimated number of segregating variants identified by *GATK* was very close to the expectation. This was indicated by observed/expected ratios of close to 1 (Germany / Spain / Poland / Sweden: 0.94 / 0.92 / 0.91 / 1.00) and by the fact that expected and observed values were highly correlated across populations (R^2=0.90, p=0.0507, df=2). Under the assumption that genetic diversity of introns is representative of most of the genome, this provided a first indication of accurate and efficient variant calling.

To gain further insights into variant calling and genotype accuracy, we compared individual genotypes from the re-sequencing data set with genotypes from 53 loci (totalling 30 kb) sequenced with the traditional Sanger method for the same individuals. Genotypes from the Sanger dataset could be compared to genotypes inferred by the *GATK* pipeline described above for the same 30 individuals sampled in Spain (10 carrion), Poland (10 hooded) and Sweden (10 hooded), respectively. Genotypes from the Sanger dataset identical to genotypes from the shotgun sequencing data were considered true positives. Conflicting genotypes were considered false positives, and sites that were detected as heterozygotes in the Sanger sequences, but not in the re-sequencing data, were considered false negatives. 99.95% of genotype calls were true positives, 0.031% were false positives and the remaining 0.024% were false negatives.

Ancestral state reconstruction

Several population genetic statistics require information on the ancestral state of segregating variants. Therefore, we genotyped five American crows, two rooks and two jackdaws using *GATK* for all sites segregating in any of the crow populations. A custom *perl* script was used to loop over these genotype calls for each site to identify the ancestral state. 73.39% of the sites were found to have the same homozygous genotype in all American crows, rooks and jackdaws, and could be directly used as the ancestral state. 9.53% of the sites were polymorphic in American crow, but fixed in rooks and jackdaws. These sites most likely reflect ancestral polymorphism prior to the split of American and carrion/hooded crows and are uninformative

about the ancestral state of carrion and hooded crows. 3.86% of the sites were fixed in American crow and rooks, but polymorphic in jackdaws. This could either be due to a mutation in the lineage leading to jackdaw (homoplasy), or due to incorrectly mapped paralogs, and these sites were therefore excluded. Also discarded from further analysis were tri-allelic sites (constituting 7.36%) and sites that lacked genotyping information or were polymorphic in rooks (constituting 5.86%).

4. Population genomics

Principle component analysis

Principle component analysis (PCA) was performed on all SNPs using the software *Eigensoft* (56). *Eigensoft* pruned the initial variant set to 7,524,126 SNPs and identified one Swedish hooded crow (S02) as an outlier individual, which was subsequently removed from the analysis. Coordinates from the first principle components were projected on geographical coordinates using Procrustes analysis as implemented in the *shapes R* package.

Sliding windows

We partitioned the genome into 22,072 non-overlapping sliding windows, 50 kb in size. Windows were distributed on a per-scaffold basis starting at position 1 of a scaffold and were then oriented along *in silico* chromosome builds. For a window to be included in the downstream analysis, it was required to pass a quality threshold based on the quantification of the proportion of sites that was of sufficiently good quality and of high enough coverage to be callable by the variant calling program *GATK* ("callability" criterion). Windows with extremely low values based on a 2.5% quantile cut-off were marked for exclusion. This corresponded to windows with callability of less than 29% in Spain, 35% in Germany, 34% in Poland and 40% in Sweden. For each window, we then calculated how many of the 60 individuals were marked for exclusion. Assuming a random distribution of low quality values across the genome and across individuals, the expected number of individuals with low quality values for a given window can be approximated by binomial sampling with success rate 2.5% (corresponding to our cut-off). Exclusion in more individuals than expected by chance point towards systematic problems with a window (due to e.g. high GC content, high paralogue content). Consequently, all windows with low callability in more than six individuals ($p<0.001$) and windows with extreme repeat content exceeding 20.4% (2.5% upper quantile) were removed from subsequent analyses. In total, this resulted in the exclusion of 1,818 windows.

For each of the remaining 20,254 windows, we quantified a series of population genetic parameters, as will be described in the following sections.

Site frequency spectrum estimation

Nielsen *et al.* (57) suggested a probabilistic method implemented in the software package *ANGSD* to obtain direct estimates of the site frequency spectrum (SFS) from short-read sequencing data. Using default parameters and genotype likelihoods based on the *GATK* genotyping model (52), we estimated the unfolded SFS in steps of 50 kb windows and derived the following summary statistics that are direct functions of the SFS for each population: population mutation rate θ (Watterson's estimator), Tajima's D, Fu and Li's D and Fu's Fs. *ANGSD* requires the input of one outgroup genome for the estimation of the unfolded SFS. To make use of the information of all our three outgroups, we generated the consensus sequence for each of the five American crows, two rooks and two jackdaws. All nine consensus sequences were joined using the *samtools* "mpileup" command. A custom *perl* script was used to loop over these consensus sequences to identify the ancestral base at each position. 93% of the bases had an ancestral state in agreement with American crows and rooks. The rest of the bases were marked as N and consist of tri-allelic sites (5.57%) and those that are polymorphic only in rooks or polymorphic both in jackdaws and American crows but not in rooks (1.33%). The inferred ancestral genome sequence jointly considering three outgroup species was then presented as the outgroup sequence to *ANGSD*.

F-statistics, Dxy, Da, df

To estimate genetic differentiation we calculated hierarchical F-statistics as implemented in the *HierFstat R* package (58, 59). *HierFstat* uses the Weir & Cockerham (60) F_{ST} estimator in a nested analysis of variance framework for hierarchical population structure (61). F_{ST} was estimated separately for each locus of sufficient sequence coverage (reliable genotypes for ≥ 7 individuals per population) in a two-level hierarchy between populations and between taxa. Window-based F_{ST} estimates were then calculated by averaging the variance components across loci.

Custom scripts were used to calculate the number of segregating sites (S), mean nucleotide diversity (π), the average number of nucleotide substitutions (D_{xy}), the number of net nucleotide substitutions per site (D_a), and the number of fixed differences (df) (cf. 62). All parameters

(except for the number of segregating sites) were calculated on a per-locus basis and were averaged to obtain window-based estimates. Accurate estimates of π, D_{xy} and D_a require standardization by the total number of available sites per window passing the initial quality filters (*i.e.* not within repeat-masked regions and identified in a minimum of 7 individuals per population). We therefore applied *samtools* "mpileup" (53) to recalibrated *.bam* files using the same quality filters for non-segregating sites as for segregating sites.

Haplotype statistics

SNP data from each population was phased separately using the program *beagle* version 3.3.2 (63) with 10 iterations each. Sites that could be phased in all 15 individuals of a population were used to calculate haplotype statistics indicative for variation in linkage disequilibrium across the genome. These included iHHa, iHHd, iES and iHS which were calculated using the *R* package *REHH* (64).

ABBA-BABA test

To test for differential gene flow between populations of carrion and hooded crows we used the four taxon "ABBA-BABA" testing procedure based on the phylogeny ((H1:H2)H3)H4. An excess of the ABBA pattern is interpreted as gene flow between H2 and H3, whereas an excess of the BABA pattern is interpreted as gene flow between H1 and H3. To quantify the excess, we calculated the D-statistic as defined in Durand et al. 2011 (65) and implemented in the software ANGSD (57). To assess gene flow patterns between all four populations [(Sp)ain, (Ge)rmany, (Po)land, (Sw)eden] we considered all 12 possible three-way combinations between H1-H3. One high coverage individual from each of the four crow populations was used for the test. As an outgroup (O), the ancestral sequence from three outgroup species (as described above) was used, corresponding to H4. Because linkage disequilibrium decays rapidly in crows (55), we chose to use a smaller block size (50 kb) than default, which also allowed the inclusion of regions from the smaller scaffolds. Using a larger block size did not qualitatively change the results. The significance of the deviation of the D-statistic from 0 indicating gene flow is ascertained by the Z-score, which is based on jack-knife estimates of standard deviation of the D-statistic.

There was no evidence for gene flow for the trees ((Po:Sw)Sp)O and ((Po:Sw)Ge)O, supporting the view that Poland and Sweden are genetically closer to each other than to both Germany or Spain, as suggested by the principle component analysis (see above and results in main text). Yet,

in cases where Germany and Spain formed a clade (H1:H2), there was significant gene flow between Germany and both Poland and Sweden. In cases where Germany and Poland or Sweden formed a clade, there was evidence for gene flow between Spain and Germany (see **table S6**).

The D-statistic has also been defined in Patterson et al. 2012 (66) based on population allele frequencies rather than character states of single individuals. We used the implementation of this definition in *egglib* (67, 68) and calculated the D-statistic in windows of 20 Kb each following the default. The genome-wide mean values of the D-statistic (see **table S6**) were similar to those obtained from the previous definition.

Admixture analyses

To quantify genome-wide admixture between carrion and hooded crow populations, we estimated ancestry of each individual using the genome-wide SNP dataset and the model-based assignment software program ADMIXTURE 1.23 (69). ADMIXTURE was run for each possible group number (K= 1 to 4) with 200 bootstrap replicates to estimate parameter standard errors used to determine the optimal group number (K). Under the constraint of two population clusters (CVerror = 0.358), the German population is represented as the product of admixture between Spanish carrion crows (~0.19 assigned ancestry) and hooded crows (~0.81 assigned ancestry) (**fig. S6**). The inference of admixed individuals in the German population is consistent with the ABBA-BABA inference of gene flow between Germany and Spain, and between Germany and both hooded crow populations.

Gene flow estimation with IMa2

To estimate rates of gene flow across the hybrid zone, we used the program *IMa2* (70) on previously published Sanger sequencing data of 20 German carrion and 23 Polish hooded crows (55). Within one gene, one locus from pairs of loci with significant linkage disequilibrium was chosen, and loci with significant deviations from neutrality were excluded. For each of the selected 77 loci, the longest stretch with no evidence for recombination was taken. *IMa2* uses MCMC simulations, and for the final runs we used 100 Metroplis Coupled chains, a burn-in of 250,000 steps, and a final run length of 1,000,000 steps. Prior boundaries were determined using preliminary runs and set at 0-1 for divergence time, 0-100 for gene flow rates (m) and 0-7 for all three population sizes (θ). This analysis suggested significant gene flow from the Polish hooded crow population into the German carrion crow population and less, but non-zero gene flow in the other direction (**fig. S7**).

Substitution rate estimation

Heterogeneity in genetic variation across the genome can be caused by selection or by differences in mutation rate. In neutral regions of the genome, substitution rate estimates can be used as a proxy for inference of mutation rates. Four-way codon-based alignments including open reading frames in crow and corresponding orthologous genes in zebra finch, chicken and turkey (Ensembl 72) were generated with *GUIDANCE-HoT* (71) using *PRANK's* (72) progressive alignment algorithm. From 9,589 resulting reliable gene-alignments, substitution rate estimates were obtained at 4-fold degenerate sites (d4) using model 0 for codons as implemented in *CODEML* from *PAML* version 4.7 (73). Genes with d4 estimates larger than 1.5 were removed, as such high substitution rates may reflect incorrectly inferred orthology or reading frame rather than biological reality. For each window, substitution rate estimates were averaged across all genes present in that window. The mean genome-wide substitution rate (d_S) was 0.1755. Substitutions rates did not differ between regions of elevated differentiation between taxa ("peaks"see below, d_S = 0.1618) and the genome-wide average (d_S = 0.1753, Kruskal-Wallis p-value = 0.985). This led us to conclude that differences in mutation rate were not responsible for the observed differences in genetic diversity among peak and non-peak regions.

5. Outlier screens

Window-based

Estimates of F_{ST} between carrion and hooded crows on a per-window basis ranged from -0.0323 to 0.6957 (mean = 0.0620, see **table S5** for pairwise F_{ST} comparisons between all four populations). Given that sex chromosomes differ from autosomes in several properties that can affect population genetic estimates (such as effective population size, mutation rate, and recombination rate), we partitioned the genome into autosomally-linked scaffolds and scaffolds linked to the Z-chromosome to test for outlier windows. Randomization tests were used to determine statistical significance of empirical window-based F_{ST} estimates. In 100 randomizations using all SNPs, we shuffled per locus variance components and diversity estimates without replacement and re-calculated windows-based averages. This strategy does not alter the hierarchical population structure, nor change the number of variants per window. Using a type I error level corresponding to the 99th percentile of the randomized distribution, 3,693 windows were identified as exceeding random genome-wide levels of differentiation ($F_{ST} \geq$ 0.0830). Since this was more than 1% of all windows, we chose to use the more conservative empirical cut-off. 204 windows exceeded the 99th percentile of the F_{ST} empirical distribution for

autosomes ($F_{ST}\geq$ 0.2040) and 16 for the Z-chromosome ($F_{ST}\geq$ 0.2950), and were accordingly classified as statistical outliers.

Of the 220 windows (189 after removing low quality windows) with extreme F_{ST} values, 52 windows were singletons. The vast majority of windows (137) were clustered in pairs (called peaks hereafter). Fourteen windows were clustered in pairs of two, 12 in pairs of three, and 73 windows were clustered in large peaks comprising more than three adjacent windows. Clusters were mainly located at the end of scaffolds, a pattern also observed in flycatchers (5).

A comparison of population-level pairwise F_{ST} comparisons indicated that divergent Spanish alleles strongly contributed to outlier peaks between carrion and hooded crows, reflecting population-specific differences rather than divergence between taxa as a whole. For example, pairwise F_{ST} estimates across scaffold 57 illustrate a peak present in all population comparisons with Spain, which was however not detected in comparisons between German carrion crows and either hooded crows from Poland or Sweden (**Fig, 3**). In total, five peaks were identified as taxon specific and were located heterogeneously across the genome on scaffold 29, scaffold 7, scaffold 48, scaffold 43 and between adjacent scaffolds 78 and 60 (**fig. S8A-E**). These potential "speciation islands" had elevated F_{ST} between taxa (Germany-Poland, Germany-Sweden, Spain-Poland, and Spain-Sweden), but not within taxa (Germany-Spain and Poland-Sweden).

Peak regions identified on the basis of elevated F_{ST} were usually associated with a reduction of genetic diversity (π), as well as of D_{xy}, which is largely a function of π. Net divergence, D_a, generally increased in peaks, similar to the pattern of F_{ST} in most cases (see also **table S10A, 10B**). There was generally no significant signal for Tajima's D, Fu and Li's D, or Fu's Fs in either of the taxa, which would have pointed towards recent selective sweeps. In conjunction with elevated levels of linkage disequilibrium and increased haplotype length, this pattern may be most consistent with a scenario of background selection in a region of low recombination caused e.g. by structural variation of genomic features like centromeres or alternatively with repeated, but older positive selection events.

The most extreme 48 outlier windows were located in scaffolds 78 and 60, which according to synteny in chicken, flycatcher and zebra finch genomes, are adjacent and located on chromosome 18 (**Fig. 3, fig. S8A**). These peaks were also accompanied by a marked reduction in nucleotide diversity and differentiation in all populations. However, D_a showed a distinct increase in the

beginning of the peak and a decrease in the latter half of the peak, suggesting that divergence in this region might be oldest or least influenced by shared polymorphism. Most interestingly, the length of haplotype blocks as estimated by the iES statistic increased significantly towards the end of the first scaffold. This increase in haplotype length coincided with minimal values of Tajima's D and with even more extreme minimum of the Fu and Li's D statistic in the Polish and Swedish population, indicating strong recent selection in hooded crows.

Mitochondrial differentiation

Similar to the lack of species differentiation given previous use of mitochondrial DNA (74, 75), the 449 variants detected in the mitochondrial genome did not fully cluster based on population or taxa. A phylogenetic tree reconstructed using the neighbor-joining method in *SplitsTree4* version 4.13.1 demonstrates population structure between carrion and hooded crows, but no clear separation (**fig. S4**). The F_{ST} between carrion and hooded crows was 0.13 and thereby similar to that of the autosomes.

Testing for divergence hitchhiking

Under the model of divergence hitchhiking it is assumed that the peak increases in width under the influence of divergent selection under conditions of gene flow (19). We therefore compared peak width and height between both hooded crow populations with the Spanish (no gene flow) and German population (substantial gene flow) of scaffold 78, respectively. The peak was wider in the comparison of both Poland and Sweden with Spain encompassing 27 consecutive windows, whereas it stretched across 24 consecutive windows in the comparisons of both hooded crow populations with Germany. While peak width is difficult to compare across different levels of background differentiation, reduced width in the comparison between populations connected by gene flow across the hybrid zone is not consistent with a scenario of divergence hitchhiking promoting speciation. It rather suggests stability or slight erosion of the peak margins. Mean and maximum peak height (measured as F_{ST}) were similar, but slightly elevated in the comparisons with Germany (Sweden: 0.514, 0.718; Poland: 0.530, 0.724) than with Spain (Sweden: 0.450; 0.700; Poland: 0.464; 0.720), which might be interpreted as a signal of divergent selection under gene flow acting on loci within the peak. Similar levels of linkage disequilibrium across the entire peak point towards a selection-migration equilibrium.

Cactus-based analyses

Window-based analysis run the risk of averaging out local patterns in the data and thus conceal

regions of interest. In addition to the windows-based population genetic outlier screens, we therefore used a method implemented in the *Saguaro* software that combines a Hidden Markov Model (HMM) with a Self Organising Map (SOM), to characterise local phylogenetic relationships among aligned sequences of all 60 individuals (76). The local relationships, termed 'cacti', are distance matrices built from a tree-like scoring algorithm that best fit the phylogenetic topology for each individual region in the genome, and are computed from the data and without any *a priori* input hypothesis. *Saguaro* was run for ten iterations, after which the set of cacti modelling the entire genome was saturated, and identified ten different cacti that could be grouped into two main classes based on their ability to differentiate the taxa **(fig. S9**). The majority of cacti, eight out of ten, covering approximately 99.7% of the genome covered by cacti (**table S11**), did not adhere to taxon classification: either German carrion crows clustered more closely with both hooded crow populations than with Spanish carrion crows, resembling the genome-wide PCA, or there was complete panmixia among populations, resembling previous findings using neutral genetic markers (55, e.g. 77). By contrast, taxon-specific phylogenic patterns, with clear separation between hooded and carrion crows, were generated by cactus 5 and cactus 6, containing approximately 0.28% of the genome assigned to cacti (**fig. S9**).

6. Inversion detection

Inversion polymorphisms are known to occur in different frequencies in natural populations, and by acting as barriers to gene flow are thought to play an important role in adaptation and speciation (20). Detecting inversions with certainty in population genomic data sets is, however, still challenging. Therefore, we combined several independent lines of evidence to characterize potential inversions that are (near-) fixed variants between carrion and hooded crows. We focused on scaffold 78, which by population genetic parameters (F_{ST}, D_{xy}, D_a, π, iES) has been identified as a strong candidate region for an inversion (see above).

Linkage disequilibrium

We used sites that could be phased into haplotypes in all four populations and screened the entire genome for signs of an inversion using the R package *inveRsion* (78) which implements the inversion-detection model suggested by Sindi *et al.* (79). We parametrized the inversion detection algorithm with a window size of 0.1, block size of 30 and minimum allele frequency of 0.3. While none of the predictions showed a positive BIC score (according to the definition in (78)) which would be taken as strong evidence for an inversion, the top candidates with the 1% highest scores predicted inversions for scaffold 29 and scaffold 78 both corresponding to F_{ST}

peaks. Based on this threshold, the inversion prediction on scaffold 78 was localized to positions 2,043,084-2,047,280.

We further calculated pairwise linkage disequilibrium on scaffold 78 for each of the populations using the *Haploview* program on phased genotypes (80). Linkage disequilibrium was generally higher in peak regions compared to non-peak regions (peak/non-peak for Spain, Germany, Poland, Sweden: D': 0.81 / 0.73 , 0.79 / 0.73 , 0.81 / 0.79 , 0.82 / 0.81, and r^2: 0.12 / 0.06, 0.15 / 0.07, 0.09 / 0.07, 0.08 / 0.06). The difference in peak vs. non-peak regions was significant in all comparisons (Kruskal-Wallis test: p-value < 2.2e-16). Linkage disequilibrium in peak regions was particularly increased in the German and Spanish populations. Scaffold 78 showed an exceptional pattern of linkage disequilibrium in which Spanish and German populations had higher linkage for non-adjacent sites than for adjacent sites (**fig. S10**). Such a pattern of linkage disequilibrium can be caused by an inversion and has been used to predict the breakpoints of inversions in human populations (81). Applying this criterion, we predicted breakpoints at positions 1,883,205, 2,459,610 and 2,494,964 in Spain and breakpoints at positions and breakpoints around 1,888,990, 2,450,595, and 2,481,577 in Germany.

Paired-end / mate-pair read coverage

Information from paired-end reads from re-sequencing data can be used to predict inversions (82). We mapped paired-end reads of all 60 individuals to the hooded crow reference genome using the *bwa* read mapper and re-aligned using *GATK* as described previously. These *.bam* files were used to obtain predictions of inversions from the program *delly* (83). Using all four populations, three smaller inversions were predicted on scaffold 78 between breakpoints at positions [905,147 – 906,125], [2,474,917 – 2,509,736] and [2,337,455 – 2,372,516] (**fig. S10**).

We further calculated paired-read coverage for each of the individuals to identify potential breakpoints. Regions with zero paired-end coverage in carrion crows but non-zero paired-end coverage in hooded crow individuals suggest an inversion in carrion crows prevents mapping at breakpoints and provides potential confidence intervals around the breakpoints. This approach predicted breakpoint ranges between positions 2,012,307 – 2,012,321 as well as between 2,387,613 – 2,387,626, broadly corresponding to a large inversion inferred from the linkage disequilibrium based tests.

Mate-pair libraries with a mean insert size of 2.5Kb were constructed from a carrion crow

individual from the Spanish population. It was sequenced to a sequence coverage of 37X using Solid sequencing platform. Reads were mapped in color-space to the genome using *BioScope*™ 1.3.1 (SOLID™) and separately using *bwa* 0.7.4 (28) in both color-space and base-space. Paired-coverage of the mate-pairs was calculated for each base to identify potential inversion breakpoints. Regions with zero paired-end coverage suggest an inversion in carrion crows relative to the hooded-crow assembly. This approach predicted breakpoint ranges between positions [1,717,096 – 1,717,203], [2,482,232 – 2,482,367] and [2,484,608 – 2,484,660]. Based on the same mate-pair data *delly* 0.3.2 (56) predicted three inversion breakpoints along scaffold_78 at positions 2,339,649, 2,474,473 and 2,477,663.

Combining the evidence from the above mentioned approaches, we tentatively predict one major inversion of around 400 kb with breakpoints near 2,000,000 and 2,400,000 potentially accompanied by a larger inversion encompassing the entire peak and smaller nested smaller inversions (**fig. S10**).

F_{ST} based prediction

Inversions are expected to negatively impact inter-population gene flow and leave a signature of increased differentiation, especially around the inversion breakpoints (11). Accordingly, increased values of F_{ST} near breakpoints have been used to weed out false positives obtained from other methods (84). In our case, we observed two peaks of elevated F_{ST} on scaffold 78, that were connected by a saddle of slightly decreased differentiation. The center of the right peak coincides with a predicted region of an inversion breakpoint (**fig. S10**). Nearly all fixed differences between taxa (81/83) were located in distinct clusters, and the cluster of hooded-derived variants coincided with one predicted inversion breakpoint. Such a pattern has previously been described to be associated with an inversion polymorphism in *Drosophila* (85).

Ancestry polarized haplotype runs

We looked at the ancestral state of each of the fixed differences identified on scaffold 78 and the adjacent scaffold 60 in Rook, Jackdaw and American crow genomes. Most of these sites were monomorphic (66 of 81) in the ancestral samples and were largely consistent across all 3 outgroup species (62 of 81). In the first half of the prominent peak on scaffold 78, most of the fixed sites (50 of 81) up to 1.9 Mb had an ancestral state resembling the hooded crow followed by a run of fixed sites (18 of 81) that had an ancestral state resembling the carrion crow. Such a pattern of correlated haplotype runs of ancestral / derived states has been seen in the inversion on

human chromosome 17q21 and has been interpreted as a region prone to repeated inversions (86).

7. Common garden experiment

Animal husbandry

Crow nestlings were obtained from two populations of each taxon during May-June 2010 as described above (see also map in **Fig. 1**). For the gene expression study, only a single male individual was sampled per nest. In one case of two German individuals, using siblings could not be avoided. The choice of only using males was made in order to focus on gene expression differences between populations without having to account for additional variance introduced by sex-specific gene expression.

Nestlings were hand-fed at the sampling location until all samples had been obtained and then transported to a large bird keeping facility at the Max Planck Institute for Ornithology in Radolfzell in Germany either by car (German, Spanish, Polish populations) or by plane (Swedish population). The crows were initially hand-raised in cardboard boxes and at around 6 weeks of age moved to $12m^2$x2.5m aviaries where they were raised from June-October 2010 in taxon specific groups, two to three crows each. Drinking water and suitable food was available *ad libitum*. The animals were all fed the same diet including grain, a protein-rich food (cat food and cottage cheese), vegetables and fruit. The crows were regularly weighed, measured, and treated with Praziquantel against tapeworms and Fenbendazol against roundworms. Between July 25 and August 5, 12 crows died of what was eventually diagnosed as coccidiosis; from then on, the remaining crows were also treated with Toltrazuril against coccidiae.

Permission for keeping of crows and the experimentation described below was granted by *Regierungspräsidium Freiburg* (Aktenzeichen 35-9185.81/G-10/23).

Tissue sampling

One of the main goals of this study was to monitor gene expression activity in feather follicles at an early stage where melanin is produced in melanocytes, and deposited into keratinocytes. Obtaining sufficient material of homogeneous skin samples with feather follicles at a comparable developmental stage is challenging during natural moult, when feathers are shed and grown at different times for different parts of the the body. We therefore stimulated synchronized feather regrowth after the completion of the first juvenile moult by plucking feathers on 3x3cm patches on the central back and belly, and on the crown and throat. To reduce stress and pain this was

done under inhalational anesthesia using isofluran.

We then allowed feathers to regrow for 10-13 days (six days for one individual with particularly fast regrowth), at which stage the previously plucked areas of skin contained densely spaced feather shafts with the first parts of feathers about to or just protruding from the shafts. In a subset of individuals we confirmed that the regrowing feathers were identical in colour and shape to the original feathers. After this period of regrowth, a total of 19 crows (six from Germany, two from Spain, seven from Sweden and four from Poland) were euthanized by desanguination after preceding anesthesia by stroke. Upon death, tissue samples were excised, cut into small pieces, immediately submerged into RNA*later* (Ambion, Inc., Austin, Texas), and subsequently kept at -20°C. The following tissues were sampled: skin from back, skin from belly, skin from crown, skin from throat, gonads, eye, forebrain, pituitary and hypothalamus (brain), cerebellum (brain), liver, heart, and spleen.

8. Gene expression analysis

mRNA sequencing

For each sample, a total of 20-40 mg tissue was homogenized by mechanic disruption (TissueRuptor, Qiagen) in 1 ml TRIzol® reagent (Invitrogen). Total RNA was then extracted using either the Qiagen RNeasy Mini Kit or RNeasy lipid tissue kit with the inclusion of an on-column DNA digestion step (RNase free DNase I, Qiagen). All extractions used for sequencing yielded sufficient amounts (>1 µg) of high quality RNA (mean RIN integrity number 9.59) for subsequent library preparations. Tissue-specific libraries were prepared for all 19 individuals for liver, forebrain, gonads and skin (with RNA extraction failing for one sample: gonads in one Swedish individual). Skin RNA from throat and crown was pooled at equal proportions and combined in one library "skin from head" (black in both taxa), and skin from belly and back was pooled as "skin from torso" (black in carrion and grey in hooded crows, see **Fig. 1**). Libraries for heart, eye, cerebellum, pituitary/hypothalamus, and spleen were produced for a subset of individuals (see **table S3**).

All 120 libraries (insert size: ~180 bp) were prepared using Illumina's truSeq RNA Sample Prep Kit v1 according to the manufacturer's instructions. This protocol includes a poly-A mRNA enrichment step and cDNA preparation through random hexamer primed reverse transcription followed by ligation of barcoded adapters and PCR amplification (12-15 cycles). The individually barcoded libraries were pooled during the sequencing step, with 12 libraries being

sequenced per lane on an Illumina HiSeq 2000 machine. Care was taken to spread tissues and taxa across lanes and positions within lanes (blocked statistical design), to avoid lane effects with the potential to bias the results. Read depth coverage is given in **table S3**.

Sequence data processing

RNA-seq paired end reads from each of 120 libraries were mapped to the hooded crow reference genome using the *tophat (version 2.0.2)* read mapper. Subsequently, read counts per gene and transcript were obtained by running *RSEM* (87) (version 1.2.0) on the sequence alignment files using the '--no-polyA' option, since strand information was not available for a number of transcripts.

Statistical analyses

In total, 42,014 crow genes (i.e. 86 less than for the entire RNA-seq data set, see above) were found to be expressed among our main tissues of interest (two types of skin, forebrain, liver, gonads). 28,411 of these were expressed in all five tissues, and 1,759 were exclusively expressed in one tissue. The number of expressed genes in any single tissue varied between 30,905 (for liver) and 40,429 (for gonads).

Differential gene expression was inferred using the *R/Bioconductor* packages *edgeR* (88) and *EBseq* (89). For each of the five main tissues, we tested for differential expression between carrion and hooded crows. Gene counts from *RSEM* were normalized for differences in library sizes and RNA composition with the “calcNormFactors()” function in *edgeR* which uses the trimmed mean of M-values (TMM) method (90) and with the “QuantileNorm()” function in *EBseq*, which uses Upper-Quantile-Normalization (91). Since biological replicates were present for all analysis, variance was estimated on a tagwise (gene-by-gene) basis. This is done automatically in *EBseq* when executing the EBTest() function, whereas in *edgeR* this was done by first estimating dispersion first for all genes simultaneously (common dispersion, “estimateCommonDisp()” function), and then separately for each gene or transcript unit (tagwise dispersion, “estimateTagwiseDisp()” function). For all analyses, nominal p-values were adjusted according to the Benjamini & Hochberg (92) method, using as a threshold for significance a false discovery rate (FDR) of 0.05. Results from *edgeR* and *EBseq* were combined such that genes were considered differentially expressed only when the FDR was <0.05 according to both approaches.

Candidate genes from the melanogenesis pathway

We have previously screened the literature on existing information on genes involved in processes related to melanogenesis and compiled a list 95 prime candidate genes that may explain colour variation in crows (55).

9. Immunohistochemistry

Differences in grey and black plumage colouration in hooded and carrion crows are caused by differences in eumelanin pigment concentration (93). Eumelanin is synthesized in specialized organelles of the melanocytes, the melanosomes, from where it is transported into keratinocytes. Keratinocytes represent the most abundant cell type in feather follicles and produce the fibrous structural protein keratin for the mature feather.

Theoretically, differing eumelanin contents could result from differences in: the activity of the melanogenesis pathway, melanosome transport or melanin deposition in keratinocytes, melanocyte abundance or a combination thereof. Our RNA-seq results of feather follicles containing active melanocytes (see above) showed clear differences in gene expression levels in the melanogenesis pathway between grey and black feathers, excluding a sole defect in melanin deposition as cause of the colour difference. However, tissue based RNA-seq data integrates across many cells and does therefore not allow to distinguish between increased expression levels due to pathway upregulation or increased cell density, in our case, melanocyte- density. To assess differences in pathway activity and melanocyte abundance, we performed comparative whole mount immunohistochemistry on matched feather follicle samples from carrion and hooded crows with an antibody against an enzyme of the melanogenesis pathway (tyrosinase) identified as differentially expressed in our RNA-seq results. Briefly, feather buds of comparable developmental stage were dissected out from PFA-fixed tissue samples from the backs of carrion and hooded crows, extensively rinsed in PBS and split along their longitudinal axis. Grey and black feathers were always processed in parallel and staining performed as follows: blocking (10% FBS in PBS, 1% Triton) for two days, primary antibody: anti-trp (abcam83774, 1:100, in blocking solution as above) for two weeks, washing in PBST for one week, secondary antibody: anti-rabbit 488 (Invitrogen, 1:200 in PBST) for two days, washing in PBST+ DAPI for 2 days; specimen were then transferred into Mowiol + 2.5% w/v DABCO and imaged on a Leica CTR 6000 microscope.

For immunohistochemistry on free-floating sections, carrion and hooded crow feathers matched

for length and developmental stage were embedded in OCT and cryo-sectioned (25 um sections). Sections were collected in PBS, incubated in blocking solution (see above) for 1 day, incubated with blocking solution with primary antibody (same as above) for 3 days, rinsed in PBS for 1 day, incubated with secondary antibody (same as above) over night at 4 degree, rinsed and mounted for imaging on a Leica SP5 confocal microscope. Image post-processing (pseudo-colouring, linear adjustment of contrast and brightness across the entire image, scale bars) was performed with the Adobe PhotoShop suite.

Supplementary Figures S1-S13

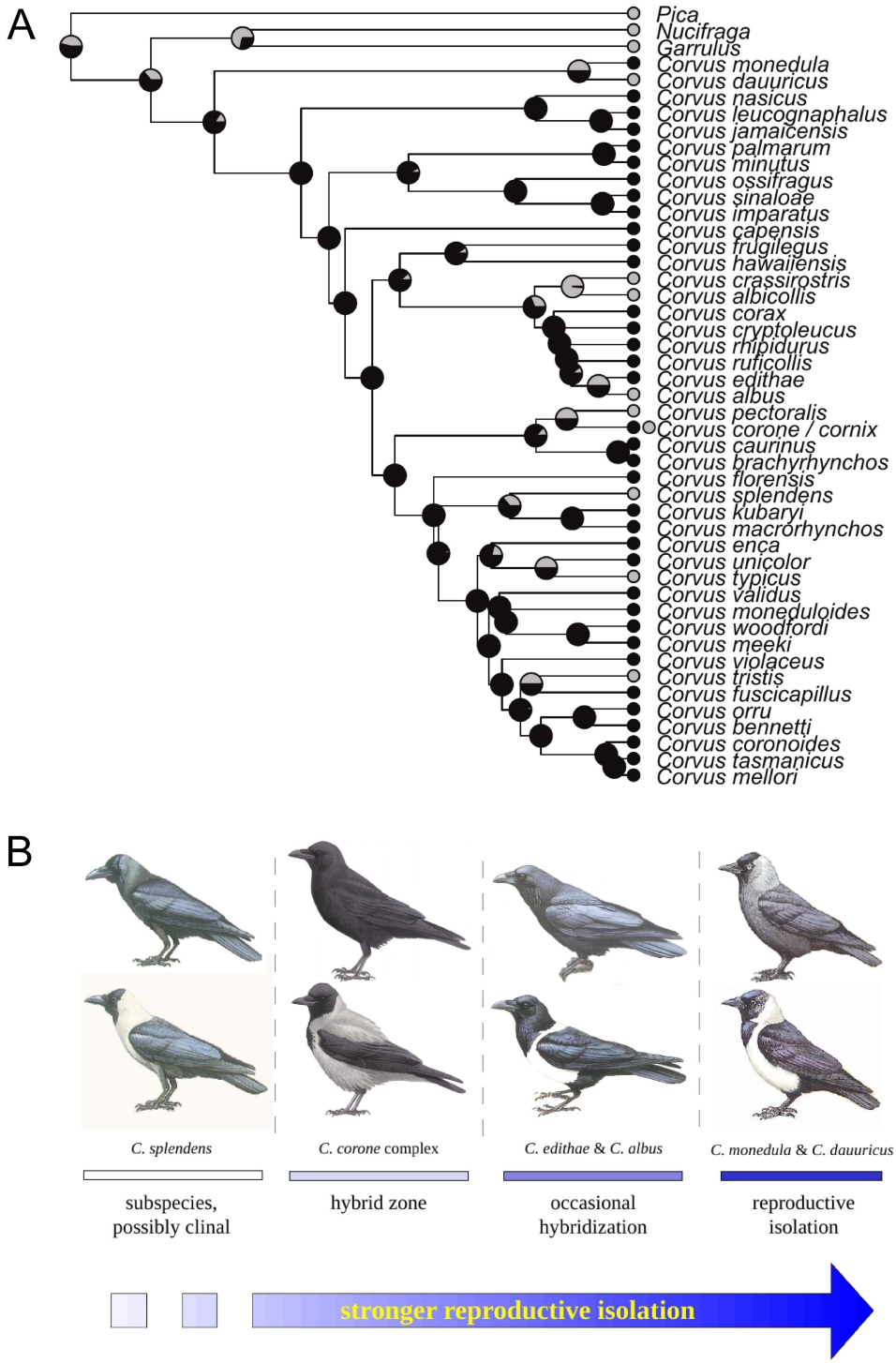

Fig. S1. (A) Most of the approximately 40 species in the genus *Corvus* are all black (black circles), but several species have partly grey or white plumage (grey circles, see panel B). Six of these species show pied plumage with grey or white dorsal and ventral patches on an otherwise

black background (C. *dauuricus*, *C. albus*, *C. pectoralis*, *C. (corone) cornix*, and *C. splendens*), all in a similar fashion to hooded crows. One species (*C. tristis*) is almost entirely grey, while the two remaining species have striking white neck-patches only (*C. crassirostris*, *C. albicollis*). The phylogenetically independent recurrence of this pattern reconstructed under a maximum likelihood framework (94, 95) suggests parallel evolution and hints at a common, simple genetic architecture. **(B)** These visual signals contribute to prezygotic isolation and have prompted the hypothesis to promote speciation via sexual/social selection (25, 96). Interestingly, several pied and black (sub-)species pairs can be placed along a conceptual line of increasing reproductive isolation. In the house crow (*C. splendens)*, colour phenotypes can segregate along broad-scale clines and as polymorphisms within populations. Grey-coated hooded crows (*C. (corone) cornix*) and all-black carrion crows are a textbook example of incipient speciation along narrow hybrid zones in Europe (*C. (corone) corone*) and Russia (*C. (corone) orientalis)* and are the subject of this study. The pied crow (*Corvus albus*) is widespread in Africa and hybridizes with the dwarf raven (*Corvus edithae*) in parts of its range. Finally, hybridization between Daurian and Eurasian jackdaws (*C. dauuricus and C. monedula*) has only rarely been recorded. *Phylogenetic tree adapted from Jønsson et al.* (97). *Crow* i*mages with permission from Bloomsbury Publishers.*

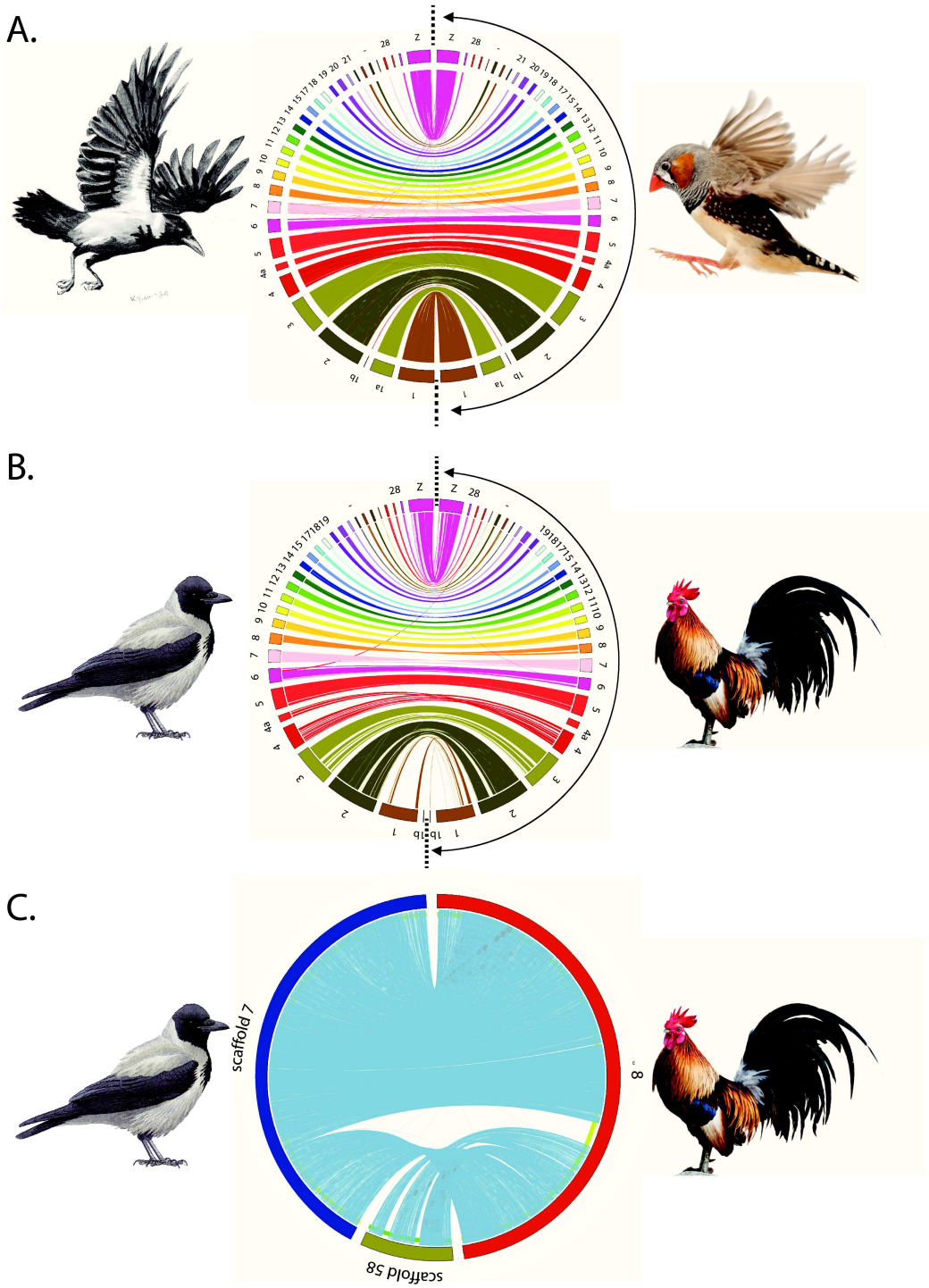

Fig. S2. Chromosomal synteny plot between hooded crow and **(A)** zebra finch and **(B)** chicken genome assemblies. The 100 largest scaffolds were aligned to chicken with *Satsuma*. **(C)** Example for a rearrangement between hooded crow and chicken on chromosome 8. *Chicken image by courtesy of Sattapapan Tratong/123RF. Zebra finch image by courtesy of Eric Isselee1/123RF. Drawings by courtesy of Dan Zetterström.*

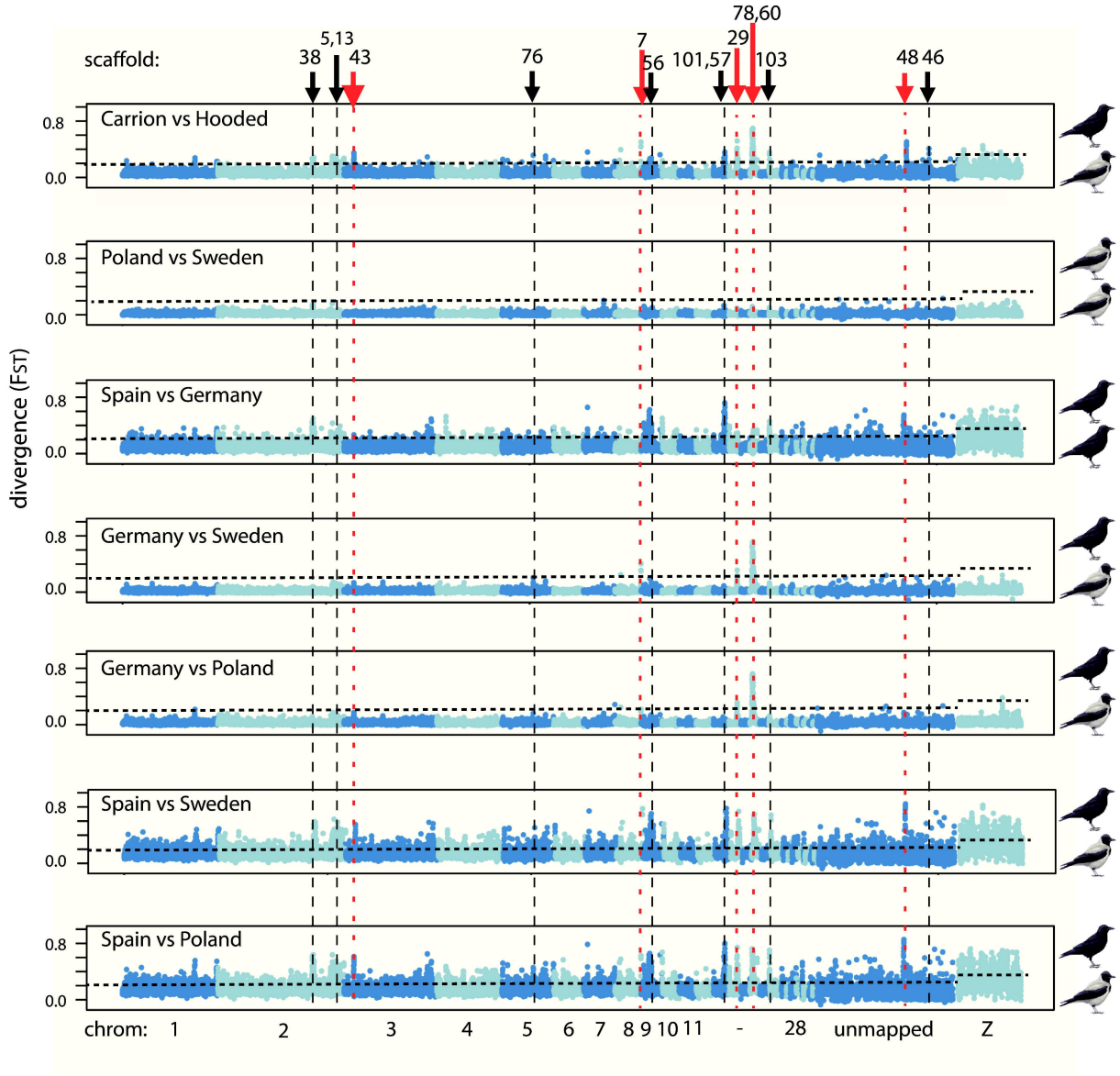

Fig. S3. Manhattan plots of genome-wide pairwise F_{ST} estimated in 50-kb sliding windows between taxa (top panel), between carrion crow populations (second panel), between carrion crow populations (third panel), and between carrion and hooded crows in all possible combinations (following four panels). Alternating colours paint the different chromosomes, the dotted horizontal lines mark the 99th percentile F_{ST} estimates of autosomes and Z chromosome from the taxon-specific comparison, respectively. The red arrows indicate taxon-specific peaks of consecutive elevated F_{ST} windows, while the black arrows highlight outlier peaks that are specific to comparisons with the refugial Spanish population.

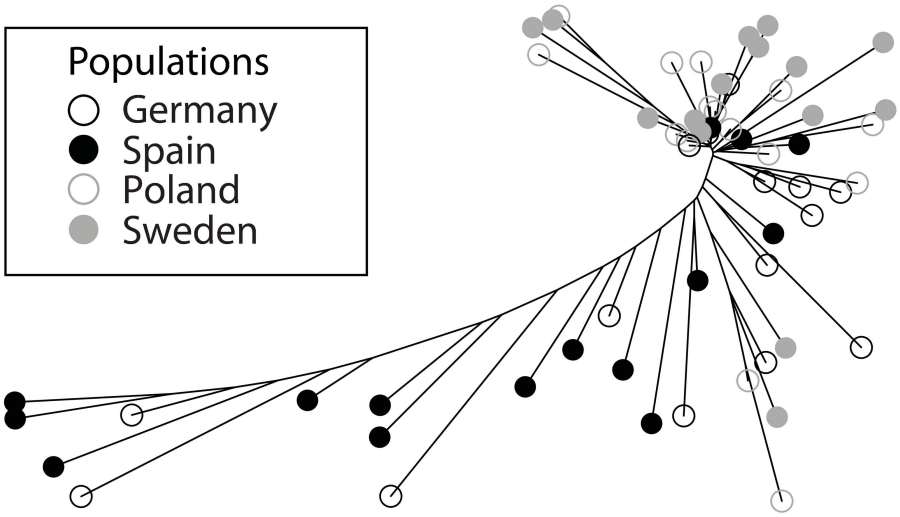

Fig. S4. Phylogenetic neighbor-joining tree reconstructed for all 60 individuals from the 449 SNPs detected in the mitochondrial genome.

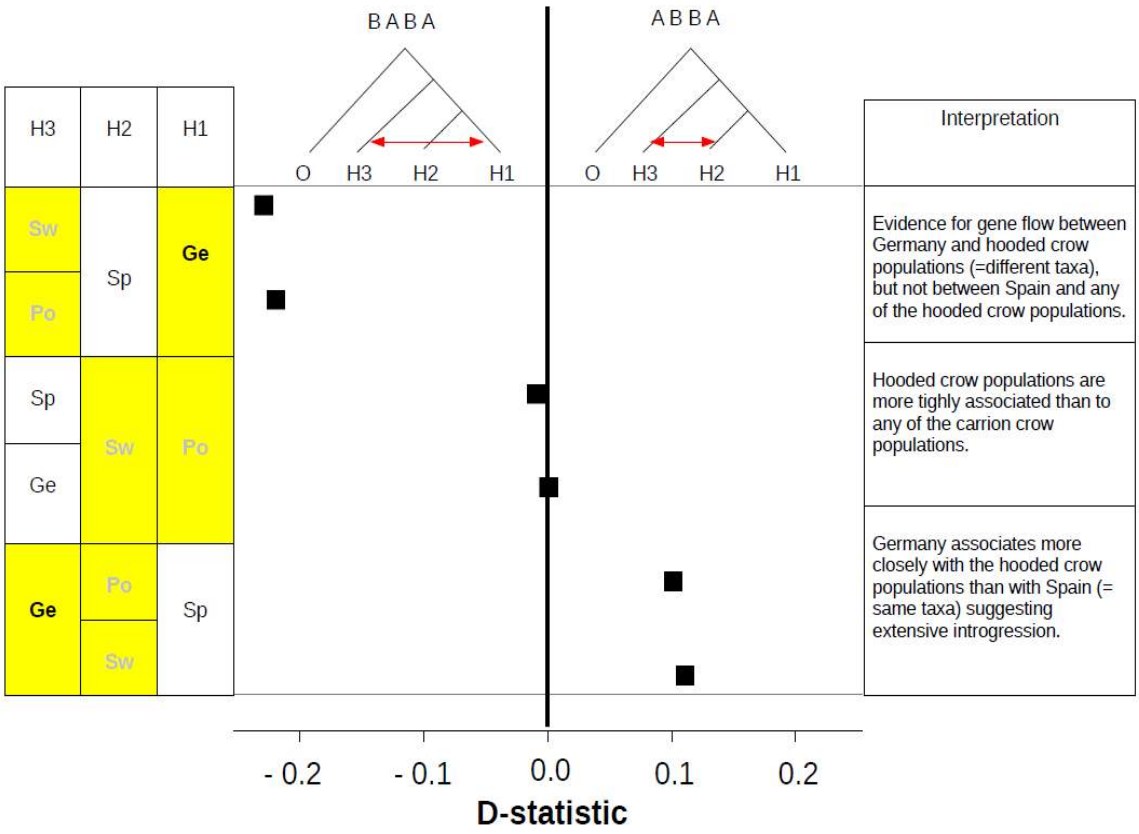

Fig. S5. Visual representation of the decisive ABBA-BABA hypotheses to test for gene flow between carrion and hooded crow populations. Negative values of the D-statistic show that the BABA pattern occurred more often than the ABBA pattern, which is often interpreted as gene flow between populations H1 and H3 (arrow). Positive values of the D-statistic show that the ABBA pattern occured more often than the BABA pattern, supporting gene flow between H2 and H3. When testing for gene flow between hooded and carrion crows with the topology [(O(Hooded crow(Spain:Germany)))], the BABA pattern was found to occur more often than the ABBA pattern. This provides evidence for gene flow between German carrion crows and hooded crow populations (i.e. between the different taxa), but not between Spanish carrion crows and any of the hooded crow populations. When the [(O(Carrion crow(Sweden: Poland)))] topology was used, the ABBA and BABA patterns were not found to be significantly different. Hence, hooded crow populations were more tightly associated with one another than with any of the carrion crow populations. When the [(O(Germany(Hooded crow: Spain)))] topology was used, the ABBA pattern was found to be more common than the BABA pattern. This suggests that the German carrion crow population associates more closely with any of the hooded crow populations than with Spain (=same taxa), in turn suggesting extensive introgression. For all possible population comparisons, see **table S6**.

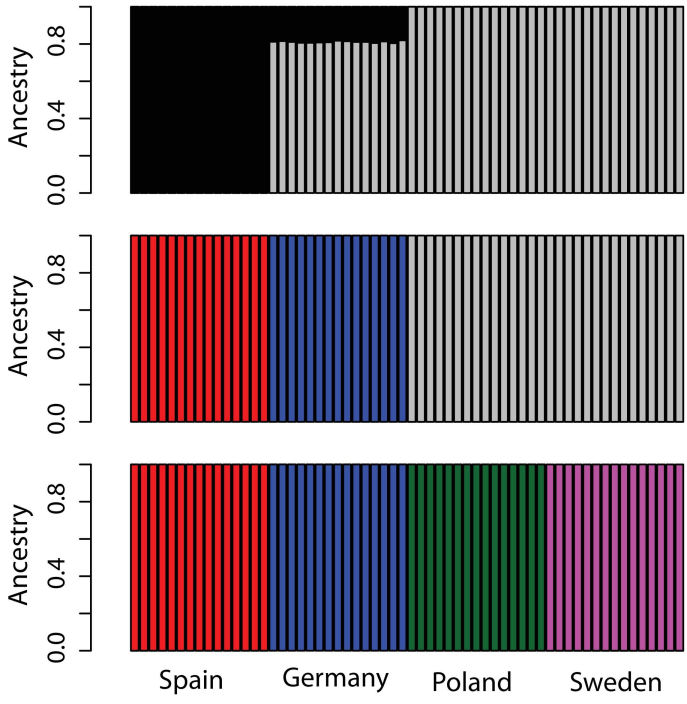

Fig. S6. Genome-wide admixture analyses inferred from 8.4 million SNPs analyzed in ADMIXTURE 1.23 (69). Cross validation methods indicate the optimal number of genetic clusters (illustrated in different colors) is two, suggesting the German carrion crows are a mixture of Spanish carrion crows and (Polish and Swedish) hooded crows.

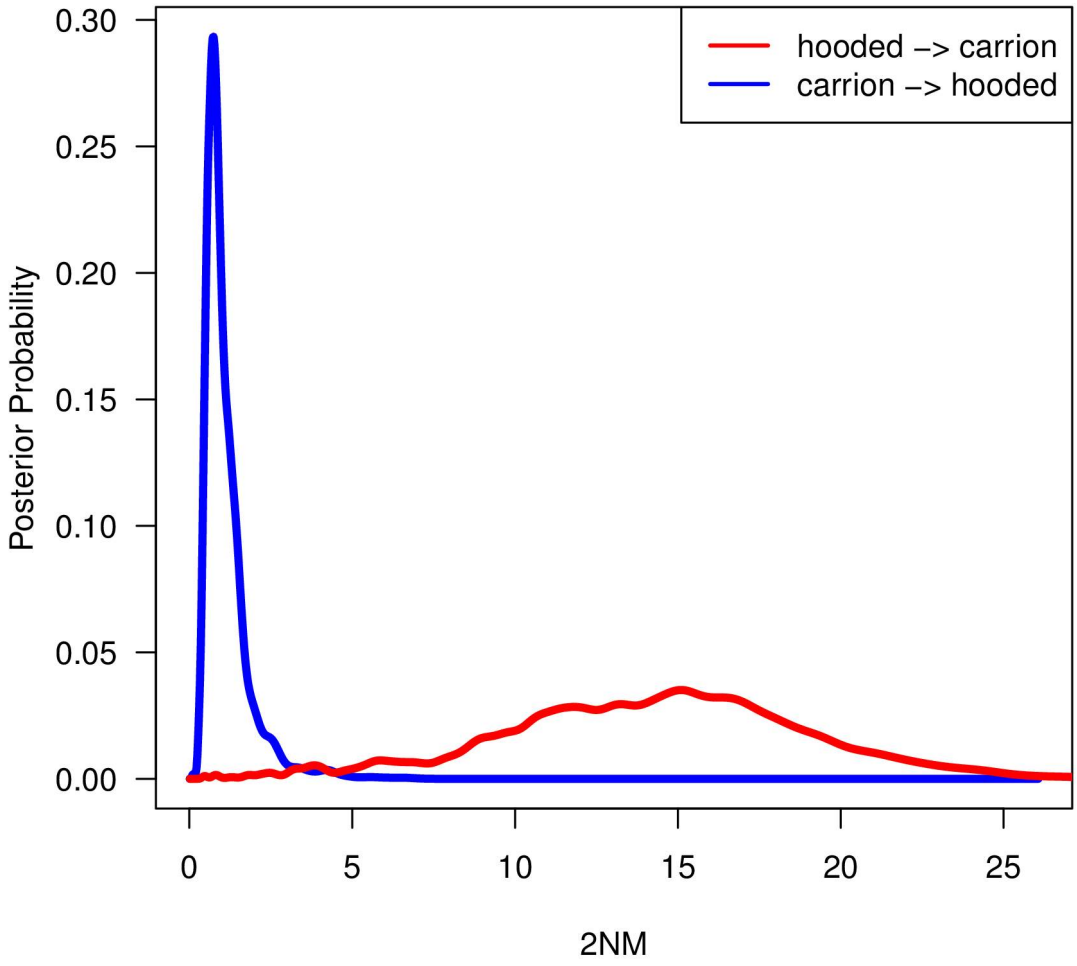

Fig. S7. Inference of levels of gene flow between German carrion and Polish hooded crows using the program *IMa2*. Gene flow is expressed as the population migration rate, 2NM. Highest posterior probabilities were found for 2NM = 1.49 from carrion to hooded crows and for 2NM = 30.20 from hooded to carrion crows. 95% confidence intervals for the *m* parameter (i.e. m divided by the mutation rate) that is directly estimated by *IMa2* exclude zero in both directions.

Fig. S8. Distribution of population genomic parameters along (**A**) scaffolds 78 and 60 on chromosome 18, (**B**) scaffold 29 on chromosome 15, (**C**) scaffold 7 on chromosome 8, **(D)** scaffold 43 on chromosome 3 and **(E)** an unmapped scaffold 48. At the top, examples of class I cacti (differentiating taxa) and class II cacti (not differentiating taxa) are given. The top panel shows the distribution of cacti along scaffolds for class I cactus 5 (red) and cactus 6 (blue), and class II cacti (black). The remaining panels show average values per 50 kb window for the following statistics: (from top to bottom): F_{ST} between taxa, total sequence divergence between taxa (Dxy), proportion of shared polymorphisms among sites polymorphic in at least one taxon (Da). And for each population: nucleotide diversity (π), Fu and Li's D, the integrated extended haplotype homozygosity statistic (iES). Population-specific estimates are drawn in red for Spanish carrion crows, in blue for German carrion crows, in green for Polish hooded crows, and in magenta for Swedish hooded crows.

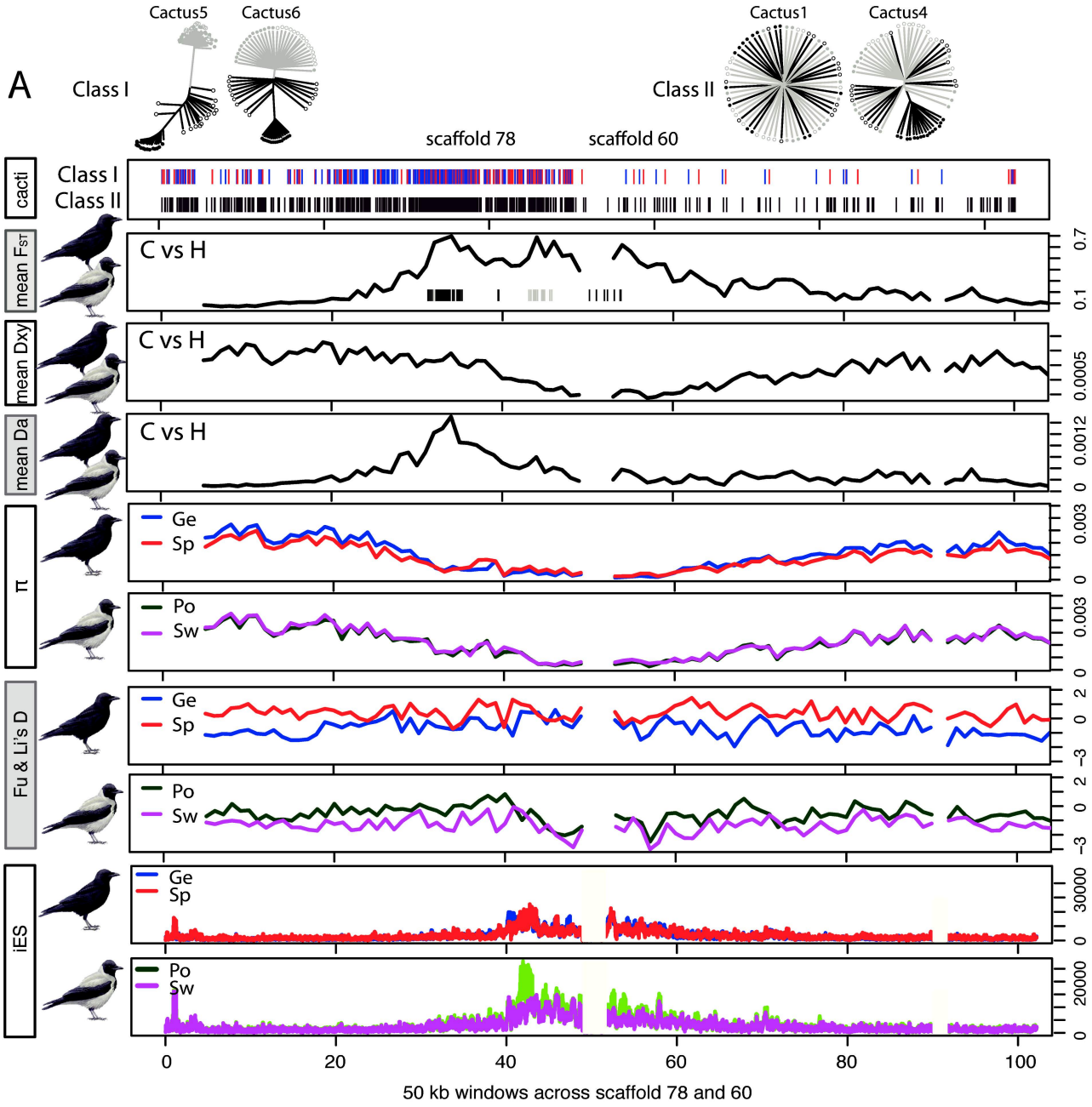

Fig. S8A. Scaffold 78 and 60 on chromosome 18

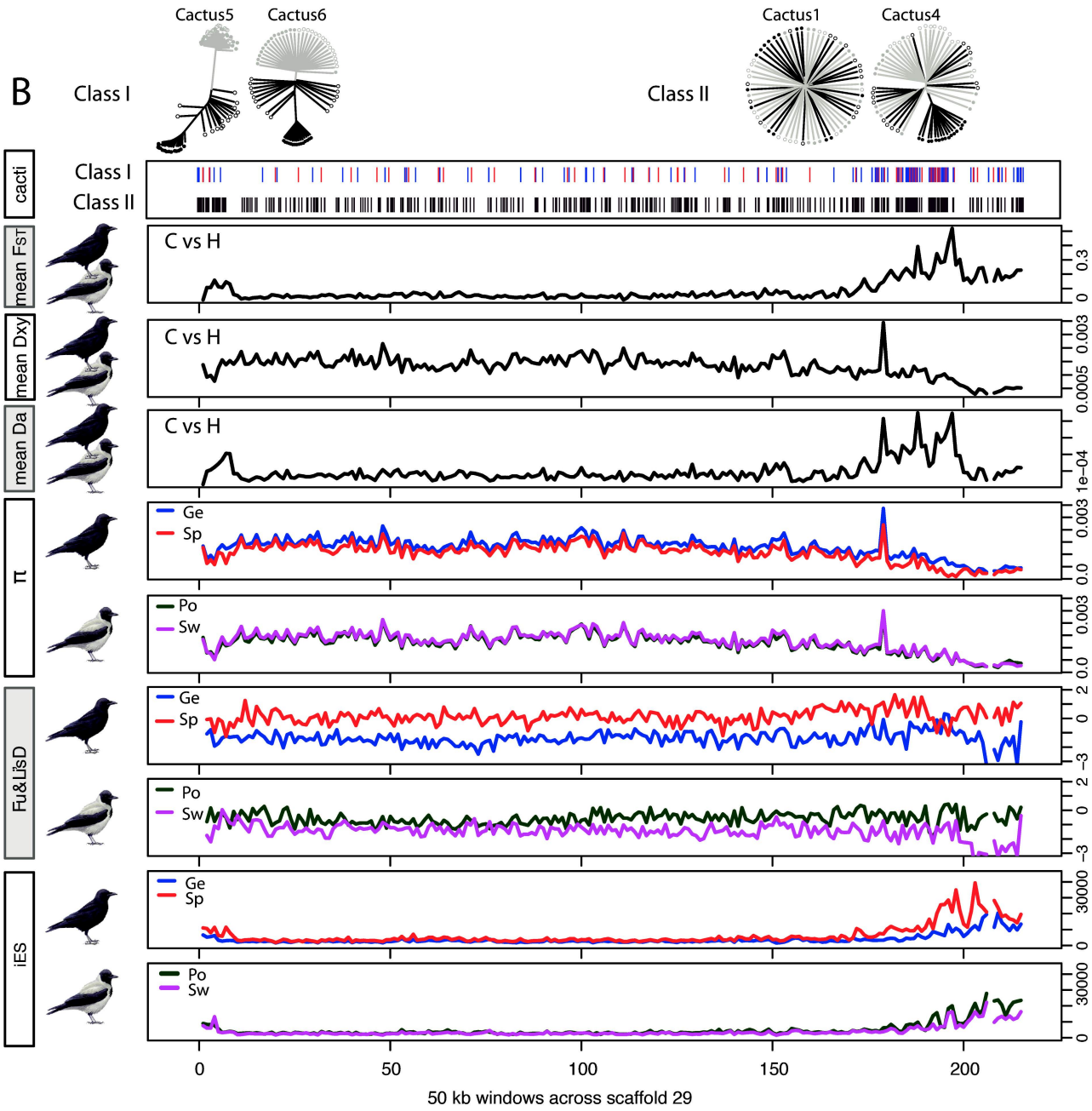

Fig. S8B. Scaffold 29 on chromosome 15

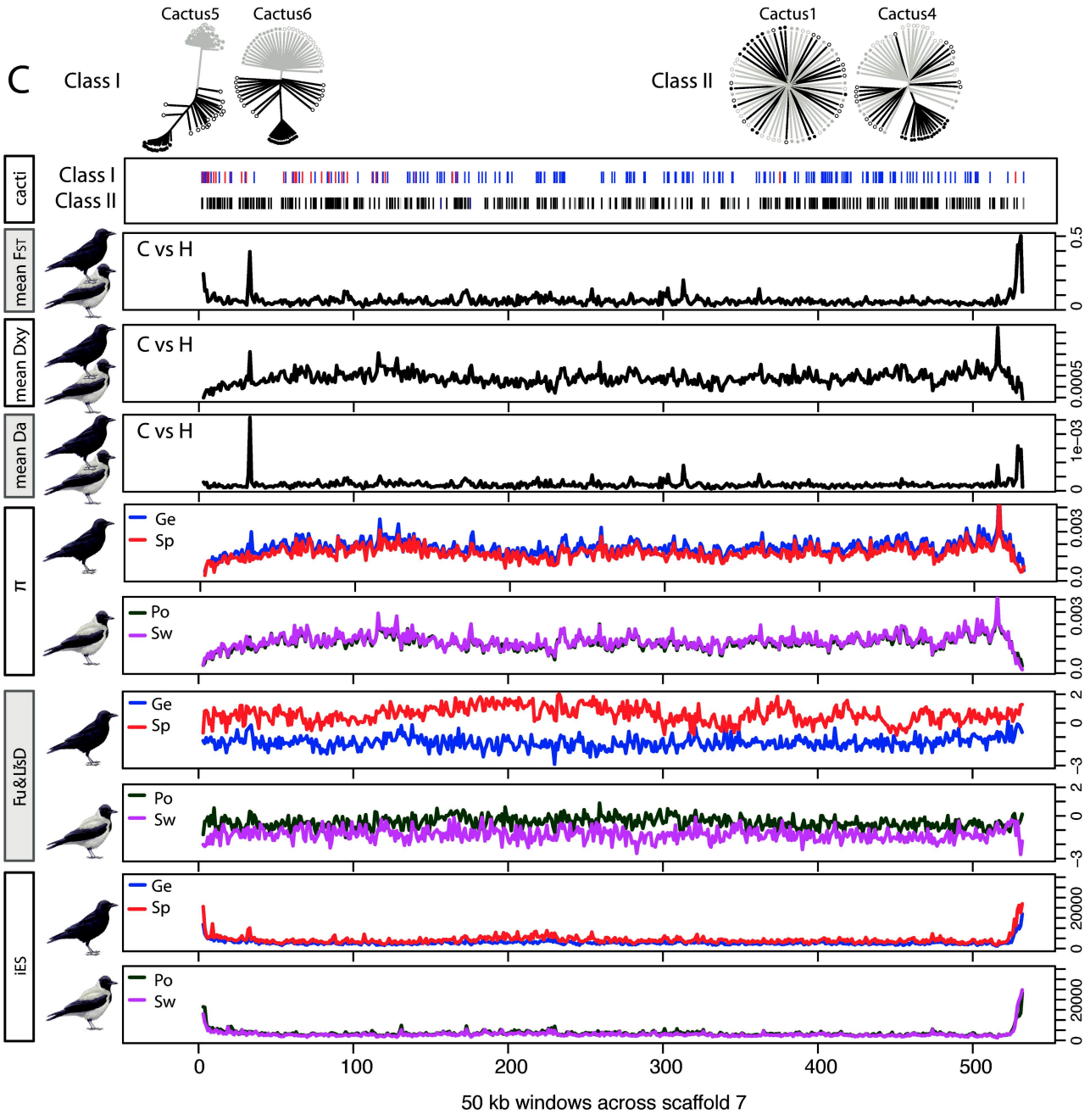

Fig. S8C. Scaffold 7 on chromosome 8

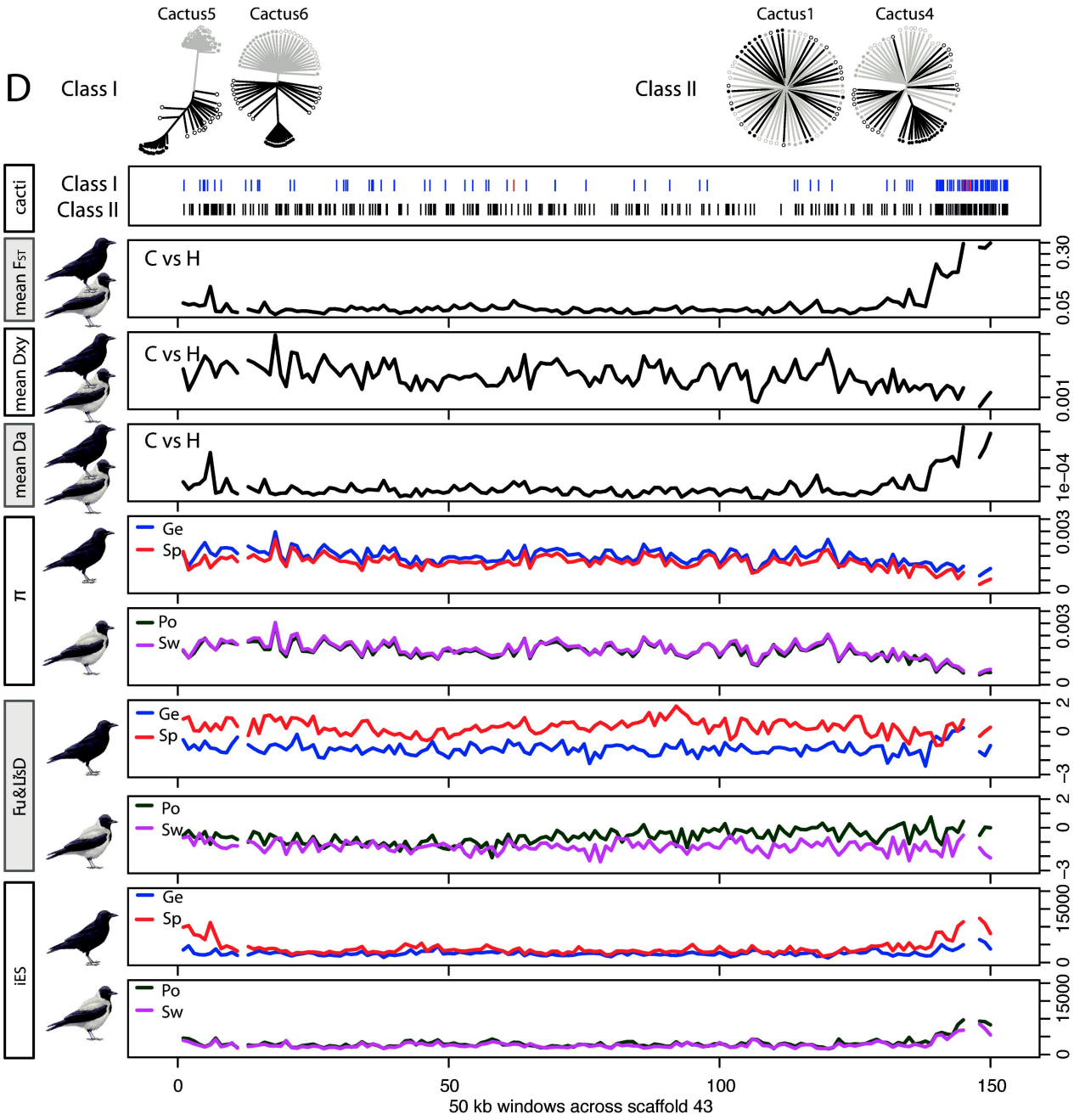

Fig. S8D. Scaffold 43 on chromosome 3

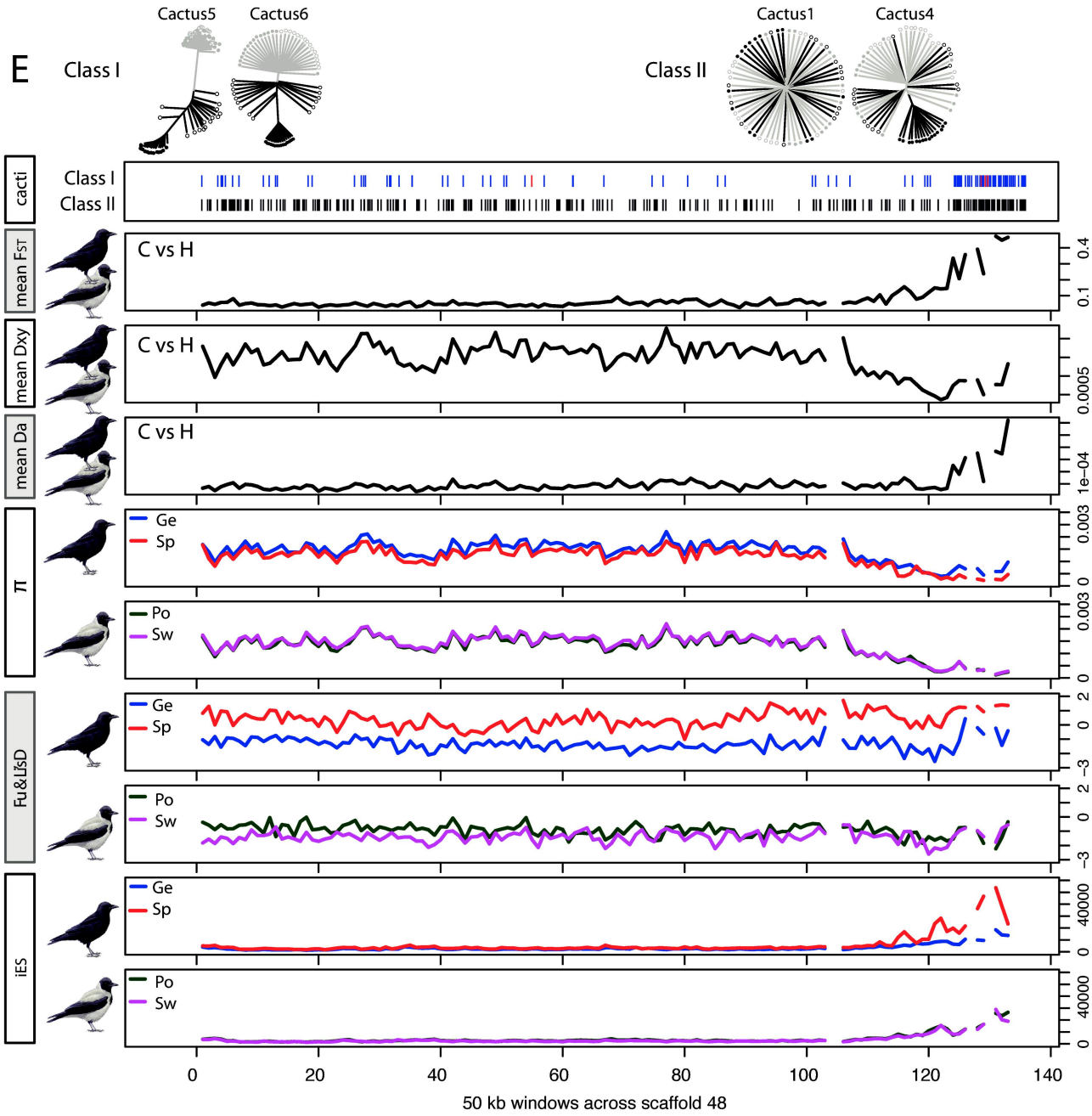

Fig. S8E. Scaffold 48 for which no chromosome could be confidently assigned.

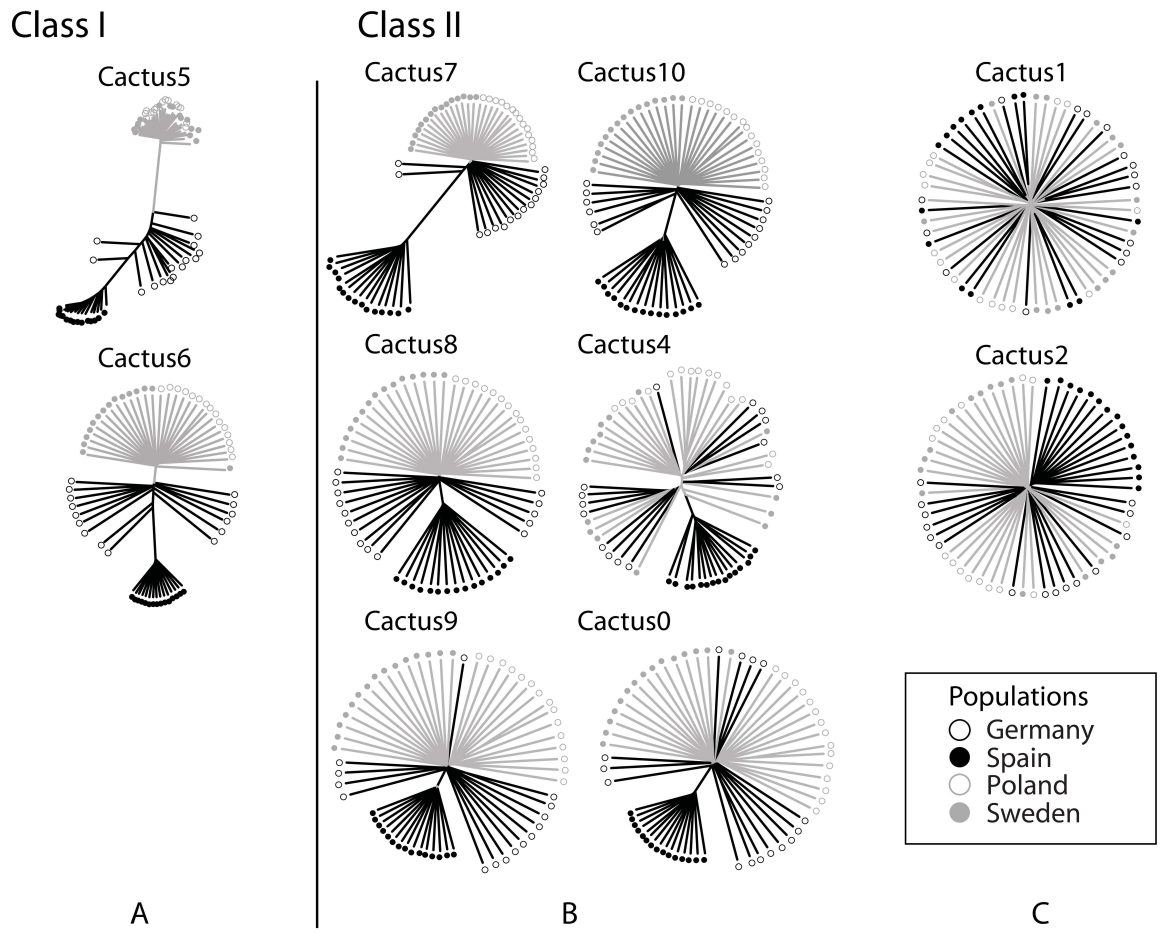

Fig. S9. Genomes were analyzed using the SOM-HMM based method implemented in *Saguaro* to identify localized phylogenetic patterns within the genome (called cacti). (**A**) Two taxon-specific cacti (class I cacti) were detected, covering approximately 0.28% of the genome. The remaining part of the genome generated two broad types of class II cacti: (**B**) a pattern with a divergent Spanish population, and (**C**) a pattern of complete admixture among populations.

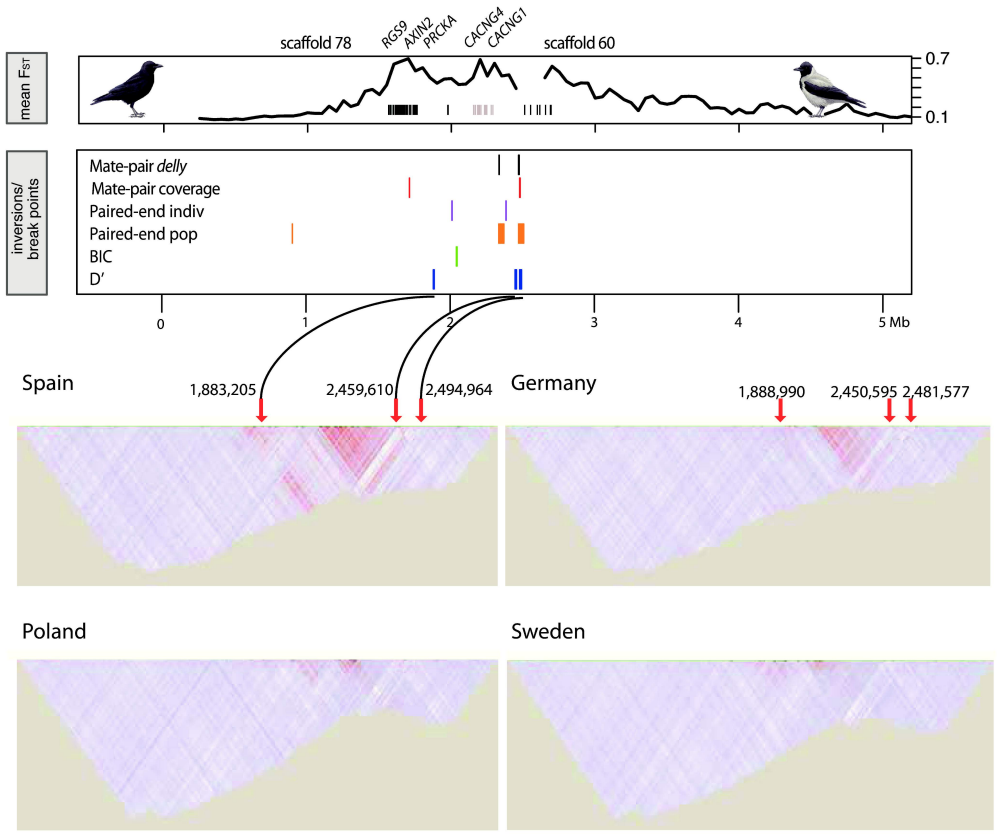

Fig. S10. (A) Upper panel: Distribution of the F_{ST} statistic in 50 kb windows across scaffold 78 and 60 on chromosome 18. Fixed differences between taxa are indicated as lines below the peak region; grey indicates hooded derived fixed differences. Mid-panel: Summary of inversion breakpoint prediction from a variety of approaches on scaffold 78 (see Supplementary Methods). **(B)** Linkage disequilibrium heatmap plots drawn for a subset of the region in (A) showing pairwise D' for part of the scaffold 78 with elevated F_{ST} for each of the four populations based on phased SNP data. Suspected inversion breakpoints are denoted by arrows.

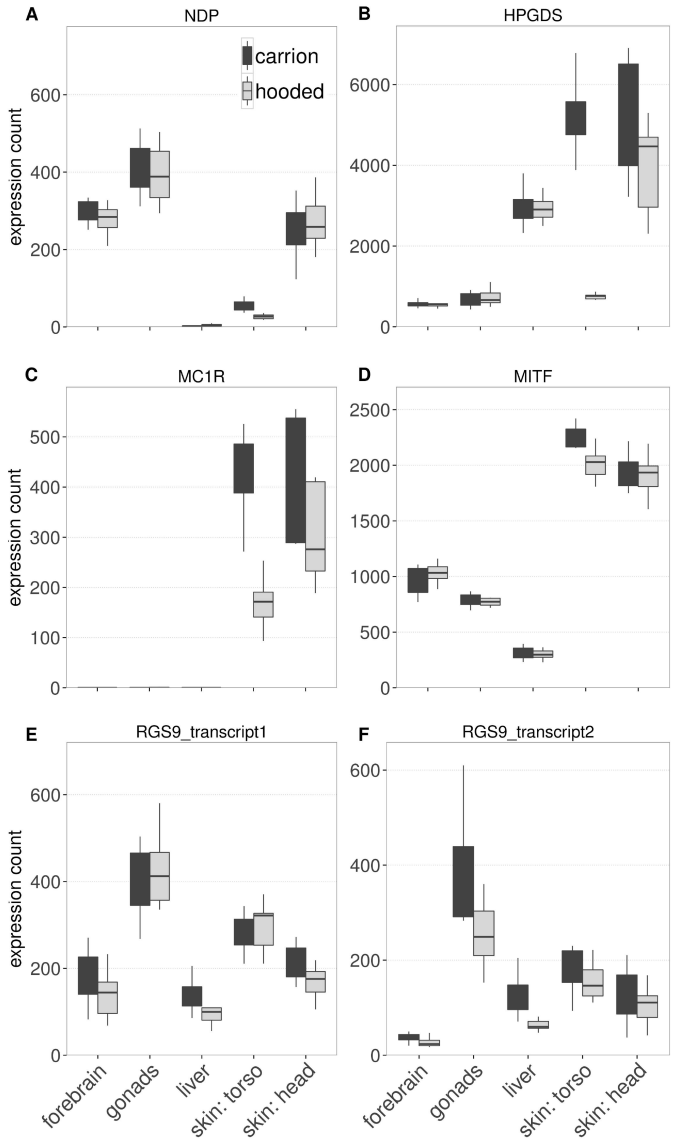

Fig. S11. Examples of gene expression profiles across tissues for a set of six genes. **(A)** norrrin, **(B)** hematopoietic prostaglandin D synthase, **(C)** melanocortin 1 receptor, **(D)** microphthalmia-associated transcription factor, **(E and F)** regulator of G-protein signaling 9.

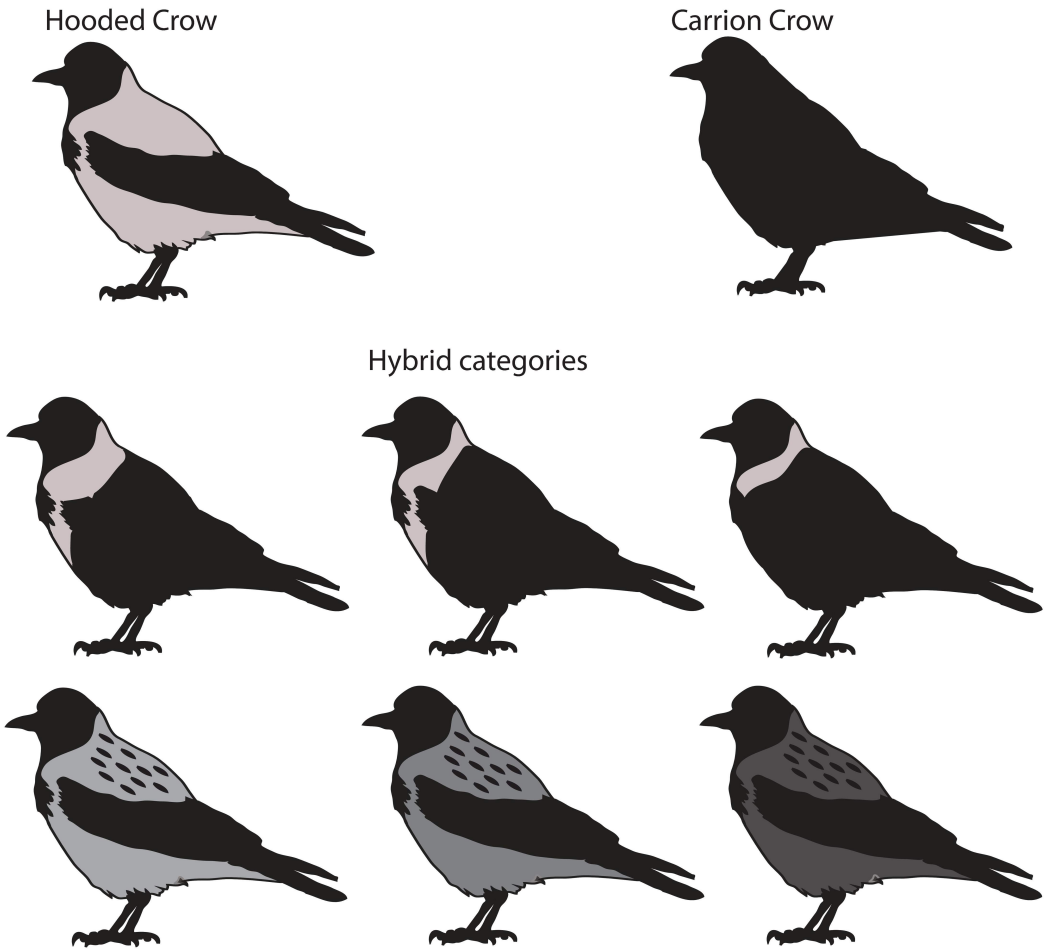

Fig. S12. Plumage characteristics of pure hooded and carrion crows (above) and a series of categorical hybrids from the hybrid zone. Hybrid categories are idealized types and blend into one another. Hybrid types include “grey neck” hybrids with varying size of the grey collar and “diffuse” hybrids that show gradation to darker forms, sometimes with mottling (dark and gray feathers intermixing).

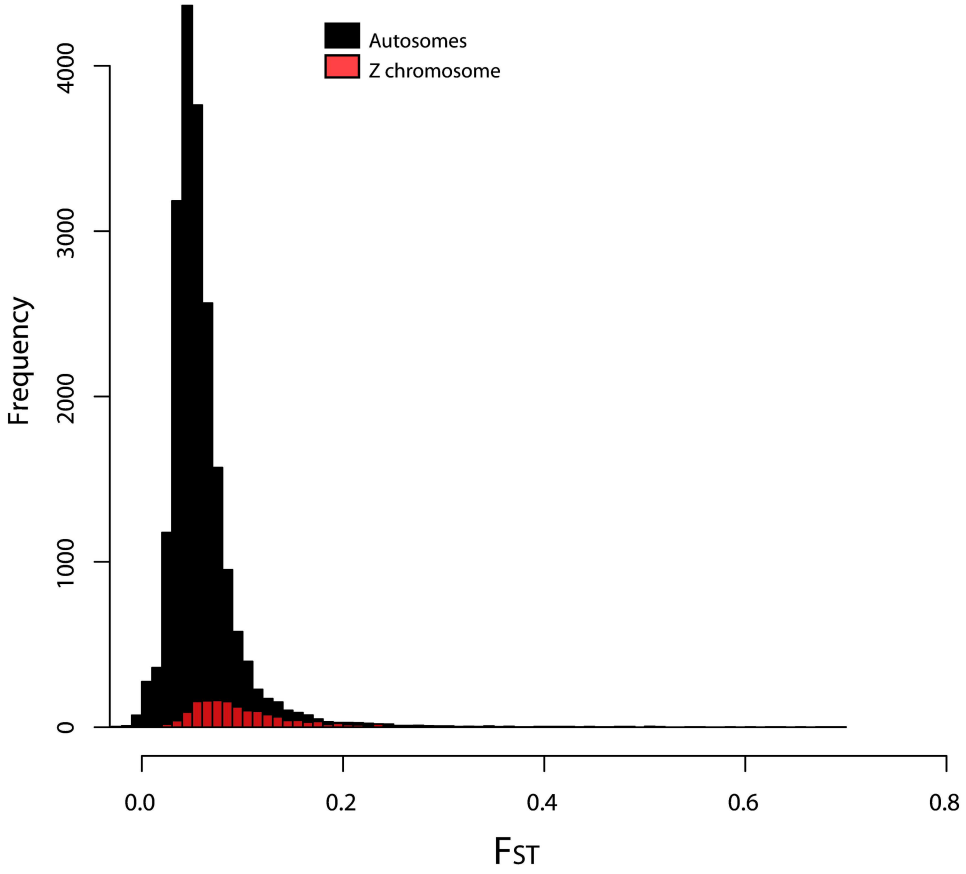

Fig. S13. Genome-wide distribution of genetic differentiation (F_{ST}) between carrion and hooded crows derived from 50 kb non-overlapping sliding windows across all autosomes (dark grey) and the sex chromosome (red).

Supplementary Tables S1 - S10

Table S1. Overview of sequencing data used for genome assembly of a male hooded crow. Coverage is given as the multiple of the expected genome size of 1.21 Gb. Libraries not used for the final assembly are shown in brackets.

Library	Library type	Insert size [bp]	Read length [bp]	Amount of sequence [Gb]		Sequence coverage [x-times]		Physical coverage [x-times]	
				raw	used	raw	used	raw	used
1	paired-end	142 +- 10	100	22.48	20.8	18.74	17.33	13.31	12.31
2	paired-end	152 +- 10	100	24.85	22.97	20.71	19.13	15.75	14.55
3	paired-end	162 +- 10	100	17.09	15.15	14.25	12.62	11.53	10.23
4	paired-end	164 +- 10	100	22.26	20.11	18.55	16.75	15.22	13.73
(5)	paired-end	173 +- 10	100	21.02	*NA*	17.52	*NA*	15.18	*NA*
(6)	paired-end	180 +- 10	100	16.09	*NA*	13.4	*NA*	12.12	*NA*
Total fragment libraries				***124***	***79***	***103***	***66***	***83***	***50***
(7)	mate pair	934 +/- 640	50	4.46	*NA*	3.72	*NA*	34.71	*NA*
8	mate pair	2046 +/- 164	50	11.01	8.92	9.17	7.43	186.76	151.37
9	mate pair	2,144 +/- 180	50	10.56	8.58	8.8	7.15	186.5	151.63
10	mate pair	2,339 +/- 155	50	6.72	5.3	5.6	4.42	130.78	103.13
11	mate pair	2,369 +/- 157	50	16.17	12.01	13.47	10.01	318.83	236.84
(12)	mate pair	2,940 +/- 295	50	3.32	*NA*	2.77	*NA*	81.37	*NA*
13	mate pair	4,980 +/- 229	50	9.93	8.13	8.28	6.78	411.35	336.89
14	mate pair	5,004 +/- 248	50	7.33	5.83	6.11	4.85	305.3	242.72
15	mate pair	5,303 +/- 225	50	15.56	9.79	12.97	8.16	687.87	432.92
16	mate pair	19,902 +/- 1,600	50	19.81	0.63	16.5	0.52	3,284.47	103.93
Total jump libraries				***105***	***59***	***87***	***49***	***5,628***	***1,759***
Final assembly				**229**	**138**	**191**	**115**	**5,711**	**1,809**

Table S2. Assembly statistics for the hooded crow draft genome.

parameter	value
Number of contigs	39,738
Number of contigs per Mb	37.8
Number of scaffolds	1,341
Total contig length	1,012,615,042
Total scaffold length, with gaps	1,049,934,819
N50 contig size in kb	62.5
N50 scaffold size in kb	15,932
N50 scaffold size in kb, with gaps	16,359
Number of scaffolds per Mb	1.28
Median size of gaps in scaffolds	581
Percent of bases in captured gaps	3.52
Percent of bases in negative gaps	0.03
Percent of ambiguous bases	15.23
Number of libraries used	36
Number of N's	36,971,193
Number of Gaps	39,425
AllpathsLG version used	41,687
Percent of assembly covered by contigs > 1MB	97
Number of contigs > 1MB	110
N90	3,342,196

Table S3. Overview of 100 bp paired-end mRNA sequencing data with average insert size of 180 bp used for assembly, annotation and analysis of the crow transcriptome (Ge: Germany, Sp: Spain, Sw: Sweden, Po: Poland).

Tissue	Nr. libraries/individuals				Raw data (total) [million read pairs]				Mapped data (total) [million read pairs]			
	Carrion		Hooded		Carrion		Hooded		Carrion		Hooded	
	Ge	Sp	Sw	Po	Ge	Sp	Sw	Po	Ge	Sp	Sw	Po
skin from torso	6/6	2/2	11/7	5/4	117	19	240	82	76	14	88	50
skin from head	6/6	2/2	10/7	4/4	101	31	142	71	70	23	83	48
forebrain	7/6	4/2	8/7	6/4	99	65	160	87	71	23	82	46
liver	6/6	2/2	8/7	4/4	118	34	122	79	76	20	85	49
gonads	6/6	2/2	6/6	4/4	120	31	85	78	74	25	71	48
heart	1/1	0	1/1	0	11	0	28	0	11	0	13	0
spleen	1/1	0	1/1	1/1	20	0	16	9	9	0	9	8
hypothalamus /pituitary	1/1	0	1/1	0	11	0	18	0	9	0	11	0
cerebellum	0	0	1/1	0	0	0	19	0	0	0	13	0
eye	1/1	0	1/1	1/1	15	0	12	27	10	0	10	10
Total	**35**	**12**	**48**	**25**	**612**	**180**	**842**	**433**	**406**	**115**	**452**	**259**

Table S4. Overview of 100 bp paired-end re-sequencing data with average insert size of 400 bp for all individuals from the European crow and the three outgroup species used for population genomic analysis (Ge: Germany, Sp: Spain, Sw: Sweden, Po: Poland).

		nr. of individuals	Raw data [Gb]		Final mapped data [Gb]	
			mean	range	mean	range
Carrion crow	**Ge**	15	16.4	8.8-29.5	15.3	8.2-28.1
	Sp	15	13.9	9.2-30.7	13.0	8.6-29.4
Hooded crow	**Sw***	15	16.7	9.5-34.6	15.5	8.8-32.2
	Po	15	12.1	8.5-17.4	11.3	8.2-15.9
American crow		5	13.5	10.4-16.9	12.3	8.8-15.8
Rook		2	12.7	12.3-13.1	11.8	11.3-12.3
European jackdaw		2	13.9	13.9-14.0	12.4	12.3-12.5

*not including the genome individual

Table S5. Population genetic summary statistics.

Nucleotide diversity π (column 3), Tajima's D statistic (column 4) and mean of window-based pairwise F_{ST} values between all four populations (columns 4-8) for autosomes (above the diagonal) and Z chromosome (below the diagonal) separately (cf. **fig. S13**). The number of pairwise fixed differences is given in parentheses. Asterisks symbolize standard type I error probabilities thresholds.

(Sub)-Species	**Population**	**Π** [x 10^{-3}]	**Tajima's D**	**F_{ST}** (fixed SNPs)			
		Auto-/gonosome	**Auto-/gonosome**	**Spain**	**Germany**	**Sweden**	**Poland**
Carrion Crow	**Spain**	1.42 / 1.05	-0.134 / 0.448	-	0.1013 (6)	0.1460 (409)	0.1554 (434)
	Germany	1.66 / 1.30	-1.021 / -0.620	0.2103 (104)	-	0.0172 (168)	0.0260 (144)
Hooded Crow	**Sweden**	1.63 / 1.23	-0.963 / -0.616	0.2657 (982)	0.0267 (0)	-	0.0153 (5)
	Poland	1.56 / 1.18	-0.709 / -0.364	0.2760 (930)	0.0364 (1)	0.0258 (1)	-

Table S6. ABBA-BABA test statistics based on the phylogeny ((H1:H2)H3)H4). H1, H2 and H3 represent 3 crow populations at a time [out of (Sp)ain, (Ge)rmany, (Po)land, (Sw)eden] with the same outgroup (H4). A Z-score value of 3 and above is considered significant (66). A positive value of the D-statistic reflects higher gene flow between H2 and H3 than with H1 and a negative value means reflects higher gene flow between H1 and H3 than with H2 (see **fig. S5**). Based on the significance of the test populations with evidence for gene flow are highlighted in bold. The D-statistic was estimated in *ANGSD* and *Egglib*. n.ABBA/n.BABA: number of ABBA/BABA sites, JackEst: corrected estimate of D. The standard error (SE) is rounded to 2 decimal places. Decisive population comparisons highlighted in grey are discussed in the Supplementary text and illustrated in **fig. S5**.

Populations			**ANGSD**						**Egglib**
H1	**H2**	**H3**	**n.ABBA**	**n.BABA**	**D**	**JackEst**	**SE**	**Z**	**D**
Sp	**Po**	**Ge**	137846	112037	0.1	0.1	0	**34.31**	**0.08**
Sp	**Sw**	**Ge**	138804	112003	0.11	0.11	0	**35.43**	**0.08**
Po	**Sw**	Ge	114733	113777	0	0	0	1.5	0.00
Ge	Po	**Sp**	88529	112037	-0.12	-0.12	0	**-39.6**	**-0.10**
Ge	Sw	**Sp**	87649	112003	-0.12	-0.12	0	**-41.56**	**-0.10**
Po	**Sw**	Sp	88674	89535	-0.01	-0.01	0	-1.66	0.00
Ge	Sp	**Po**	88529	137846	-0.22	-0.22	0	**-77.51**	**-0.18**
Ge	**Sw**	**Po**	127344	113777	0.06	0.06	0	**20.02**	**0.04**
Sp	**Sw**	**Po**	152483	89535	0.26	0.26	0	**95.79**	**0.21**
Ge	Sp	**Sw**	87649	138804	-0.23	-0.23	0	**-81.13**	**-0.18**
Ge	**Po**	**Sw**	127344	114733	0.05	0.05	0	**18.61**	**0.04**
Sp	**Po**	**Sw**	152483	88674	0.27	0.27	0	**97.51**	**0.21**

Table S7. Significant Gene Ontology terms among differentially expressed genes. No categories were significant for forebrain, liver or gonads.

tissue	GO category	FDR	GO description
skin from torso (carrion vs. hooded)	GO:0042438	0.00279	melanin biosynthetic process
	GO:0015991	0.00340	ATP hydrolysis coupled proton transport
	GO:0016021	0.00340	integral to membrane
	GO:0033178	0.03531	proton-transporting two-sector ATPase complex, catalytic domain
	GO:0005765	0.03531	lysosomal membrane
	GO:0016020	0.03531	membrane
	GO:0042470	0.03531	melanosome
	GO:0043473	0.03531	pigmentation
skin from head (carrion vs. hooded)	GO:0032982	0.00004	myosin filament
	GO:0003779	0.02006	actin binding
	GO:0016459	0.03073	myosin complex
	GO:0003774	0.03073	motor activity

Table S8. Differentially expressed melanogenesis genes in comparison between carrion and hooded crows. All except the upper three genes were differentially expressed in skin from torso, where carrion crows produce black and hooded crows grey feathers. "LogFC" is log-fold change in expression between comparisons, and "FDR" is the false discovery rate (genes are ordered by FDR within tissues, starting with the most significant gene). The "under" column indicates which taxon had lower expression for the gene in question. The "MITF" column indicates whether the gene is transcriptionally activated by MITF. The "cactus" and "F_{ST}" columns indicate whether the gene was detected as showing high divergence between taxa in the genome sequencing approach, using cacti and F_{ST} measures, respectively.

Zebra finch Ensembl ID	Gene name	Tissue	logFC	FDR	under	MITF	cactus	F_{ST}
ENSTGUG00000004110	WIPI1	gonads	3.51	1.21E-02	carrion	1	0	0
ENSTGUG00000013283	beta-defensin	liver	-3.08	8.71E-04	hooded	0	0	0
ENSTGUG00000010401	CAV3	skin_head	-3.79	1.08E-04	hooded	0	0	0
ENSTGUG00000003144	HPGDS	skin_torso	-2.78	2.24E-182	hooded	1	1	0
ENSTGUG00000004704	TYRP1	skin_torso	-2.15	2.22E-74	hooded	1	0	0
ENSTGUG00000010434	OCA2	skin_torso	-2.12	3.19E-58	hooded	0	0	0
ENSTGUG00000001948	SLC45A2	skin_torso	-1.47	1.11E-50	hooded	1	0	0
ENSTGUG00000012909	RAB38	skin_torso	-1.54	2.81E-42	hooded	1	0	0
ENSTGUG00000018271	MLANA	skin_torso	-1.64	1.10E-37	hooded	1	0	0
ENSTGUG00000012899	TYR	skin_torso	-1.55	1.48E-32	hooded	1	0	0
ENSTGUG00000012580	SLC24A4	skin_torso	-1.6	1.92E-29	hooded	0	0	0
ENSTGUG00000009981	RASGRF1	skin_torso	-1.53	1.88E-27	hooded	0	0	0
ENSTGUG00000008024	MC1R	skin_torso	-1.28	4.83E-14	hooded	1	0	0
ENSTGUG00000004211	CLCN7	skin_torso	-0.5	1.38E-08	hooded	1	0	0
ENSTGUG00000006354	PPM1H	skin_torso	-0.68	5.99E-07	hooded	1	0	0
ENSTGUG00000000598	CTSL	skin_torso	-0.39	1.53E-05	hooded	0	0	0
ENSTGUG00000008853	NR4A3	skin_torso	-0.98	1.65E-05	hooded	0	0	0
ENSTGUG00000006682	CASP3	skin_torso	-0.49	5.18E-05	hooded	0	0	0
ENSTGUG00000006166	NDP	skin_torso	-1.06	2.22E-04	hooded	0	0	1
ENSTGUG00000017296	ATP6V1B2	skin_torso	-0.31	1.62E-03	hooded	1	0	0
ENSTGUG00000005328	TPCN2	skin_torso	-0.38	3.11E-03	hooded	0	0	0
ENSTGUG00000005425	EDNRB2	skin_torso	0.58	1.15E-02	carrion	1	0	0
ENSTGUG00000012154	OSTM1	skin_torso	-0.29	2.14E-02	hooded	1	0	0

Table S9. Repeat content listed by repeat type of the hooded crow genome as identified by *RepeatMasker*.

Repeat type			Number of elements	Length (bp)	% of sequence
Retroelements			157,810	51170521	4.87
	SINEs		7,736	925,744	0.09
	LINEs		118,253	35,577,887	3.39
		CR1	118,154	35,559,799	3.39
		L1	22	2,568	0.0
	LTR elements		31,821	14,666,890	1.40
	Retroviral		31,737	14,649,550	1.39
DNA transposons			3,541	53,2488	0.05
	Hobo-Activator		167	55,050	0.01
	Tc1-IS630-Pogo		419	73,410	0.01
	PiggyBac		531	85,039	0.01
Unclassified			2,402	405,113	0.04
Small RNA			468	40,243	0.0
Satellites			785	115,028	0.01
Simple repeats			62,824	3,191,230	0.30
Low complexity			118,921	5,031,365	0.48

Table S10A. Arithmetic mean values (+/- standard error) for differentiation, divergence and diversity statistics calculated for peak and non-peak regions within autosomes. F_{ST} and D_{xy} were calculated between taxa (CC: carrion crow, HC: hooded crow), nucleotide diversity (π), Tajima's D, Fu and Li's D, and runs of homozygosity (iES) are listed for each population separately (Ge: Germany, Sp: Spain, Sw: Sweden, Po: Poland). Significant differences between peak and non-peak regions as dertermined by Kruskal-Wallis tests are highlighted in bold.

Autosomes			**Peak regions**	**Non-peak regions**	**p-value**
mean F_{ST}			0.3012 +/- 0.0077	0.05654 +/- 0.00019	**< 2.2e-16**
mean Dxy			0.00106 +/- 0.00004	0.00170 +/- 0.00001	**< 2.2e-16**
π	**CC**	**Ge**	0.00083 +/- 0.00004	0.00167 +/- 0.00001	**< 2.2e-16**
		Sp	0.00056 +/- 0.00003	0.00143 +/- 0.00001	**< 2.2e-16**
	HC	**Sw**	0.00068 +/- 0.00003	0.00164 +/-0.00001	**< 2.2e-16**
		Po	0.00064 +/- 0.00003	0.00157 +/- 0.00001	**< 2.2e-16**
Tajima's D	**CC**	**Ge**	-0.6041 +/- 0.0340	-1.0257 +/- 0.00163	**< 2.2e-16**
		Sp	-0.2452 +/- 0.0385	-0.1330 +/- 0.00217	**0.0020**
	HC	**Sw**	-0.9196 +/- 0.0336	-0.9632 +/- 0.00168	0.2309
		Po	-0.6302 +/- 0.0340	-0.7095 +/- 0.00183	0.0127
Fu and Li's D	**CC**	**Ge**	-0.9024 +/- 0.0552	-1.2262 +/- 0.00329	**1.29e-15**
		Sp	0.3233 +/- 0.0517	0.2739 +/- 0.00392	0.0746
	HC	**Sw**	-1.3780 +/- 0.0566	-1.2449 +/ 0.00322	0.0084
		Po	-0.6231 +/- 0.0540	-0.6021 +/- 0.00356	0.6475
iES	**CC**	**Ge**	9,134 +/- 496.613	2,924 +/- 11.660	**< 2.2e-16**
		Sp	16,939 +/- 841.158	4,361 +/- 21.165	**< 2.2e-16**
	HC	**Sw**	9,853 +/- 475.003	3,033 +/- 12.838	**< 2.2e-16**
		Po	12,123 +/- 604.401	3,513 +/- 16.118	**< 2.2e-16**

Table S10B. Arithmetic mean values (+/- standard error) for differentiation, divergence and diversity statistics calculated for peak and non-peak regions within the Z chromosome. F_{ST} and D_{xy} were calculated between taxa (CC: carrion crow, HC: hooded crow), nucleotide diversity (π), Tajima's D, Fu and Li's D, and runs of homozygosity (iES) are listed for each population separately (Ge: Germany, Sp: Spain, Sw: Sweden, Po: Poland). Significant differences between peak and non-peak regions as dertermined by Kruskal-Wallis tests are highlighted in bold.

Z chromosome			**Peak regions**	**Non-peak regions**	**p-value**
mean F_{ST}			0.3434 +/- 0.0115	0.099233 +/- 0.00128	**5.58e-12**
mean Dxy			0.00153 +/- 0.00008	0.0014108 +/- 0.00002	0.05409
π	**CC**	**Ge**	0.00126 +/- 0.00007	0.0013014 +/- 0.00002	0.89280
		Sp	0.000584 +/- 0.00006	0.0010540 +/- 0.00002	**1.36e-6**
	HC	**Sw**	0.00081 +/- 0.00006	0.0012360 +/- 0.00002	**2.34e-5**
		Po	0.000779 +/- 0.00007	0.0011863 +/- 0.00002	**9.03e-5**
Tajima's D	**CC**	**Ge**	-0.18413 +/- 0.0880	-0.6249 +/- 0.00818	**4.09e-6**
		Sp	0.1659 +/- 0.1815	0.4510 +/- 0.01204	0.05967
	HC	**Sw**	-0.7681 +/- 0.1224	-0.6138 +/- 0.00859	0.32050
		Po	-0.35455 +/- 0.1302	-0.3636 +/- 0.00911	0.81310
Fu and Li's D	**CC**	**Ge**	-0.33677 +/- 0.1012	-0.6241 +/- 0.01498	0.02117
		Sp	0.5624 +/- 0.1687	0.8281 +/- 0.0159	0.16780
	HC	**Sw**	-1.0896 +/- 0.186	-0.8455 +/ 0.0148	0.04560
		Po	-0.1649 +/- 0.170	-0.2931 +/- 0.0147	0.41330
iES	**CC**	**Ge**	6,211 +/- 675.02	4,003 +/- 51.915	**0.00023**
		Sp	13,312 +/- 1027.93	8,523 +/- 120.904	**8.85e-6**
	HC	**Sw**	8,728 +/- 1066.055	4,290 +/- 54.960	**2.50e-7**
		Po	13,970 +/- 2141.428	5,227.76 +/- 80.688	**1.31e-7**

Table S11. The number of regions, base-pair length and percent of genome associated with local phylogenetic signals, called cacti, detected using *Saguaro*. Class 1 cacti include regions that produce a taxon-specific phylogeny, whereas Class 2 cacti do not cluster based on taxonomy.

	Number of regions > 300 bp	**Cumulative length (bp)**	**Percent**
Class 1			
Cactus 5	2,087	1,593,644	0.156
Cactus 6	1,980	1,286,989	0.126
Class 2			
Cactus 0	197	140,067	0.014
Cactus 1	302	215,259	0.021
Cactus 2	207	269,130	0.026
Cactus 4	139	117,806	0.012
Cactus 7	4,218	2,855,291	0.280
Cactus 8	38,530	1,011,121,650	99.110
Cactus 9	1,185	711,364	0.070
Cactus 10	2,954	1,888,678	0.185

References

1. R. Abbott, D. Albach, S. Ansell, J. W. Arntzen, S. J. Baird, N. Bierne, J. Boughman, A. Brelsford, C. A. Buerkle, R. Buggs, R. K. Butlin, U. Dieckmann, F. Eroukhmanoff, A. Grill, S. H. Cahan, J. S. Hermansen, G. Hewitt, A. G. Hudson, C. Jiggins, J. Jones, B. Keller, T. Marczewski, J. Mallet, P. Martinez-Rodriguez, M. Möst, S. Mullen, R. Nichols, A. W. Nolte, C. Parisod, K. Pfennig, A. M. Rice, M. G. Ritchie, B. Seifert, C. M. Smadja, R. Stelkens, J. M. Szymura, R. Väinölä, J. B. Wolf, D. Zinner, Hybridization and speciation. *J. Evol. Biol.* **26**, 229–246 (2013). doi:10.1111/j.1420-9101.2012.02599.x Medline

2. S. H. Martin *et al*., Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.* **23**, 1817–1828 (2013). doi:10.1101/gr.159426.113

3. S. Sankararaman, S. Mallick, M. Dannemann, K. Prüfer, J. Kelso, S. Pääbo, N. Patterson, D. Reich, The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354–357 (2014). doi:10.1038/nature12961 Medline

4. Heliconius Genome Consortium, Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94–98 (2012). doi:10.1038/nature11041

5. H. Ellegren, L. Smeds, R. Burri, P. I. Olason, N. Backström, T. Kawakami, A. Künstner, H. Mäkinen, K. Nadachowska-Brzyska, A. Qvarnström, S. Uebbing, J. B. Wolf, The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* **491**, 756–760 (2012). doi:10.1038/nature11584

6. W. Meise, Die Verbreitung der Aaskrähe (Formenkreis *Corvus corone* L.). *J. Ornithol.* **76**, 1–203 (1928).

7. G. Hewitt, The genetic legacy of the Quaternary ice ages. *Nature* **405**, 907–913 (2000). doi:10.1038/35016000 Medline

8. J. B. W. Wolf, T. Bayer, B. Haubold, M. Schilhabel, P. Rosenstiel, D. Tautz, Nucleotide divergence vs. gene expression differentiation: Comparative transcriptome sequencing in

natural isolates from the carrion crow and its hybrid zone with the hooded crow. *Mol. Ecol.* **19** (suppl. 1), 162–175 (2010). doi:10.1111/j.1365-294X.2009.04471.x Medline

9. C. Randler, Assortative mating of carrion *Corvus corone* and hooded crows *C. cornix* in the hybrid zone in eastern Germany. *Ardea* **95**, 143–149 (2007). doi:10.5253/078.095.0116

10. See supplementary materials on *Science* Online.

11. R. F. Guerrero, F. Rousset, M. Kirkpatrick, Coalescent patterns for chromosomal inversions in divergent populations. *Philos. Trans. R. Soc. London Ser. B* **367**, 430–438 (2012). doi:10.1098/rstb.2011.0246 Medline

12. M. J. Hoogduijn, I. S. Hitchcock, N. P. Smit, J. M. Gillbro, K. U. Schallreuter, P. G. Genever, Glutamate receptors on human melanocytes regulate the expression of MiTF. *Pigment Cell Res.* **19**, 58–67 (2006). doi:10.1111/j.1600-0749.2005.00284.x Medline

13. C. L. Farnsworth, N. W. Freshney, L. B. Rosen, A. Ghosh, M. E. Greenberg, L. A. Feig, Calcium activation of Ras mediated by neuronal exchange factor Ras-GRF. *Nature* **376**, 524–527 (1995). doi:10.1038/376524a0 Medline

14. B. Moniot, A. Farhat, K. Aritake, F. Declosmenil, S. Nef, N. Eguchi, Y. Urade, F. Poulat, B. Boizet-Bonhoure, Hematopoietic prostaglandin D synthase (H-Pgds) is expressed in the early embryonic gonad and participates to the initial nuclear translocation of the SOX9 protein. *Dev. Dyn.* **240**, 2335–2343 (2011). doi:10.1002/dvdy.22726 Medline

15. K. A. Martemyanov, V. Y. Arshavsky, Biology and functions of the RGS9 isoforms. *Prog. Mol. Biol. Transl. Sci.* **86**, 205–227 (2009). doi:10.1016/S1877-1173(09)86007-9

16. N. Saino, L. Scatizzi, Selective aggressiveness and dominance among carrion crows, hooded crows and hybrids. *Boll. Zool.* **58**, 255–260 (1991). doi:10.1080/11250009109355762

17. P. P. M. Schnetkamp, The SLC24 gene family of Na^+/Ca^{2+}-K^+ exchangers: From sight and smell to memory consolidation and skin pigmentation. *Mol. Aspects Med.* **34**, 455–464 (2013). doi:10.1016/j.mam.2012.07.008 Medline

18. D. T. Parkin, M. Collinson, A. J. Helbig, A. G. Knox, G. Sangster, The taxonomic status of carrion and hooded crows. *Br. Birds* **96**, 274–290 (2003).

19. J. L. Feder, S. P. Egan, P. Nosil, The genomics of speciation-with-gene-flow. *Trends Genet.* **28**, 342–350 (2012). doi:10.1016/j.tig.2012.03.009 Medline

20. A. A. Hoffmann, L. H. Rieseberg, Revisiting the impact of inversions in evolution: From population genetic markers to drivers of adaptive shifts and speciation? *Annu. Rev. Ecol. Evol. Syst.* **39**, 21–42 (2008). doi:10.1146/annurev.ecolsys.39.110707.173532 Medline

21. F. C. Jones, M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli, J. Johnson, R. Swofford, M. Pirun, M. C. Zody, S. White, E. Birney, S. Searle, J. Schmutz, J. Grimwood, M. C. Dickson, R. M. Myers, C. T. Miller, B. R. Summers, A. K. Knecht, S. D. Brady, H. Zhang, A. A. Pollen, T. Howes, C. Amemiya, J. Baldwin, T. Bloom, D. B. Jaffe, R. Nicol, J. Wilkinson, E. S. Lander, F. Di Palma, K. Lindblad-Toh, D. M. Kingsley; Broad Institute Genome Sequencing Platform & Whole Genome Assembly Team, The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55–61 (2012). doi:10.1038/nature10944 Medline

22. K. Kunte, W. Zhang, A. Tenger-Trolander, D. H. Palmer, A. Martin, R. D. Reed, S. P. Mullen, M. R. Kronforst, doublesex is a mimicry supergene. *Nature* **507**, 229–232 (2014). doi:10.1038/nature13112 Medline

23. H. Thorneycroft, Cytogenetic study of white-throated sparrow *Zonotrichia albicollis* (Gmelin). *Evolution* **29**, 611–621 (1975).

24. S. Renaut, C. J. Grassa, S. Yeaman, B. T. Moyers, Z. Lai, N. C. Kane, J. E. Bowers, J. M. Burke, L. H. Rieseberg, Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nat. Commun.* **4**, 1827 (2013). doi:10.1038/ncomms2833 Medline

25. C. A. McLean, D. Stuart-Fox, Geographic variation in animal colour polymorphisms and its role in speciation. *Biol. Rev. Camb. Philos. Soc.* 10.1111/brv.12083 (2014). doi:10.1111/brv.12083 Medline

26. S. Gnerre, I. Maccallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, A. M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E. S. Lander, D. B. Jaffe, High-quality draft assemblies

of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 1513–1518 (2011). doi:10.1073/pnas.1017351108 Medline

27. R. Li, W. Fan, G. Tian, H. Zhu, L. He, J. Cai, Q. Huang, Q. Cai, B. Li, Y. Bai, Z. Zhang, Y. Zhang, W. Wang, J. Li, F. Wei, H. Li, M. Jian, J. Li, Z. Zhang, R. Nielsen, D. Li, W. Gu, Z. Yang, Z. Xuan, O. A. Ryder, F. C. Leung, Y. Zhou, J. Cao, X. Sun, Y. Fu, X. Fang, X. Guo, B. Wang, R. Hou, F. Shen, B. Mu, P. Ni, R. Lin, W. Qian, G. Wang, C. Yu, W. Nie, J. Wang, Z. Wu, H. Liang, J. Min, Q. Wu, S. Cheng, J. Ruan, M. Wang, Z. Shi, M. Wen, B. Liu, X. Ren, H. Zheng, D. Dong, K. Cook, G. Shan, H. Zhang, C. Kosiol, X. Xie, Z. Lu, H. Zheng, Y. Li, C. C. Steiner, T. T. Lam, S. Lin, Q. Zhang, G. Li, J. Tian, T. Gong, H. Liu, D. Zhang, L. Fang, C. Ye, J. Zhang, W. Hu, A. Xu, Y. Ren, G. Zhang, M. W. Bruford, Q. Li, L. Ma, Y. Guo, N. An, Y. Hu, Y. Zheng, Y. Shi, Z. Li, Q. Liu, Y. Chen, J. Zhao, N. Qu, S. Zhao, F. Tian, X. Wang, H. Wang, L. Xu, X. Liu, T. Vinar, Y. Wang, T. W. Lam, S. M. Yiu, S. Liu, H. Zhang, D. Li, Y. Huang, X. Wang, G. Yang, Z. Jiang, J. Wang, N. Qin, L. Li, J. Li, L. Bolund, K. Kristiansen, G. K. Wong, M. Olson, X. Zhang, S. Li, H. Yang, J. Wang, J. Wang, The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311–317 (2010). doi:10.1038/nature08696 Medline

28. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009). doi:10.1093/bioinformatics/btp324 Medline

29. A. E. Vinogradov, Genome size and GC-percent in vertebrates as determined by flow cytometry: The triangular relationship. *Cytometry* **31**, 100–109 (1998). doi:10.1002/(SICI)1097-0320(19980201)31:2<100::AID-CYTO5>3.0.CO;2-Q Medline

30. G. Venturini, R. D'Ambrogi, E. Capanna, Size and structure of the bird genome—I. DNA content of 48 species of neognathae. *Comp. Biochem. Physiol. B* **85**, 61–65 (1986). doi:10.1016/0305-0491(86)90221-X

31. E. M. Rasch, DNA "standards" and the range of accurate DNA estimates by Feulgen absorption microspectrophotometry. *Prog. Clin. Biol. Res.* **196**, 137–166 (1985). Medline

32. C. B. Andrews, S. A. Mackenzie, T. R. Gregory, Genome size and wing parameters in passerine birds. *Proc. R. Soc. B* **276**, 55–61 (2009). doi:10.1098/rspb.2008.1012

33. G. Marçais, C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011). doi:10.1093/bioinformatics/btr011 Medline

34. G. V. Roslik, A. P. Kriukov, [Karyological study of some Corvine birds (Corvidae, aves)]. *Russ. J. Genet.* **37**, 962–973 (2001). Medline

35. M. G. Grabherr, P. Russell, M. Meyer, E. Mauceli, J. Alföldi, F. Di Palma, K. Lindblad-Toh, Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics* **26**, 1145–1151 (2010). doi:10.1093/bioinformatics/btq102 Medline

36. A. F. A. Smit, R. Hubley, P. Green, Repeat Masker Open-3.0, www.repeatmasker.org (2010).

37. G. Lunter, M. Goodson, Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21**, 936–939 (2011). doi:10.1101/gr.111120.110 Medline

38. C. Alkan, S. Sajjadian, E. E. Eichler, Limitations of next-generation genome sequence assembly. *Nat. Methods* **8**, 61–65 (2011). doi:10.1038/nmeth.1527 Medline

39. L. Ye, L. W. Hillier, P. Minx, N. Thane, D. P. Locke, J. C. Martin, L. Chen, M. Mitreva, J. R. Miller, K. V. Haub, D. J. Dooling, E. R. Mardis, R. K. Wilson, G. M. Weinstock, W. C. Warren, A vertebrate case study of the quality of assemblies derived from next-generation sequences. *Genome Biol.* **12**, R31 (2011). doi:10.1186/gb-2011-12-3-r31 Medline

40. M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011). doi:10.1038/nbt.1883 Medline

41. H. Ellegren, Evolutionary stasis: The stable chromosomes of birds. *Trends Ecol. Evol.* **25**, 283–291 (2010). doi:10.1016/j.tree.2009.12.004 Medline

42. B. M. Skinner, D. K. Griffin, Intrachromosomal rearrangements in avian genome evolution: Evidence for regions prone to breakpoints. *Heredity* **108**, 37–41 (2012). doi:10.1038/hdy.2011.99 Medline

43. A. E. Darling, B. Mau, N. T. Perna, progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLOS ONE* **5**, e11147 (2010). doi:10.1371/journal.pone.0011147 Medline

44. M. Bernt, A. Donath, F. Jühling, F. Externbrink, C. Florentz, G. Fritzsch, J. Pütz, M. Middendorf, P. F. Stadler, MITOS: Improved de novo metazoan mitochondrial genome annotation. *Mol. Phylogenet. Evol.* **69**, 313–319 (2013). doi:10.1016/j.ympev.2012.08.023 Medline

45. B. Nabholz, H. Ellegren, J. B. W. Wolf, High levels of gene expression explain the strong evolutionary constraint of mitochondrial protein-coding genes. *Mol. Biol. Evol.* **30**, 272–284 (2013). doi:10.1093/molbev/mss238 Medline

46. O. Keller, F. Odronitz, M. Stanke, M. Kollmar, S. Waack, Scipio: Using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. *BMC Bioinformatics* **9**, 278 (2008). doi:10.1186/1471-2105-9-278 Medline

47. A. Roberts, C. Trapnell, J. Donaghey, J. L. Rinn, L. Pachter, Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* **12**, R22 (2011). doi:10.1186/gb-2011-12-3-r22 Medline

48. C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, L. Pachter, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012). doi:10.1038/nprot.2012.016 Medline

49. C. Trapnell, L. Pachter, S. L. Salzberg, TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009). doi:10.1093/bioinformatics/btp120 Medline

50. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). doi:10.1093/bioinformatics/btq033 Medline

51. R. Griffiths, M. C. Double, K. Orr, R. J. G. Dawson, A DNA test to sex most birds. *Mol. Ecol.* **7**, 1071–1075 (1998). doi:10.1046/j.1365-294x.1998.00389.x Medline

52. M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, M. J. Daly, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011). doi:10.1038/ng.806 Medline

53. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). doi:10.1093/bioinformatics/btp352 Medline

54. E. Garrison, G. Marth, http://arxiv.org/abs/1207.3907 (2012).

55. J. W. Poelstra, H. Ellegren, J. B. W. Wolf, An extensive candidate gene approach to speciation: Diversity, divergence and linkage disequilibrium in candidate pigmentation genes across the European crow hybrid zone. *Heredity* **111**, 467–473 (2013). doi:10.1038/hdy.2013.68 Medline

56. A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, D. Reich, Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006). doi:10.1038/ng1847 Medline

57. R. Nielsen, T. Korneliussen, A. Albrechtsen, Y. Li, J. Wang, SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing data. *PLOS ONE* **7**, e37558 (2012). doi:10.1371/journal.pone.0037558 Medline

58. J. Goudet, hierfstat, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* **5**, 184–186 (2005). doi:10.1111/j.1471-8286.2004.00828.x

59. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2012; www.r-project.org).

60. B. S. Weir, C. C. Cockerham, Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984). doi:10.2307/2408641

61. R.-C. Yang, Estimating hierarchical F-statistics. *Evolution* **52**, 950–956 (1998). doi:10.2307/2411227

62. M. Nei, *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).

63. S. R. Browning, B. L. Browning, Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007). doi:10.1086/521987 Medline

64. M. Gautier, R. Vitalis, rehh: An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* **28**, 1176–1177 (2012). doi:10.1093/bioinformatics/bts115 Medline

65. E. Y. Durand, N. Patterson, D. Reich, M. Slatkin, Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011). doi:10.1093/molbev/msr048 Medline

66. N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, D. Reich, Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012). doi:10.1534/genetics.112.145037 Medline

67. S. De Mita, M. Siol, EggLib: Processing, analysis and simulation tools for population genetics and genomics. *BMC Genet.* **13**, 27 (2012). doi:10.1186/1471-2156-13-27 Medline

68. S. H. Martin, J. W. Davey, C. D. Jiggins, http://biorxiv.org/content/early/2013/12/11/001347 (2013).

69. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009). doi:10.1101/gr.094052.109 Medline

70. J. Hey, R. Nielsen, Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 2785–2790 (2007). doi:10.1073/pnas.0611164104 Medline

71. G. Landan, D. Graur, Local reliability measures from sets of co-optimal multiple sequence alignments. *Pac. Symp. Biocomput.* **13**, 15–24 (2008). Medline

72. A. Löytynoja, N. Goldman, Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**, 1632–1635 (2008). doi:10.1126/science.1158395 Medline

73. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007). doi:10.1093/molbev/msm088 Medline

74. E. Haring, A. Gamauf, A. Kryukov, Phylogeographic patterns in widespread corvid birds. *Mol. Phylogenet. Evol.* **45**, 840–862 (2007). doi:10.1016/j.ympev.2007.06.016 Medline

75. A. P. Kryukov, H. Suzuki, Phylogeography of carrion, hooded, and jungle crows (Aves, Corvidae) inferred from partial sequencing of the mitochondrial DNA cytochrome b gene. *Russ. J. Genet.* **36**, 922–929 (2000).

76. N. Zamani, P. Russell, H. Lantz, M. P. Hoeppner, J. R. Meadows, N. Vijay, E. Mauceli, F. di Palma, K. Lindblad-Toh, P. Jern, M. G. Grabherr, Unsupervised genome-wide recognition of local relationship patterns. *BMC Genomics* **14**, 347 (2013). doi:10.1186/1471-2164-14-347 Medline

77. F. Haas, M. A. Pointer, N. Saino, A. Brodin, N. I. Mundy, B. Hansson, An analysis of population genetic differentiation and genotype-phenotype association across the hybrid zone of carrion and hooded crows using microsatellites and MC1R. *Mol. Ecol.* **18**, 294–305 (2009). doi:10.1111/j.1365-294X.2008.04017.x Medline

78. A. Cáceres, S. S. Sindi, B. J. Raphael, M. Cáceres, J. R. González, Identification of polymorphic inversions from genotypes. *BMC Bioinformatics* **13**, 28 (2012). doi:10.1186/1471-2105-13-28 Medline

79. S. S. Sindi, B. J. Raphael, Identification and frequency estimation of inversion polymorphisms from haplotype data. *J. Comput. Biol.* **17**, 517–531 (2010). doi:10.1089/cmb.2009.0185 Medline

80. J. C. Barrett, B. Fry, J. Maller, M. J. Daly, Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005). doi:10.1093/bioinformatics/bth457 Medline

81. V. Bansal, A. Bashir, V. Bafna, Evidence for large inversion polymorphisms in the human genome from HapMap data. *Genome Res.* **17**, 219–230 (2007). doi:10.1101/gr.5774507 Medline

82. J. I. Lucas Lledó, M. Cáceres, On the power and the systematic biases of the detection of chromosomal inversions by paired-end genome sequencing. *PLOS ONE* **8**, e61292 (2013). doi:10.1371/journal.pone.0061292 Medline

83. T. Rausch, T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, J. O. Korbel, DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012). doi:10.1093/bioinformatics/bts378 Medline

84. R. B. Corbett-Detig, C. Cardeno, C. H. Langley, Sequence-based detection and breakpoint assembly of polymorphic inversions. *Genetics* **192**, 131–137 (2012). doi:10.1534/genetics.112.141622 Medline

85. L. S. Stevison, K. B. Hoehn, M. A. F. Noor, Effects of inversions on within- and between-species recombination and divergence. *Genome Biol. Evol.* **3**, 830–841 (2011). doi:10.1093/gbe/evr081 Medline

86. M. P. Donnelly, P. Paschou, E. Grigorenko, D. Gurwitz, S. Q. Mehdi, S. L. Kajuna, C. Barta, S. Kungulilo, N. J. Karoma, R. B. Lu, O. V. Zhukova, J. J. Kim, D. Comas, M. Siniscalco, M. New, P. Li, H. Li, V. G. Manolopoulos, W. C. Speed, H. Rajeevan, A. J. Pakstis, J. R. Kidd, K. K. Kidd, The distribution and most recent common ancestor of the 17q21 inversion in humans. *Am. J. Hum. Genet.* **86**, 161–171 (2010). doi:10.1016/j.ajhg.2010.01.007 Medline

87. B. Li, C. N. Dewey, RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011). doi:10.1186/1471-2105-12-323 Medline

88. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010). doi:10.1093/bioinformatics/btp616 Medline

89. N. Leng, J. A. Dawson, J. A. Thomson, V. Ruotti, A. I. Rissman, B. M. Smits, J. D. Haag, M. N. Gould, R. M. Stewart, C. Kendziorski, EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **29**, 1035–1043 (2013). doi:10.1093/bioinformatics/btt087 Medline

90. M. D. Robinson, A. Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010). doi:10.1186/gb-2010-11-3-r25 Medline

91. J. H. Bullard, E. Purdom, K. D. Hansen, S. Dudoit, Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94 (2010). doi:10.1186/1471-2105-11-94 Medline

92. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).

93. G. E. Hill, K. J. McGraw, *Bird Coloration: Function and Evolution* (Harvard Univ. Press, Cambridge, MA, 2006).

94. M. Pagel, Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proc. R. Soc. London Ser. B* **255**, 37–45 (1994). doi:10.1098/rspb.1994.0006

95. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004). doi:10.1093/bioinformatics/btg412 Medline

96. T. Londei, Alternation of clear-cut colour patterns in *Corvus* crow evolution accords with learning-dependent social selection against unusual-looking conspecifics. *Ibis* **155**, 632–634 (2013). doi:10.1111/ibi.12074

97. K. A. Jønsson, P.-H. Fabre, M. Irestedt, Brains, tools, innovation and biogeography in crows and ravens. *BMC Evol. Biol.* **12**, 72 (2012). doi:10.1186/1471-2148-12-72 Medline