Supplementary Materials for

Genomic diversity and evolution of the head crest in the rock pigeon

Michael D. Shapiro, Zev Kronenberg, Cai Li, Eric T. Domyan, Hailin Pan, Michael Campbell, Hao Tan, Chad D. Huff, Haofu Hu, Anna I. Vickrey, Sandra C.A. Nielsen, Sydney A. Stringham, Hao Hu, Eske Willerslev, M. Thomas P. Gilbert, Mark Yandell, Guojie Zhang, Jun Wang

correspondence to: mike.shapiro@utah.edu, wangj@genomics.org.cn

This PDF file includes:

Materials and Methods

Genome assembly

The DNA sample for sequencing of the reference genome was extracted from blood obtained from a single, male Danish Tumbler, bred by Anders and Hans Ove Christiansen (Danmarks Racedueforeninger, Næstved, Denmark). This breed was chosen because it is an old breed that is believed to have changed little in recent history. Seven paired-end sequencing libraries were constructed, with insert sizes of 170 bp, 500 bp, 800 bp, 2 kb, 5 kb, 10 kb and 20 kb. The libraries were sequenced using Illumina HiSeq2000 platform, yielding a total of 127.17 Gb raw data (Table S1). The raw sequences were filtered for low quality, adapter sequence, paired-end read overlap, and PCR duplicates. We also performed an error correction step on the raw reads before assembling. Filtering and error correction resulted in 81.57 Gb of clean data for genome assembly with the genome with SOAPdenovo (*26*). We used k-mer frequency to estimate genome size at 1.3 Gb (Table S3); final genome coverage was 62.75-fold. The assembly has an N50 scaffold length of 3.15 Mb and a total length of 1.11 Gb (84.6% representation), the largest contig is > 250 kb and the largest scaffold > 25.6 Mb (Table S4). Overall, the assembly of the pigeon genome surpasses the turkey genome (assembled using reads from a combination of next-generation technologies) in both contig and scaffold size (Table S6), demonstrating that deep sequencing using only short Illumina reads is sufficient to produce a useful draft avian genome assembly.

To detect contamination, we aligned the assembly against the NCBI nr databases, but we found little contamination from non-pigeon genomic sequence. We also assessed the assembly by aligning the genome to 2108 ESTs from *Columba livia* (downloaded from Genbank). Approximately 90% of these ESTs could be mapped to the assembly (Table S5), suggesting the assembly has good coverage of gene regions.

The sequencing coverage of the assembled genome sequence was evaluated by mapping the raw sequencing data back to the scaffolds using SOAPaligner (*27*). The peak sequencing depth was 60X and more than 87% of the assembly had at least 20-fold raw sequence coverage (Fig. S1). A scatter graph of GC content versus sequencing depths (Fig. S2) shows a typical distribution for Illumina data, and the GC content distribution in pigeon is similar to other avian genomes (Fig. S3).

To improve the quality of the annotation, we also sequenced six RNA-seq libraries using Illlumina RNA-seq technology. RNA samples from heart and liver were extracted for sequencing from three different birds: the Danish tumbler used for the reference genome, plus an Oriental frill and and a racing homer. These RNA-seq data were used in the annotation pipeline (see below for methods and Table S2 for sequencing statistics).

Annotation

We used Tandem Repeats Finder (*28*) to identify tandem repeats across the genome. Transposable elements (TEs) were identified using an approach combining both homology-based and *de novo* predictions. We identified known TEs in the pigeon assembly using RepeatMasker (http://www.repeatmasker.org) and RepeatProteinMask with the Repbase TE library. RepeatModeler was used to perform *de novo* predictions. The percentage of repetitive content in

the pigeon genome was 8.7%, which is similar to other avian genomes (*29-31*); however, we expect that the unassembled regions of the pigeon genome are also enriched in repeats.

Homology information, transcription information from RNA-seq data, and de novo predictions were integrated to annotate the protein coding genes of pigeon genome. First, protein sequences from *Gallus gallus*, *Homo sapiens*, and *Taeniopygia guttata* were used to perform homology-based gene predictions on the pigeon assembly. The homology-based pipeline included following steps: 1) homology searching against a non-redundant collection of protein sequences using TBLASTN with a E-value cutoff of 1E-5; 2) selection of the most similar proteins for each region with homologous protein matching; 3) exclusion of regions with homologous blocks shorter than 50% of query proteins; 4) use of Genewise (version 2.0) (*32*) to generate gene structures based on the homology alignments. Output gene models with a Genewise score of less than 70 were discarded. We also did pairwise whole genome alignments for pigeon and the other species to determine the syntenic blocks between them, using LASTZ (http://www.bx.psu.edu/~rsharris/lastz/). The gene models located in syntenic blocks were considered high quality genes. Three homolog-based predictions (based on proteins from *Gallus gallus*, *Homo sapiens*, and *Taeniopygia guttata*) were merged; for a given locus, the longest gene model was selected. Gene models located in non-syntenic regions that had no known SwissProt function were discarded. The merged "homology-based gene set" served as the starting point for the additional analyses described below.

RNA-seq reads were mapped to the genome by Tophat (*33*), and then Cufflinks (*34*) was used to assemble transcripts and predict open reading frames (ORFs). Transcript-based gene models with intact ORFs that had no overlap with the merged homology-based gene set were added to a merged gene set. If a transcript-based gene model with an intact ORF covered more than one homology-based gene, we replaced the homology-based gene with the transcript-based model. Transcripts without intact ORFs were used to extend incomplete homology-based gene models to find start and stop codons. The gene set improved by transcript evidence was considered the "homology-transcript gene set".

Augustus (*35*) and Genscan (*36*) were used for *de novo* gene prediction. The predicted gene models from these two programs were then merged by GLEAN (*37*). *De novo* gene models that had a known SwissProt function and did not overlap with the homology-transcript gene set were added to the annotation.

Due to limitations of automated annotation, some genes might be missed. Potentially missed genes – those that are present in the gene sets of chicken, turkey, and zebra finch but absent from pigeon – were identified from gene family analyses (see below). Protein sequences of potentially missed genes were used by Genewise to perform an additional round of homology-based gene predictions. The output gene models with transcript support and Genewise scores >70 were added to the pigeon gene annotation set. Genes related to transposons usually have many copies and can affect the subsequent analyses. Therefore, after functional annotation (see below), we removed the gene models containing transposon-related InterPro domains from the gene set, and curated some problematic gene models.

In summary, annotation of the pigeon genome identified 17,300 genes (Table S8), and all but 1,928 gene predictions are found in other avian genomes. 817 of the 1,928 contain homology to genes outside Aves and/or contain an identifiable protein domain, 1,111 have no homology or identifiable domains, but 115 of these have at least one splice site confirmed by RNA-seq data; thus, few gene predictions are good candidates for pigeon-specific protein coding genes. Fig. S4 compares general features of genes of the pigeon and other avian species.

The CEGMA pipeline (*37*) was used to assess the quality of the protein-coding gene annotation set. Of the 248 core eukaryotic genes in CEGMA, 197 genes were predicted in pigeon, and all the predicted genes can be found in our final annotation set. We compared the gene models predicted by CEGMA with the gene models by our annotation pipeline, and calculated the overlapping ratio for each CEGMA gene model (overlapping cds length / CEGMA cds length). Of the 197 predicted CEGMA genes, 166 had an overlapping ratio of at least 80% at the CDS level (Table S9).

Functions were assigned to annotated genes based on best alignments (minimum aligned coverage ≥50%) to the SwissProt database (release 15.10) (*38*) using BLASTP (Table S10). The motifs and domains of genes were determined by searching InterPro databases (v29.0) (*39*), including Pfam, PRINTS, PROSITE, ProDom, and SMART databases. GO terms for each gene were obtained from the corresponding InterPro entry. All genes were aligned against KEGG proteins (release 60.0) (*40*), and the pathways in which the gene might be involved were derived from the best matched protein in KEGG.

We used tRNAscan-SE (*41*) and INFERNAL (*42*) to predict ncRNAs in the pigeon genome (Table S11). tRNA genes were predicted by tRNAscan-SE with eukaryote parameters. rRNA fragments were identified by aligning rRNA template sequences from human using BLASTN with an E-value cutoff of 1E-5. miRNA and snRNA genes were predicted by INFERNAL software using the Rfam database (release 15.01) (*43*). To accelerate the analysis, a rough filtering was performed before INFERNAL by using BLASTN against the Rfam sequence database with an E-value cutoff of 1.

Construction of gene families

To examine the evolution of gene families in birds, genes from four avian and one lizard species (*T. guttata*, *C. livia*, *G. gallus, M. gallopavo,* and *Anolis carolinensis*) were used to construct gene families by Treefam (*44*). First, all-versus-all BLAST was performed for the protein sequences of the five species with an E-value cutoff of 1e-7. After conjoining the fragmented alignment for each gene pair by Solar (a program in Treefam), the alignments were used to calculate the distance between two genes. Next, a hierarchical clustering algorithm was used to cluster all the genes, with the following parameters: min_weight=20, min_density=0.34, and max_size=700. We found that most gene families are shared among all four birds (Fig. S5).

Functional enrichment analyses

Functional enrichment analyses were performed based on the methods described in Huang et al. (*45*). Chi-square test and Fisher's exact test (for small samples) were used to calculate the statistical significance of enrichment. For each functional class, a p-value was calculated representing the probability that the observed numbers of counts could have resulted from

randomly distributing this class between the tested gene list and the reference gene list. The p-values were adjusted by FDR and the adjusted p-value cutoff was 0.05. For GO enrichment, to remove redundancy, if the GO terms enriched at different levels with parent-child relationship and had the same gene list, the lowest level was chosen and other levels were filtered.

We found some false positives in the enrichment analysis that were due to fragmented or partial genes (e.g., one gene is split into two or more genes because of assembly gaps or annotation errors). Fragmented/partial genes may lead to a larger size of their corresponding gene family and thus result in a false signal of over-representation in the enrichment tests. Therefore, before performing the enrichment tests, we filtered the putative fragmented genes based on the SwissProt annotation. The filtering criteria were: 1) in the alignment results against SwissProt database, the query (gene in pigeon) length was shorter than half of the target length, and 2) the percent identity of the alignment was >50% (suggesting good homology). Ultimately, 1507 genes were filtered from the gene set for enrichment analyses.

Expansion and contraction of gene families

We used CAFE (*46*) to identify the clustered gene families that have undergone expansion or contraction in the pigeon relative to other birds. Some gene families identified as expanded or contracted by CAFE might be due to artifacts (incorrect automated annotation, parameter bias during clustering, etc.). Therefore, we performed a closer check on families of interest (preliminary candidates for expansion or contraction) after running CAFE and corrected the members of each family if needed. Ultimately we found 2 expanded gene families and 2 contracted gene families in the pigeon (Tables S15-S16). We constructed phylogenetic trees of these families using the WAG model in PhyML (*47*) (Figs. S6-S9).

Gene loss

Based on the gene clustering results, a gene was considered to be lost in pigeon if it was present in the chicken, turkey, and zebra finch genomes but absent from the pigeon genome. To ensure that putative losses were not due to incorrect clustering or incomplete annotation, we realigned these genes against the integrated gene set (homology-transcript and *de novo* gene predictions) and genome assembly. Gene predictions that had good homology (Genewise score >60 and no frame shift) but failed to pass the gene prediction criteria, and had expression support (average coverage depth of RNA-seq reads ≥1), were not considered lost.

In addition to the above screening criteria, we also checked the genes that flank the putatively lost genes. For a given gene loss candidate, if the flanking genes were included in the pigeon assembly and these flanking genes had conserved synteny between pigeon and other birds, the absence of the gene from pigeon was deemed to be unlikely due to incompleteness of the assembly. Thus, we filtered the candidates that had no synteny support from any of the other 3 bird genomes. For each remaining gene loss candidate, we required that at least one of its three upstream genes and at least one of its three downstream genes in another bird be assembled in the same scaffold/contig of pigeon assembly, and that these assembled flanking genes were syntenic between pigeon and the other bird. Ultimately, 67 gene families that had no annotated homolog in pigeon, but did have at least one homolog in the other 3 birds, passed the synteny criteria. We used the homologous genes of these families (204 genes in 67 families) in the other 3 birds to perform enrichment analyses (Tables S17-S19).

Pseudogene identification

In order to identify genes that might be pseudogenized due to mutations in coding sequences, we used the gene set of the zebra finch to identify homologous genes in pigeon, and those genes with frameshifts or premature terminations were considered as candidate pseudogenes. To ensure that candidate pseudogenes were not due to assembly errors, we verified that putative mutation sites were consistent with the corresponding bases in the raw reads used in the genome assembly. If inconsistent, we filtered the corresponding candidates. Moreover, if the mutation site in the transcriptome assembly was not the same as that in genome assembly, or there was an alternative spliced form, the candidate also would be discarded. Table S20 lists putative pseudogenes in the pigeon genome, and Tables S21-S23 summarize functional enrichment of this list.

Resequencing and variant calling

Non-reference genomes were sequenced using paired-end libraries on the Illumina HiSeq2000 (all *C. livia* genomes, sequenced at BGI or the University of Utah) or Genome Analyzer IIx platform (*C. rupestris* only, CoFactor Genomics, St. Louis). Raw sequencing depth ranged from 8- to 26-fold coverage. We concatenated the reference assembly scaffolds into 9 pseudo-chromosomes to facilitate alignment of the resequenced genomes (*48*). We then aligned the raw reads to the reference by SOAPaligner (v2.21) and sorted the alignment results according to chromosome coordinates. We discarded multiple-hit alignments (reads that be mapped to more than one locus). For variant calling, we first converted SOAP alignment results into SAM format, and used a custom script to retain short (≤5 bp) indel information. Then we used the "pileup" command in SAMtools to call variants (SNPs and short indels). Next, we filtered the variants using "samtools.pl" (version 0.3.3, a helper script in SAMtools), with parameter set "varFilter -S 20 -i 50 -d 3 –D 50". We observed that some SNPs or indels were very close together, which was probably attributable to alignment errors. Therefore, we filtered variants separated by ≤5 bp. Finally, we converted the coordinates of the variants in the pseudo-chromosomes to the coordinates in scaffolds and contigs in the reference assembly.

We performed an additional round of variant processing to filter on depth and quality. First, the SAMtools pileup files were converted to Genome Variant Format (GVF). Then, the mean depth was calculated for each scaffold and contig. These data were used as lambda to model depth with the Poisson distribution. Variants with a depth less than 5 or greater than the 98% quantile were masked, as were variants with a Phred scaled quality below 20. These masked variants comprise the "no-call" category of Fig. S12. The final variant data set included 22,020,759 single nucleotide variants, 1,246,896 small (≤5 bp) insertions, and 71,495 small deletions in the 40 *C. livia* genomes compared to the Danish tumbler reference (Fig. S10F). Deletions are underrepresented in the filtered data set due to low sequence coverage in flanking regions. This bias is probably not biologically meaningful and deletion variants were not used in subsequent analyses. VAT (in the VAAST (*18*) pipeline) was used to annotate the variant effects based on a draft GFF annotation file. All 41 *C. livia* and *C. rupestris* GVF files were then condensed to the internal VAAST condenser format (CDR) by the VST function in VAAST. We used the proportion of masked variants as an estimate of the called proportion of each genome, which ranged from 71% to 93%.

Diminishing returns of novel variants

To examine the number of new variants discovered in each successive genome sequenced (Fig. S10F), genomes were randomly sampled in bin sizes ranging from 1 to 40. The quantity of novel variants was counted in each group, and this process was repeated 100 times. No-calls were treated as reference alleles and the outgroup *C. rupestris* was excluded from this analysis.

Phylogenetic tree

SNP sites with genotype data for all 41 resequenced birds were used to create a binary matrix of presence and absence data relative to the reference genome assembly (1.48 million loci). This matrix was used in the R statistical environment (*49*) to generate a neighbor-joining tree using the APE (*50*) phylogenetic library. The tree was bootstrapped by sampling the binary matrix 1,000 times using the "boot.phylo" command (Figs. 1, S16).

ADMIXTURE analysis

We used the rapid, maximum likelihood algorithm in ADMIXTURE (*13*) to estimate proportion of group membership across different values of K (number of putative ancestral clusters of allelic similarity). A PLINK (*51*) genotype file was generated from the CDR file for SNP loci with complete genotype information (no missing data). As a conservative control for linkage disequilibrium (LD), the data matrix was pruned to include sites at least 100 kb apart. A mean $r^2 << 0.2$ was observed at this distance (see Fig. S10J). This filtering resulted in a matrix of 10,026 sites for the 41 *Columba* genomes (*C. livia* and *C. rupestris*). Q-matrices generated for individuals in ADMIXTURE were displayed graphically using DISTRUCT (v1.1) (*52*) (Fig. S19). We performed a second analysis that included data from *C. livia* genomes only (Fig. S17). In the absence of *C. rupestris*, we used PLINK to exclude variant sites with MAF<0.10 to reduce the effects of rare variants on the analysis (had we done this in the complete dataset that included *C. rupestris*, we would have removed many of the unique variants that distinguished *C. rupestris* from *C. livia*). This filtered dataset included 3950 sites. We repeated the ADMIXTURE analysis, and the results of both analyses from K=1-10 are shown in Figs. S17-S20.

Linkage disequilibrium

To avoid biases in linkage disequilibrium (LD) calculations due to rare alleles, we filtered the pigeon resequencing data to include only biallelic sites and alleles with frequencies between 0.30 and 0.70. The outgroup species *C rupestris* was excluded from this analysis. Pearson's correlation (r^2) was then calculated for every pairwise SNP comparison at distances between 1 bp and 1 Mb across all contigs and scaffolds in the genome assembly (*53*). Human samples (YRI) were taken from the October 2011 release of the 1000 Genomes Project and randomly subsampled to correspond to the pigeon group size (40 individuals). The r^2 values were aggregated by distance using the mean. A 500-bp sliding window was then applied, and every 100^{th} data point was plotted in Fig. S10J to smooth the curves.

Mutation rate in the pigeon lineage

We used TBLASTX (*54*) alignments ($E < 10^{-8}$) to identify one-to-one orthologs between chicken, zebra finch, and pigeon. Chicken was aligned to zebra finch and pigeon separately. We then parsed the BLAST reports with custom perl scripts to identify four-fold degenerate codon positions shared between the three species. This procedure generated three-way alignments of 1,271,075 fourfold degenerate sites from 7690 orthologous genes. We ran MODELTEST (*55*)

using these alignments and found that the General Time Reversible (GTR) substitution model fits best with the observed data. By setting the divergence time between pigeon and zebra finch to 85.5 million years ago (*56*) and running the baseml script in the PAML package (under the GTR model) (*57*), we estimated the mutation rate in the pigeon lineage after the divergence from zebra finch to be $1.42x\ 10^{-9}$ substitutions $site^{-1}$ $year^{-1}$ ($\pm\ 2.60x10^{-12}$ SE). Concurrent estimates for the mutation rates in zebra finch and chicken lineages are $2.40x10^{-9}$ and $1.90x10^{-9}$ substitutions $site^{-1}$ $year^{-1}$, respectively, which agree well with previous estimates ($2.21x10^{-9}$ and $1.91x10^{-9}$ substitutions $site^{-1}$ $year^{-1}$) (*58*).

Time to most recent common ancestor (TMRCA)

We inferred the demographic history of the pigeon population using ∂a∂i (*59*), an inference method based on a diffusion approximation to the observed allele frequency spectrum. We started with the simplest model of a constant-size pigeon population, and then gradually switched to more complex models, using the Akaike Information Criterion (AIC) to choose the best-fit model. We found that a three-epoch model fits significantly better than less complicated models. Further increasing the complexity does not improve the model. In this three-epoch model, the effective population size for the rock pigeon increased from 95,000 to 760,000 approximately 1.50 million generations ago, then remained constant until very recently, when a large decrease in population size occurred. The 95% Confidence Interval for the decrease in population size ranges from 1 to 90 generations ago. We suspect that low-coverage depth data that we generated for the 40 resequenced genomes might bias against the discovery of rare variants, which in turn could create an effect that mimics a recent reduction in population size. However, we also suspect that the recent history of inbreeding in domesticated pigeon breeds at least partially accounts for the population size decrease in the best-fitting model. In summary, the estimated effective population size for the rock pigeon over the last 1.5 million years is approximately 760,000, but because of a very recent bottleneck, the current effective population size is estimated to be 520,563.

Under the three-epoch model, we estimated the mean TMRCA value for all 40 resequenced rock pigeons by running 10,000 coalescence simulations with the program ms (*60*) and calculating the mean TMRCA value from all simulations. This resulted in an estimated TMRCA for all rock pigeons at an average genomic locus of 1.65 million years.

To generate the confidence interval for the statistics reported here, we used two different approaches. First, starting with the maximum likelihood (ML) model that ∂a∂i derived from the observed frequency spectrum, we selected one parameter at a time, and introduced a small deviation from the ML estimate given by ∂a∂i. Under the null hypothesis that the new model describes the data equally well as the ML model, -2 x log-likelihood ratio of the two models should asymptotically conform to a chi-square distribution with 1 degree of freedom. This allows us to calculate the confidence interval of each parameter using the Composite Likelihood Ratio Test (CLRT). This approach does not account for the correlation between loci (i.e., linkage disequilibrium), but we expect the correlation to be minor given the size of the pigeon genome. To ensure that we accounted for correlation at linked sites, we employed a second approach to calculate CIs by bootstrap-sampling genomic contigs from the pigeon assembly. Within each bootstrap, we computed the frequency spectrum based on the sampled genomic regions, and used

∂a∂i to generate ML estimates for demographic parameters. In Table S28, we report combined results from the two approaches by providing conservative estimates based on both approaches.

Shared variation among crested birds

Although crested birds do not appear to form an exclusive clade or share high allelic similarity (Figs. 1, S19), it is possible that the genomes of crested birds might share variants and haplotypes at a higher rate than other, random groupings of genomes. This potential for genetic structure among crested birds could lead to an excess of false-positive or uninformative signals of shared allele frequency (e.g., F_{ST}) and extended homozygosity. To measure diversity of the head crest group and compare it to other random groups of pigeon genomes, *C. livia* genomes were randomly binned into a group size of 8 (the number of genomes from crested birds), and the numbers of shared variants were counted. This process was repeated 10,000 times, yielding a normal distribution (Fig. S25A). The 8-genome bin containing the crested birds falls well within the normal distribution, suggesting this group is not highly structured. Because the crested group contains two birds from the same breed (Indian fantail), we also repeated the analysis using a bin size of 7, so that we could assess the number of shared variants in each of the Indian fantails plus the other 6 crested breeds (Fig. S25B).

F_{ST} analysis

F_{ST} was calculated for bi-allelic sites using the method of Weir and Cockerham (*61*). We excluded SNP sites at which <50% of birds were genotyped. After sorting genomes into crested and uncrested bins, we further excluded SNP sites that had >25% no-calls in either the crested or uncrested bin. In total, 17,500,439 SNP sites were used in the analysis. The outgroup species *C rupestris* was excluded from this analysis. Distribution of F_{ST} statistics is shown in Fig. S22A.

Cross-population extended haplotype homozygosity (XP-EHH)

We converted the CDR file to BEAGLE (*62*) format, and then to XP-EHH format to generate an input data file. To calculate XP-EHH, a script was retrieved from the Prichard Lab (University of Chicago) website: http://hgdp.uchicago.edu/Software/. The program was run with default settings and treated the 8 birds with head crests as one population (archangel, English trumpeter, 2 Indian fantails, mookee, Iranian tumbler, oriental frill, Jacobin) and all other resequenced birds as another population. Genome-wide XP-EHH scores are plotted in Fig. S21, and distribution of XP-EHH statistics is shown in Fig. S22B.

Haplotype network analysis

BEAGLE was used to phase and impute genotypes for sites with no more than 7% masked data and allele frequencies between 0.30 and 0.70. Phased haplotypes around the SNP at scaffold 612:596613 (*cr* locus) were aligned manually to identify a 27.4-kb haplotype (40 SNP loci) shared by all 8 crested birds in the resequencing set (Fig. S23). Two uncrested birds were heterozygous for the derived T allele at *cr*, further refining the *cr* haplotype to 11 kb (19 SNP loci; Fig. 2). Haplotype networks were generated using TCS (v1.21) (*63*), with the connection limit equivalent to the number of variant sites in the haplotype. For subsequent visualization of the haplotype tree, CLUSTAL W (*64*) was used to create a multiple-sequence alignment dendrogram, which was found to be consistent with the network generated in TCS.

TaqMan genotyping assay for the *cr* SNP in *EphB2*

DNA was extracted from 10uL of blood of an additional 61 crested birds from 22 breeds and 69 uncrested birds from 57 breeds as described (*12*). Samples were diluted to 10 ng/μL and genotyped using TaqMan SNP genotyping assay (Applied Biosystems, Foster City, CA) on an ABI GeneAmp PCR System 9700 at the University of Utah Microarray Core Facility. Primers used to amplify the target sequence were 5'-CGGCGGGCATGAAATACCT-3' and 5'-CAGACCAGGTTGCTGTTCAC-3', and the reporter sequences were 5'-ATGTTGCGGGCAGCC-3' and 5'-ATGTTGCAGGCAGCC-3'. Birds from the following breeds were genotyped for the wild-type and *cr* variant at scaffold 612:596,613:

Crested – archangel, Bokhara trumpeter, classic oriental frill, crested Saxon field color pigeon, Danzig highflier, English trumpeter, fairy swallow, Franconian trumpeter, Indian fantail, Iranian tumbler, Jacobin, medium-faced crested helmet, mindian fantail, mookee, nun, Old Dutch capuchine, Old German owl, oriental frill, Russian tumbler, saint, schmalkaldener soorhead.

Uncrested – African owl, Altenburg trumpeter, American show racer, American flying tumbler, Berlin long-faced tumbler, Bohemian pouter, Brunner cropper, Budapest tumbler, carneau, cauchois, Chinese owl, cumulet, domestic show flight, Dragoon, Egyptian swift, English baldhead long-faced clean-legged tumbler, English carrier, English long-faced muffed tumbler, English magpie, English short-faced tumbler, fantail, French mondaine, frillback, German nose-crested trumpeter, Holle cropper, horseman pouter, ice pigeon, Italian owl, king, Lahore, Lebanon, little Spanish friar tumbler, Maltese, Marchenero pouter, Modena, Norwich cropper, oriental roller, parlor roller, Pomeranian pouter, Portuguese tumbler, racing homer, Lebanon, runt, Saxon monk, Saxon pouter, Scandaroon, Shaksharli, Spanish barb, starling, Syrian Baghdad, Texas pioneer, Thai laugher, Thuringer clean leg, Vienna medium-faced tumbler, Voorburg shield cropper, West-of-England tumbler, and zitterhals.

Whole-mount in situ hybridization

To generate probes for RNA *in situ* hybridization, RNA was isolated from four-day post-laying pigeon embryos using the RNeasy kit (Qiagen, Valencia, CA), and cDNA was synthesized using M-MLV-RT (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. Linear templates for probe synthesis were amplified by PCR using the following primers: *Cttnb1* (5'-CGATGATTAACCCTCACTAAAGGGACAATGGGTGGAACACAACAG-3' and 5'-CGATGTTAATACGACTCACTATAGGGCTAGGATCATCTGGGCGGTA-3'), *EphA4* (5'-CGATGATTAACCCTCACTAAAGGGAAGCAACCTGGTCTGCAAAGT-3' and 5'-CGATGTTAATACGACTCACTATAGGGCCACAGCCTCTAGGGTGGTA-3'), and *EphB2* (5'-CGATGATTAACCCTCACTAAAGGGACGGGACTTCTTGAGTGAAGC-3' and 5'-CGATGTTAATACGACTCACTATAGGGCTGTGCTCTCATCACCTGGA-3'). Binding sites for T3 and T7 polymerase (underlined) were incorporated into the forward and reverse primers to facilitate subsequent transcription of sense and antisense probe, respectively.

Embryos used for RNA *in situ* hybridization were dissected from eggs, and fixed overnight in 4% paraformaldehyde at 4°C on a shaking table, then dehydrated into 100% MeOH and stored at -20°C. RNA *in situ* hybridization was performed as described (*65*), with larger wash volumes

and longer wash times used to accommodate the large size of the embryos. Hybridization with a sense probe was performed as negative control.

Supplementary Text

Geographic origins of breeds

Our phylogenetic and ADMIXTURE analyses includes several breeds that were not used in previous studies of pigeon relationships (*12*), including some breeds that were only recently exported from the Middle East. These breeds provide a geographic anchor to infer the origins of other breeds. For example, the fantail breeds probably have been in India for at least 2000 years (*14*), yet they show a close genetic association with three ancient breeds from Iran: the Shakhsharli, Iranian tumbler, and Lahore (also known in Iran as the Sherazi (*14*)). This affinity suggests that the ancestors of (or major genetic contributors to) fantails might have been imported from Iran and Turan (central Asia) via longstanding trade routes between these two regions (*66*). Similarly, the owl breeds (Fig. 2a, red branches) are closely related to three ancient breeds from the eastern and southern Mediterranean region, supporting their hypothesized origins in Asia Minor and Northeast Africa (*14, 67*).

The *cr* allele segregates in a cross

The presence or absence of a head crest is often an important part of a breed standard (*68*). Breeders typically cull birds not meeting this standard because they will not be competitive at shows. However, we found a notable exception to a breed standard that segregated the *cr* allele and phenotype in a cross. We genotyped a small pedigree of American show racers, an uncrested breed, in which two uncrested parents produced both crested (n=2) and uncrested (n=1) offspring. As expected, we found that both parents were +/*cr*, the crested offspring were *cr*/*cr*, and the uncrested offspring was +/*cr*. In summary, homozygosity for the *cr* allele is perfectly associated with the crest phenotype across 79 diverse breeds of domestic pigeon (see main text) and in an unusual cross.

Author Contributions

M.D.S., G.Z., M.Y., M.T.P.G., and J.W. planned the project. Sample collection, sequencing, assembly, and annotation of the reference genome was conducted by S.C.A.N, E.W., M.T.P.G. G.Z., C.L., H.P., H.T., Haofu Hu, and supervised by M.T.P.G. and G.Z. H.P. conducted the gene content and enrichment analyses and H.T. produced the initial SNP and indel variant calls. The population study was designed and supervised by M.D.S. and M.Y. S.A.S. and M.D.S. collected and prepared samples for resequencing. Z.K. performed the F_{ST}, XP-EHH, VAAST (with E.T.D.), phylogenetic, and ADMIXTURE analyses, and developed the no-call pipeline. E.T.D., Z.K., and M.D.S. performed the haplotype analysis. M.C. and M.Y. calculated the variant statistics for the reference and resequenced genomes and performed the mutation rate analysis. C.H. and Hao Hu performed the mutation rate, TMRCA, and N_e analyses and C.H. contributed to several other population genetic analyses. E.T.D. and A.I.V. designed in situ hybridization probes and performed the gene expression analyses and TaqMan assays. M.D.S and G.Z. wrote the manuscript with input from M.T.P.G., M.Y., E.T.D., C.L., Z.K., C.H., Hao Hu, E.W., and S.C.A.N. M.D.S. and J.W. are co-senior authors.

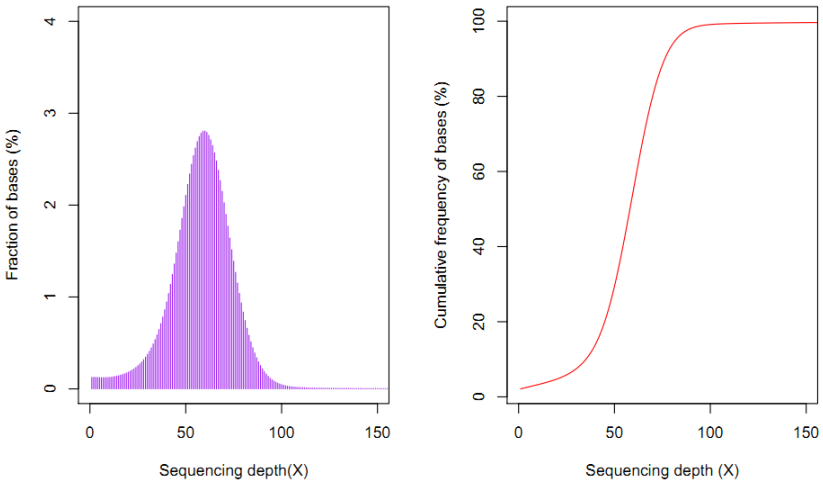

Fig. S1.

Sequencing depth distribution. The raw reads were aligned onto the assembled genome sequence using SOAPaligner, allowing 2 mismatches for 44-bp reads, 5 mismatches for the longer reads.

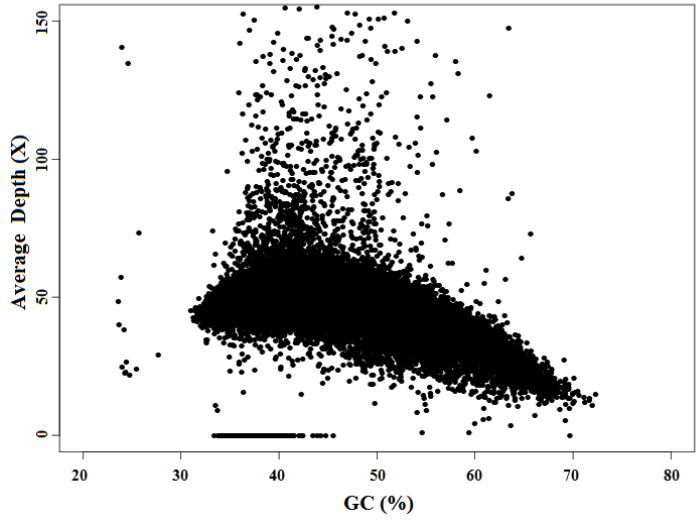

Fig. S2.

GC content versus sequencing depth. The x-axis represents GC content and the y-axis represents average depth using 10-kb non-overlapping sliding windows.

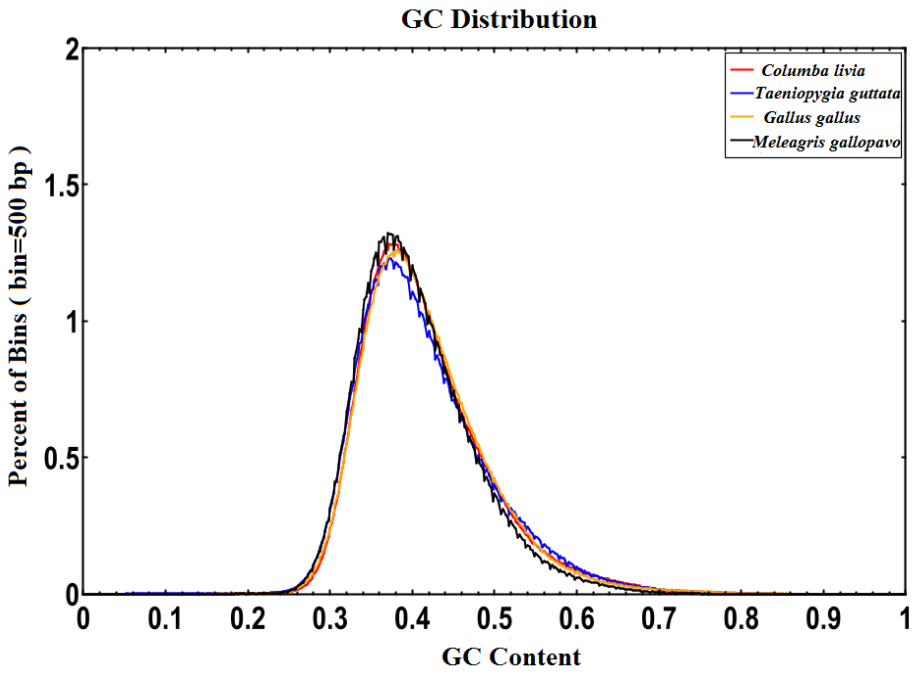

Fig. S3.

GC content distributions of 4 avian genomes. The x-axis represents GC content and the y-axis represents the percentage of 500-bp, non-overlapping, sliding windows in the genome. GC content distributions are similar among the pigeon, zebra finch, chicken, and turkey genomes.

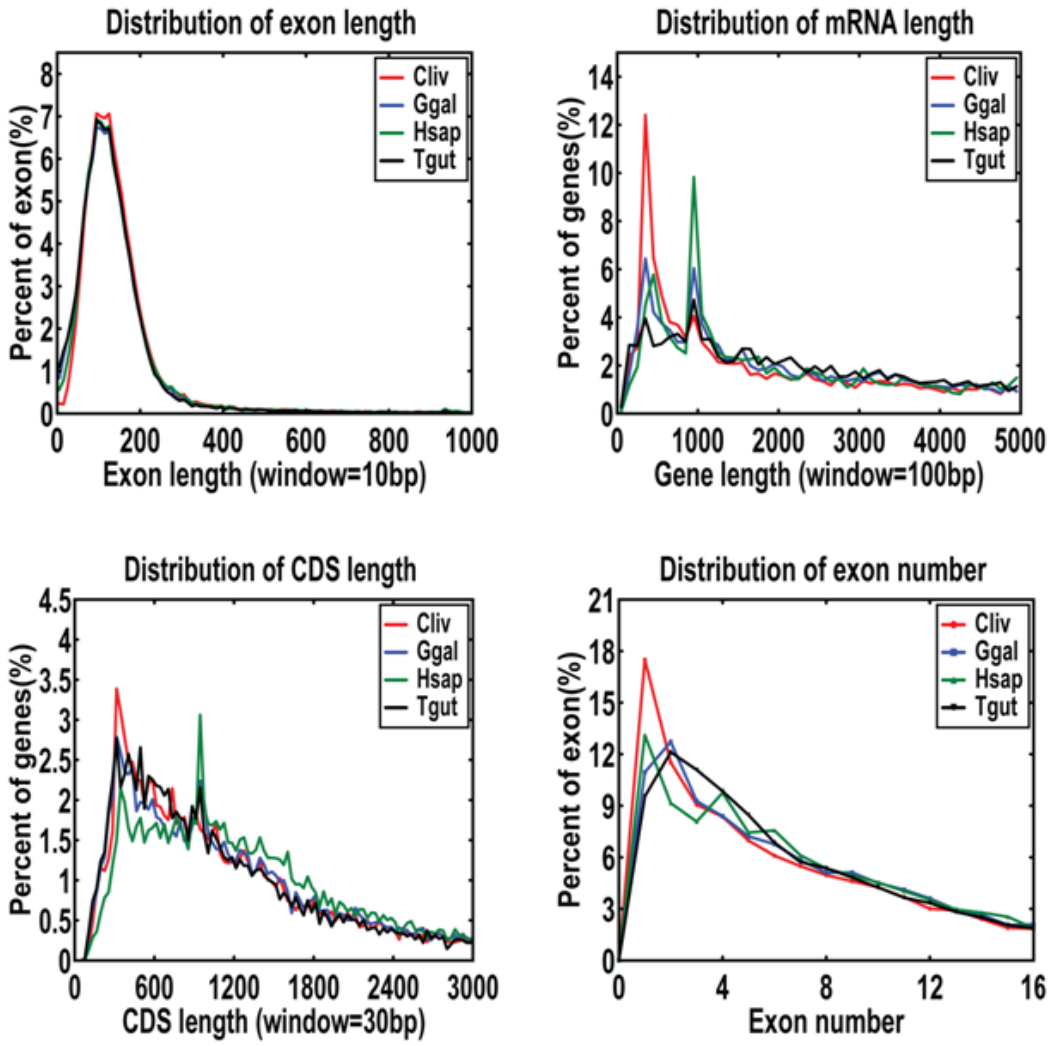

Fig. S4.

Comparison of general features of protein-coding genes. Cliv, Ggal, Hsap and Tgut are abbreviations for *C. livia*, *G. gallus*, *H. sapiens* and *T. guttata*, respectively.

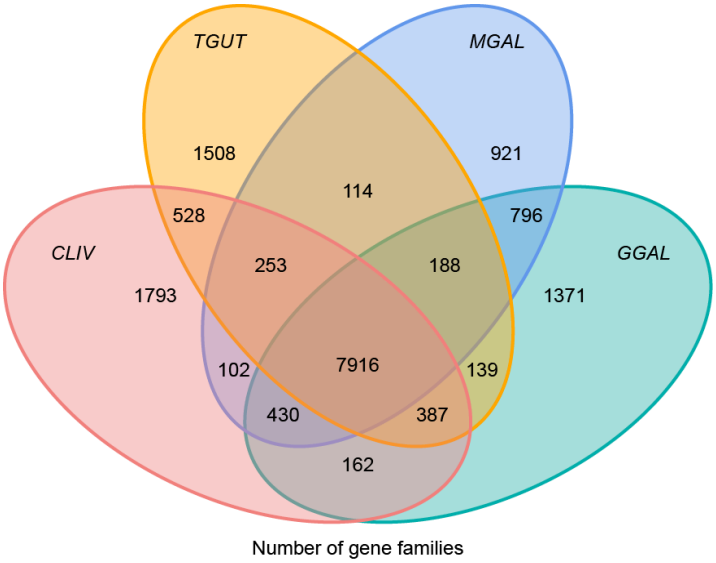

Fig. S5.

Venn diagram of gene families of four birds**.** CLIV, TGUT, MGAL and GGAL are abbreviations for *C. livia*, *T. guttata*, *M. gallopavo* and *G. gallus*, respectively.

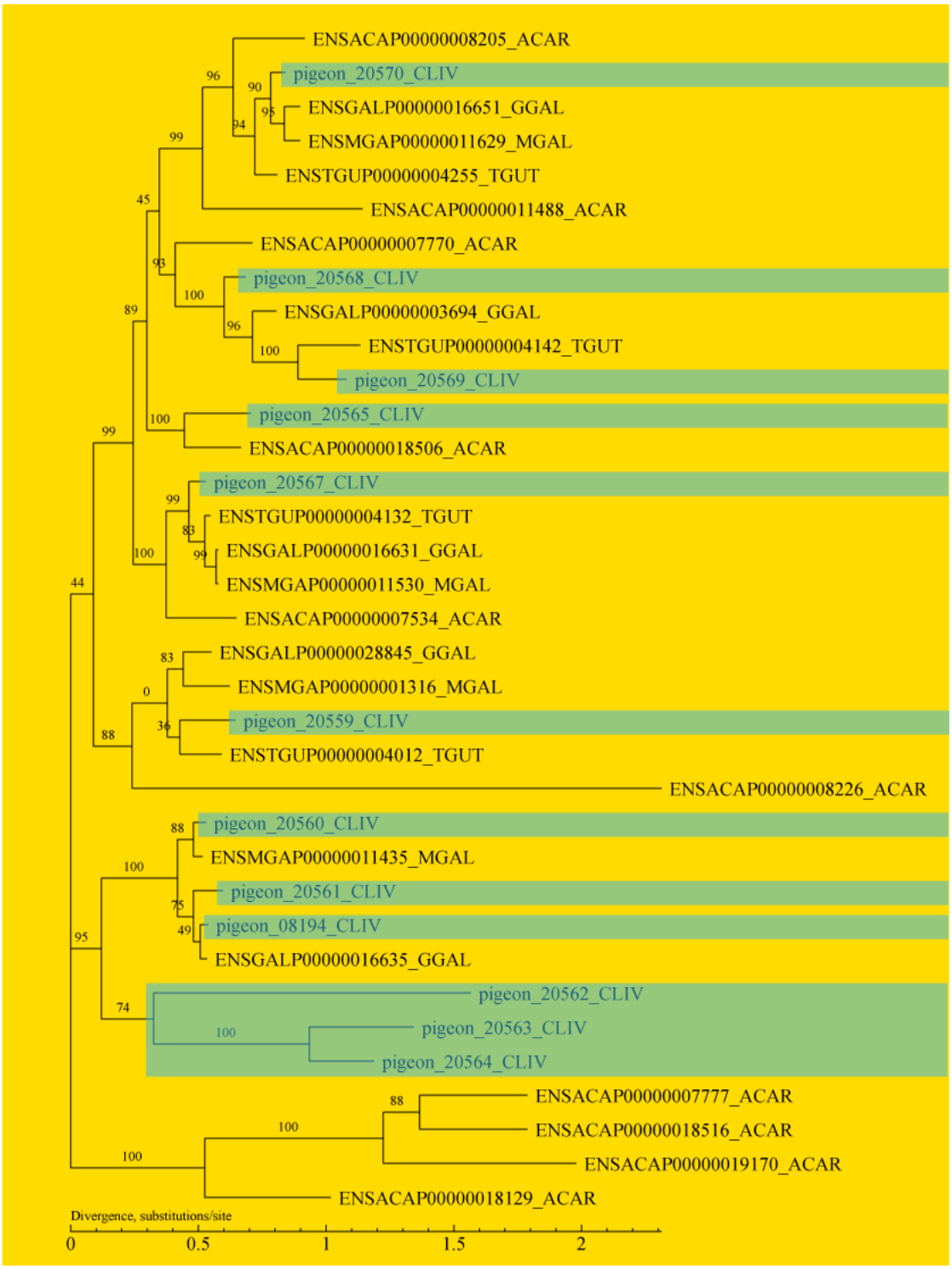

Fig. S6.

Phylogenetic tree of the gene family “type II keratin”. Tree was generated by PhyML, with parameters “-d aa -m WAG -b -4 -rates gamma”.

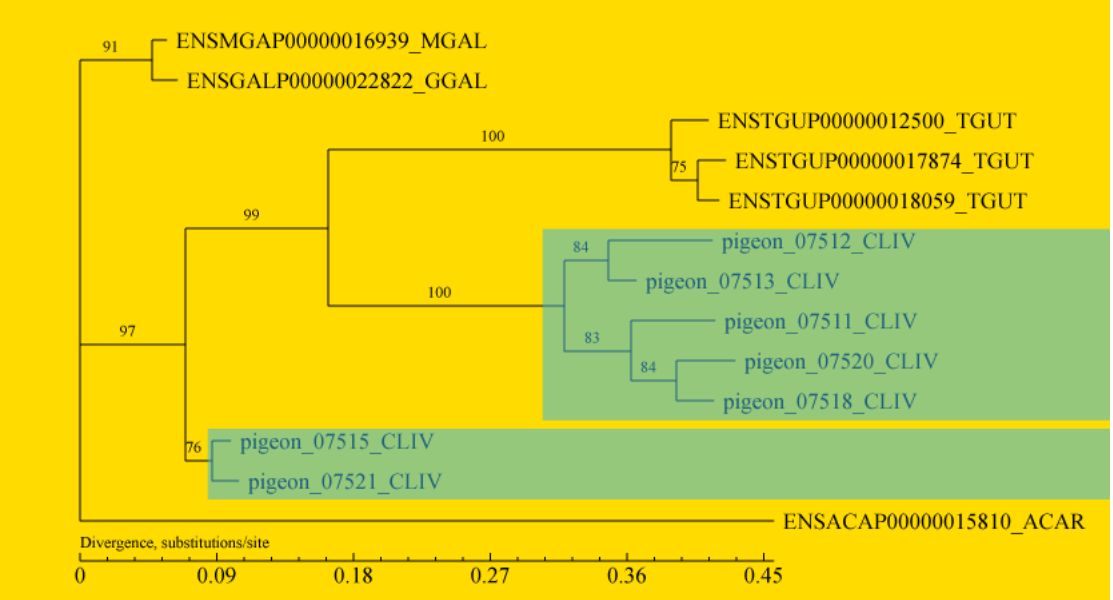

Fig. S7.

Phylogenetic tree of the gene family "lactosylceramide 4-alpha-galactosyltransferase". Tree was generated by PhyML, with parameters "-d aa -m WAG -b -4 -rates gamma".

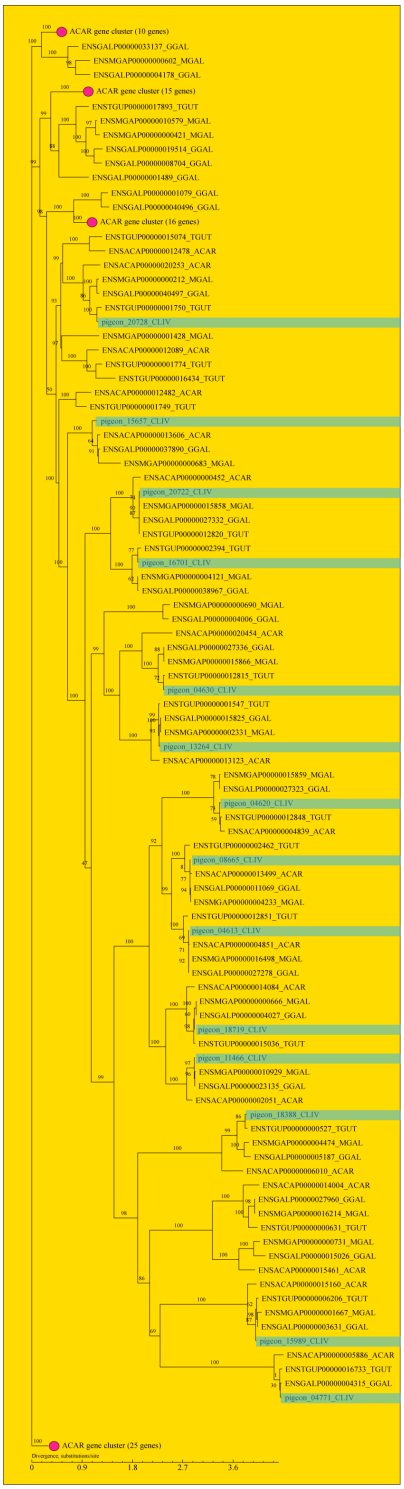

Fig. S8.

Phylogenetic tree of the gene family “protocadherin”. Tree was generated by PhyML, with parameters “-d aa -m WAG -b -4 -rates gamma”.

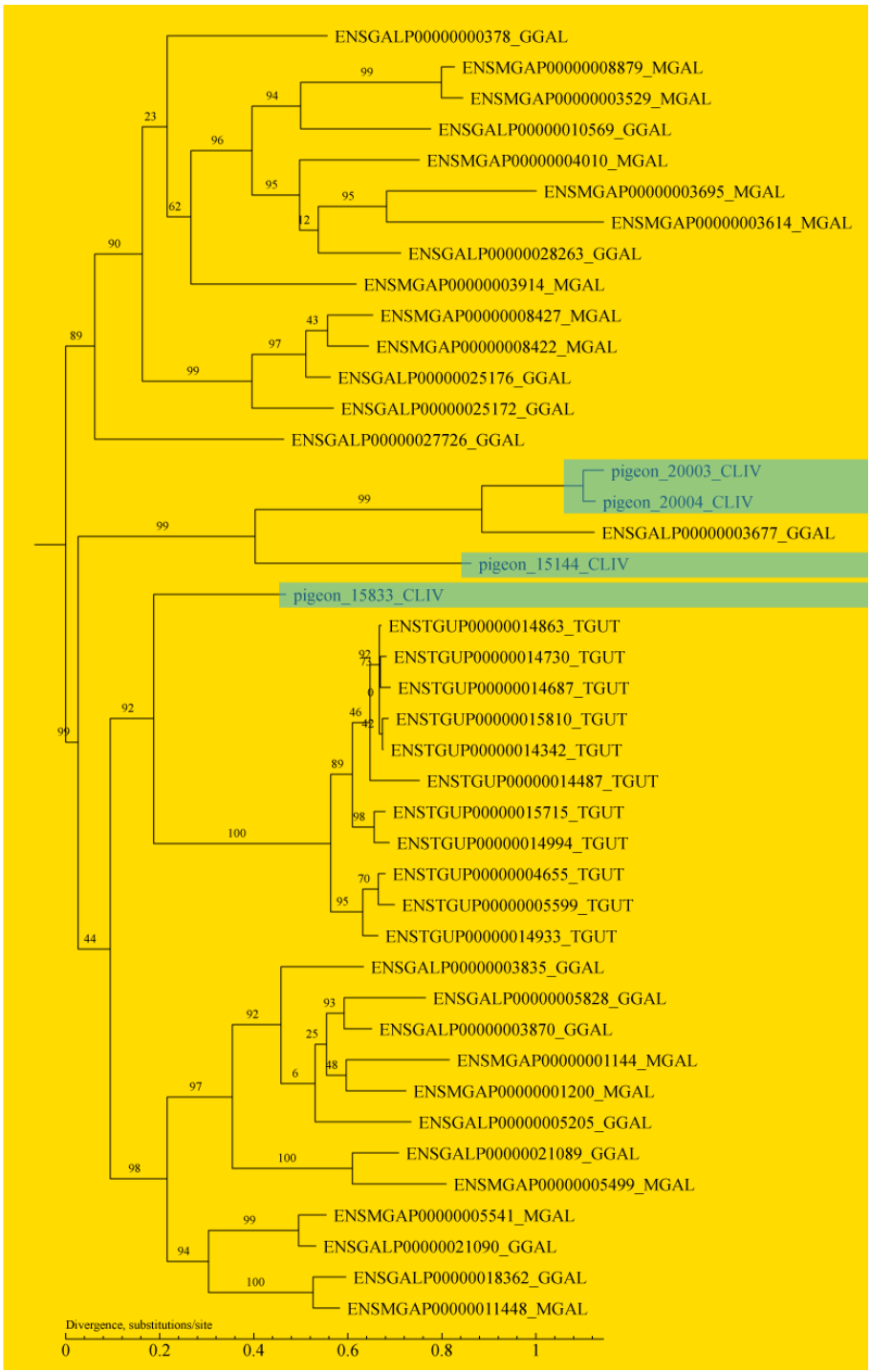

Fig. S9.

Phylogentic tree of the gene family “PHD finger protein 7”. Tree was generated by PhyML, with parameters “-d aa -m WAG -b -4 -rates gamma”. Because no homolog was found in lizard, we used the ‘root’ function in Treebest (http://treesoft.sourceforge.net/treebest.shtml) to determine the root.

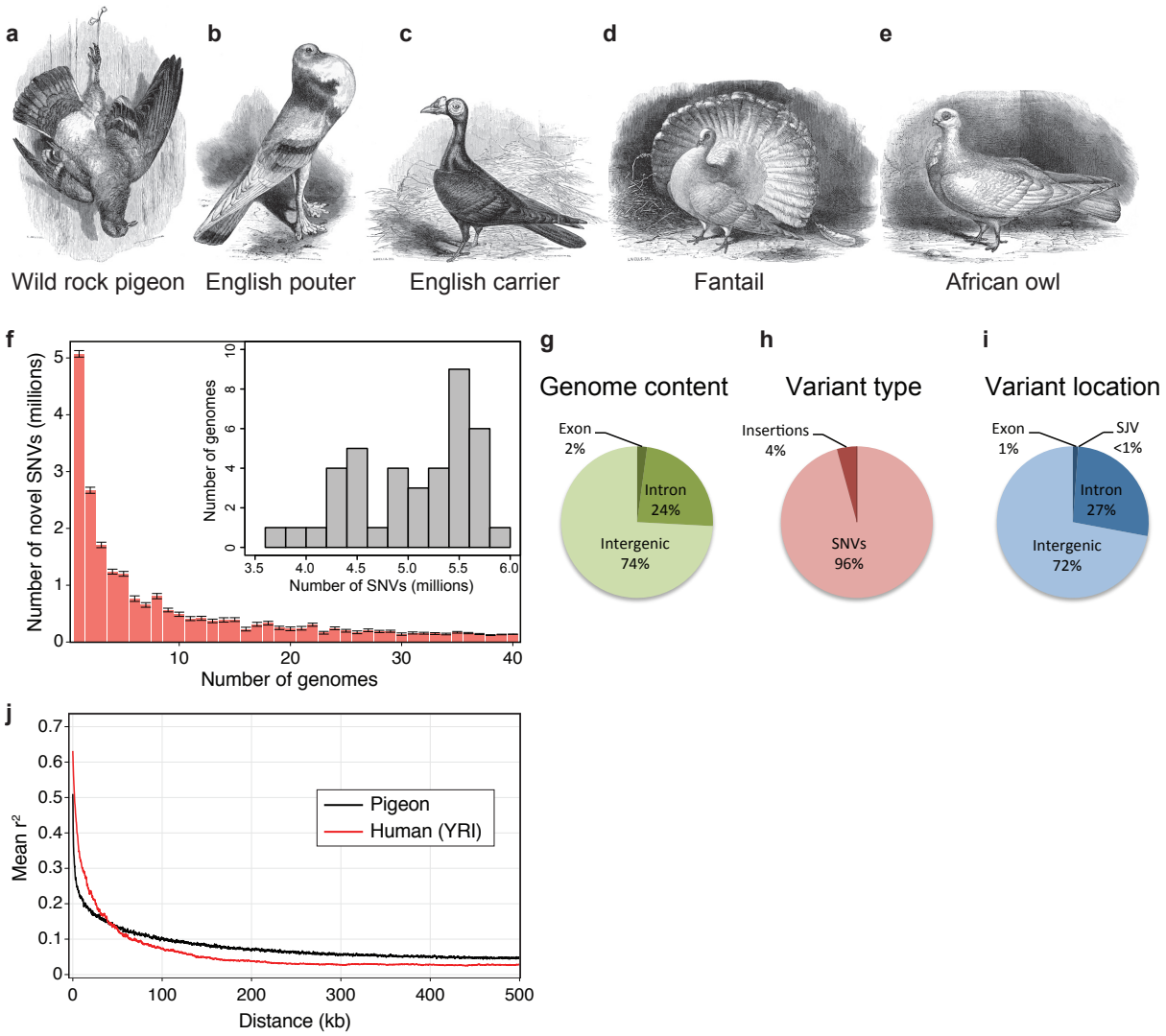

Fig. S10.

Phenotypic and genomic diversity in the rock pigeon. **a-e**, Wild type (**a**) and diverse domestic breeds of rock pigeon (**b-e**) as illustrated in Darwin's *Variation in Animals and Plants under Domestication* (*4*). **f**, The number of unique single nucleotide variants (SNVs) declines rapidly with each new rock pigeon genome sequenced, similar to a pattern observed for resequenced human genomes (*69*). Error bars are +/- SEM from 100 bootstrap replicates. Inset, frequency of SNV counts (200,000-SNV bins) across 40 resequenced *C. livia* genomes. **g**, Proportion of the pigeon reference genome composed of exon, intron, and intergenic sequence. **h**, Proportion of single-nucleotide and insertion variants in the 40 resequenced rock pigeon genomes. **i**, Location of variants in the resequenced genomes (SJV, splice junction variant). As expected, variants are found preferentially in non-coding regions of the genome. Of the variants predicted in exons, 60% are synonymous and 40% are non-synonymous. **j**, Linkage disequilibrium in the rock pigeon and an African human population. Mean r^2 across a 500-bp sliding window is plotted against genomic distance for 40 *C. livia* genomes (black trace), and 40 randomly selected genomes from the 1000 Genomes Project YRI population (red).

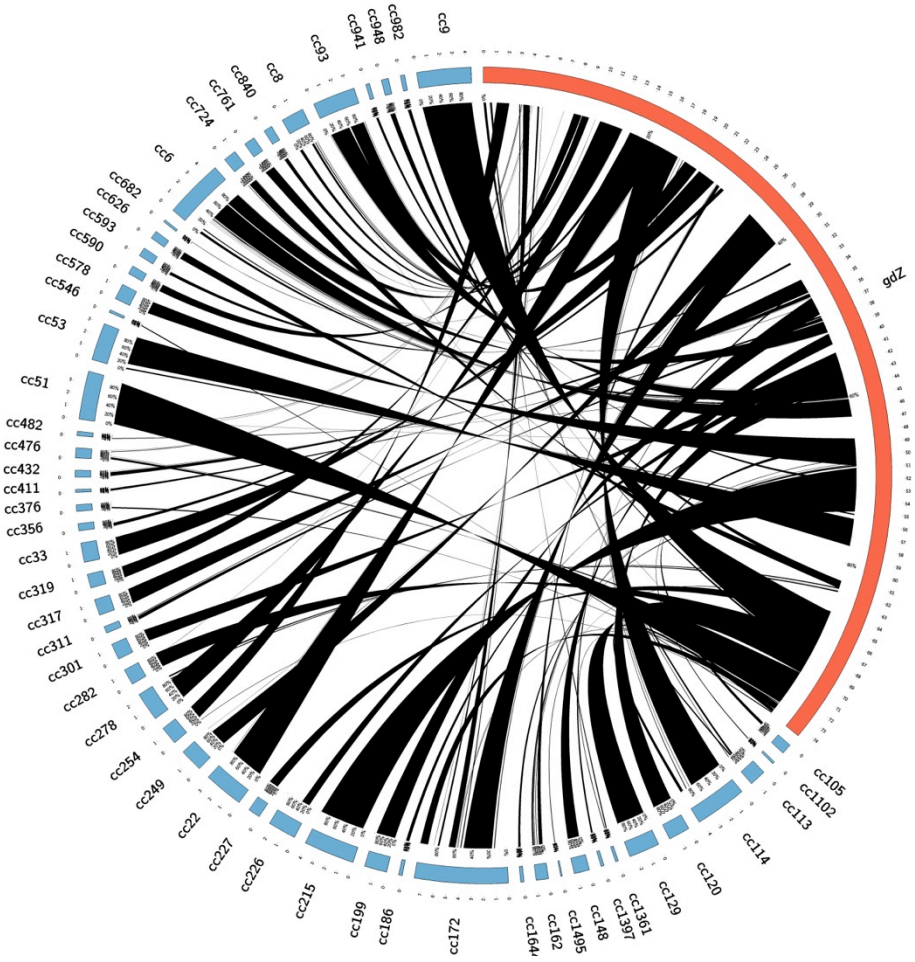

Fig. S11.

Circos plot of chicken Z-chromosome (gdZ, red) and corresponding scaffolds in the pigeon genome (blue; "cc" precedes scaffold number). Black lines show regions of high sequence conservation as aligned by BLAT (*70*). Pigeon scaffolds map to most of the chicken Z-chromosome, and most scaffolds map to a single contiguous segment of the Z-chromosome.

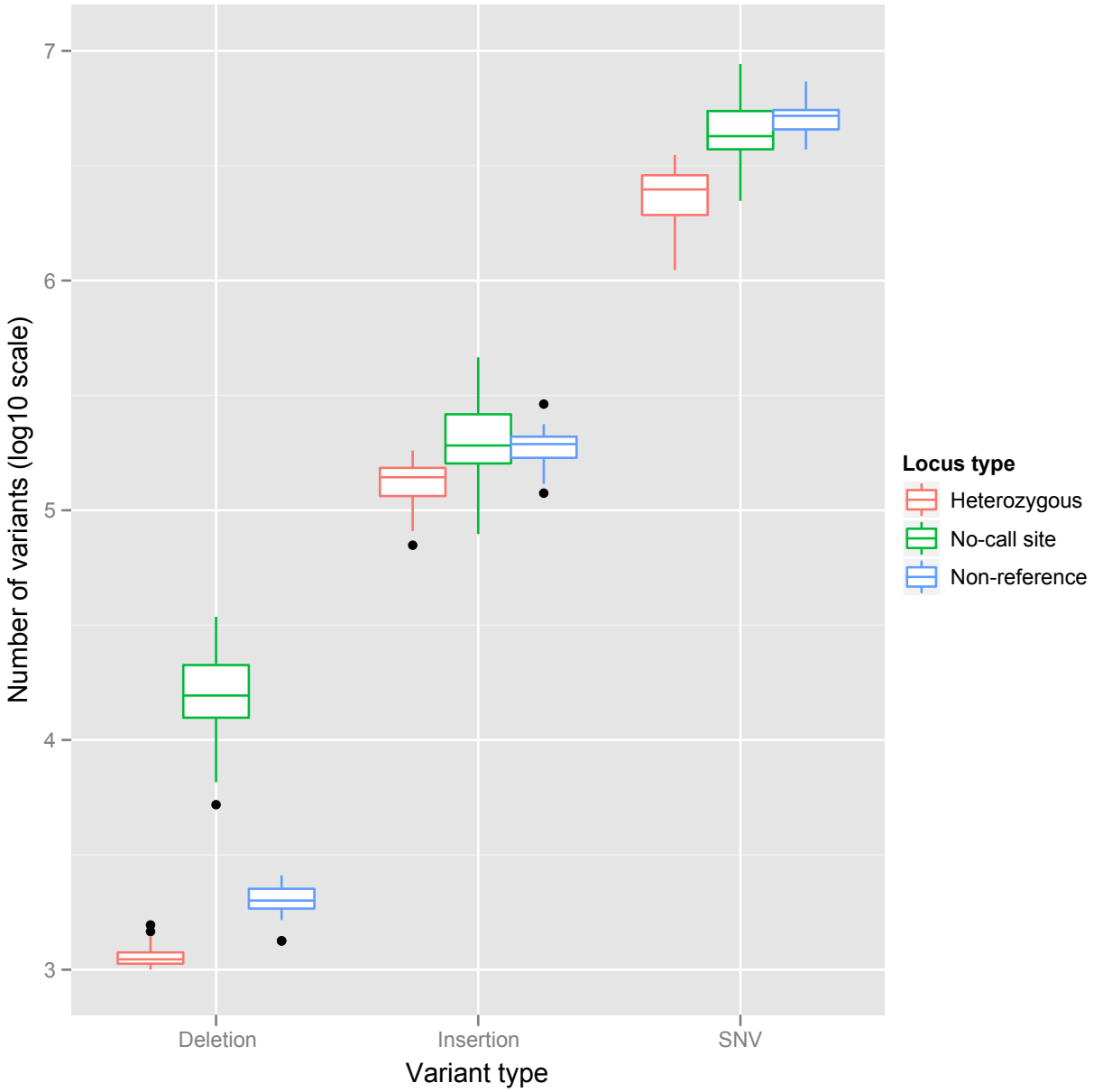

Fig. S12.

Numbers of variants in 40 rock pigeon (*C. livia*) genomes after filtering for sequencing coverage and quality. Boxes define 25% and 75% quantiles, horizontal line indicates median. A high number of "no-call" deletion sites probably resulted from low sequencing coverage in and around indel clusters.

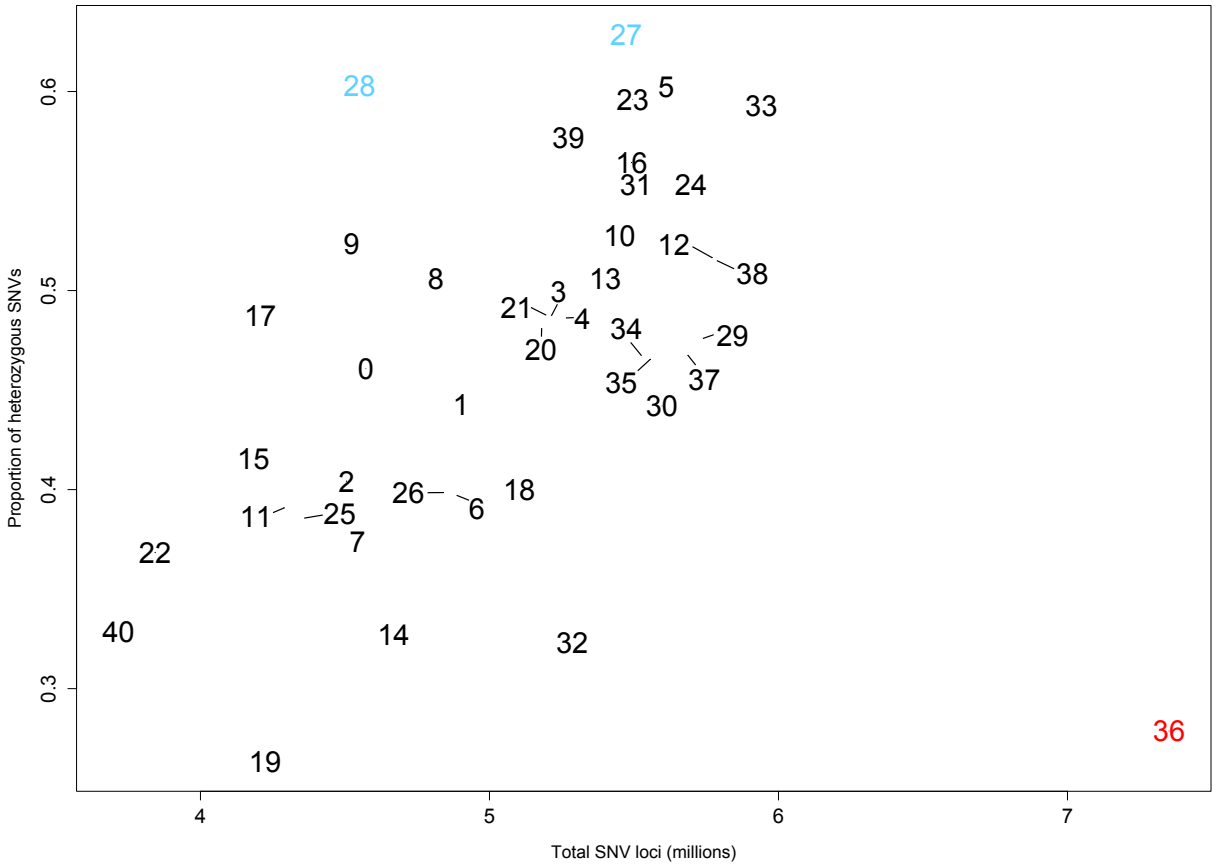

Fig. S13.

Proportion of heterozygous SNP sites in 41 resequenced genomes of *Columba livia* and *C. rupestris*. Feral birds are have a high proportion of heterozygous sites (blue numbers 27 and 28) and, as expected, the outgroup species *C. rupestris* (red number 36) has an excess of SNP loci when variants are called against the *C. livia* reference. Domestic pigeons are indicated with black numbers, which correspond to individuals from the following breeds: 0, Indian fantail; 1, African owl; 2, laugher; 3, Mookee; 4, Spanish barb; 5, starling; 6, English carrier; 7, Scandaroon; 8, Berlin long-face tumbler; 9, Birmingham roller; 10, king; 11, Chinese owl; 12, Saxon monk; 13, Syrian dewlap; 14, Shakhsharli; 15, Oriental roller; 16, Carneau; 17, English long-face tumbler; 18, English pouter; 19, Jacobin; 20, Lahore; 21, Lebanon; 22, parlor roller; 23, racing homer; 24, archangel; 25, cumulet; 26, Egyptian swift; 27, feral (Virginia, USA); 28, feral (Utah, USA); 29, ice pigeon; 30, frillback; 31, Iranian tumbler; 32, Marchenero pouter; 33, runt; 34, Saxon pouter; 35, English trumpeter; 36, *Columba rupestris* (wild, UWBM 59803); 37, fantail; 38, Indian fantail; 39, racing homer; 40, oriental frill.

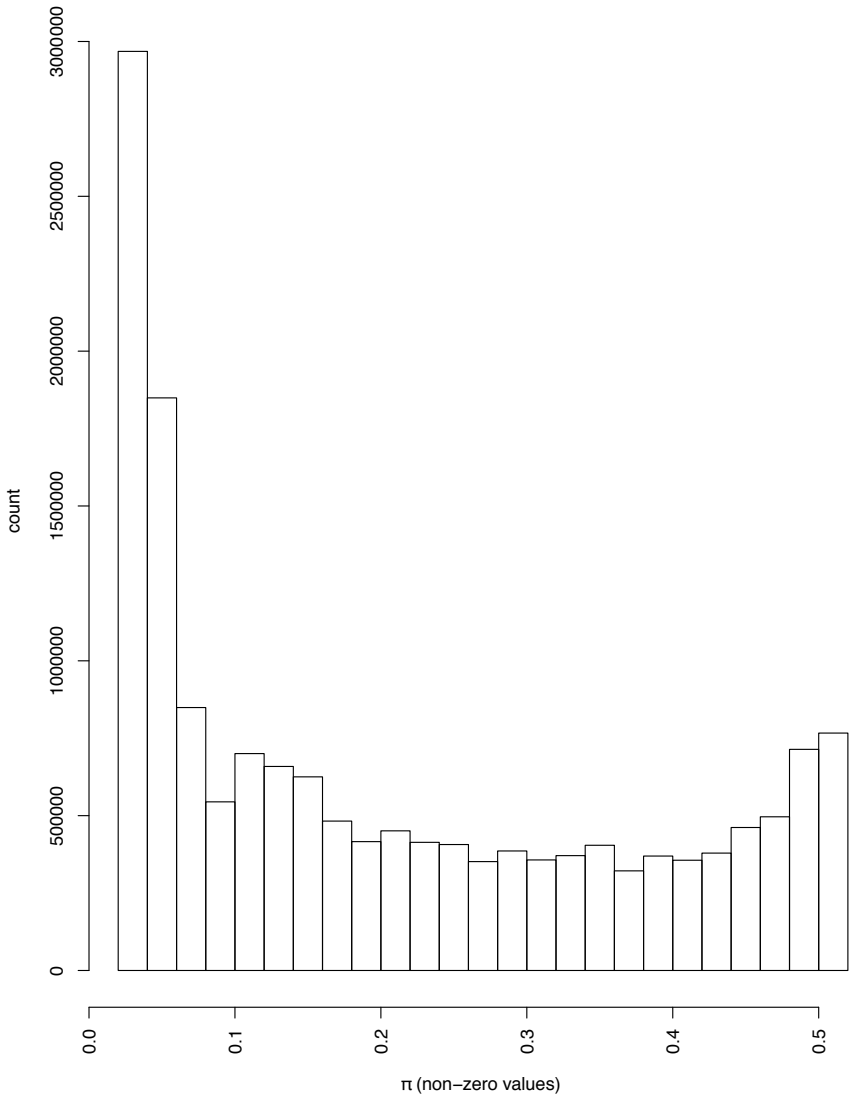

Fig. S14.
Genome-wide distribution of nucleotide diversity (π) among 40 genomes of domestic and feral rock pigeons.

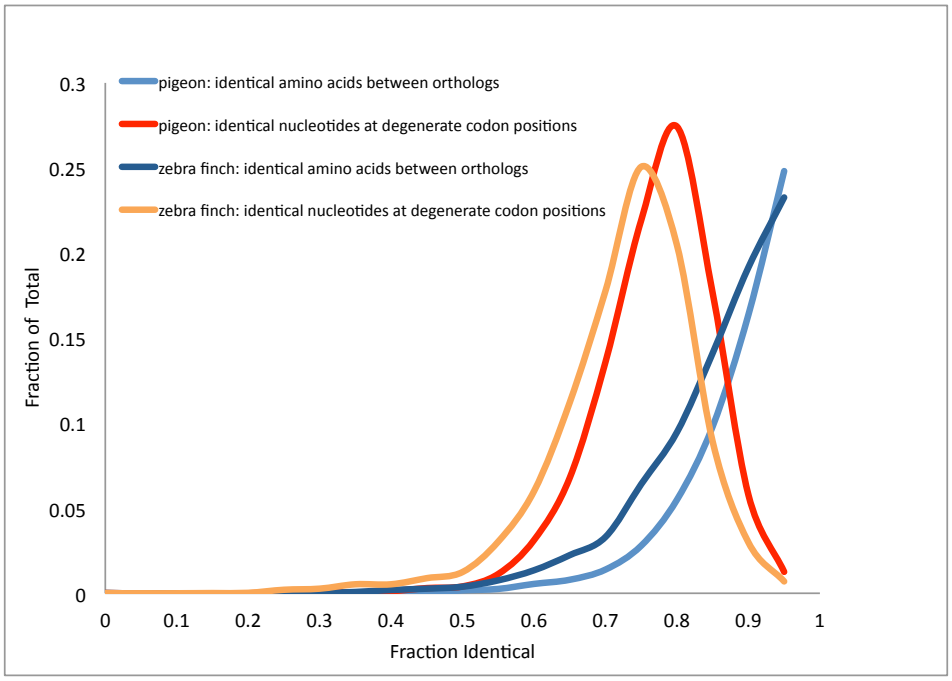

Fig. S15.

Distribution of identical amino acids and identical nucleotides at fourfold degenerate codon sites between pigeon-chicken and zebra finch-chicken orthologs.

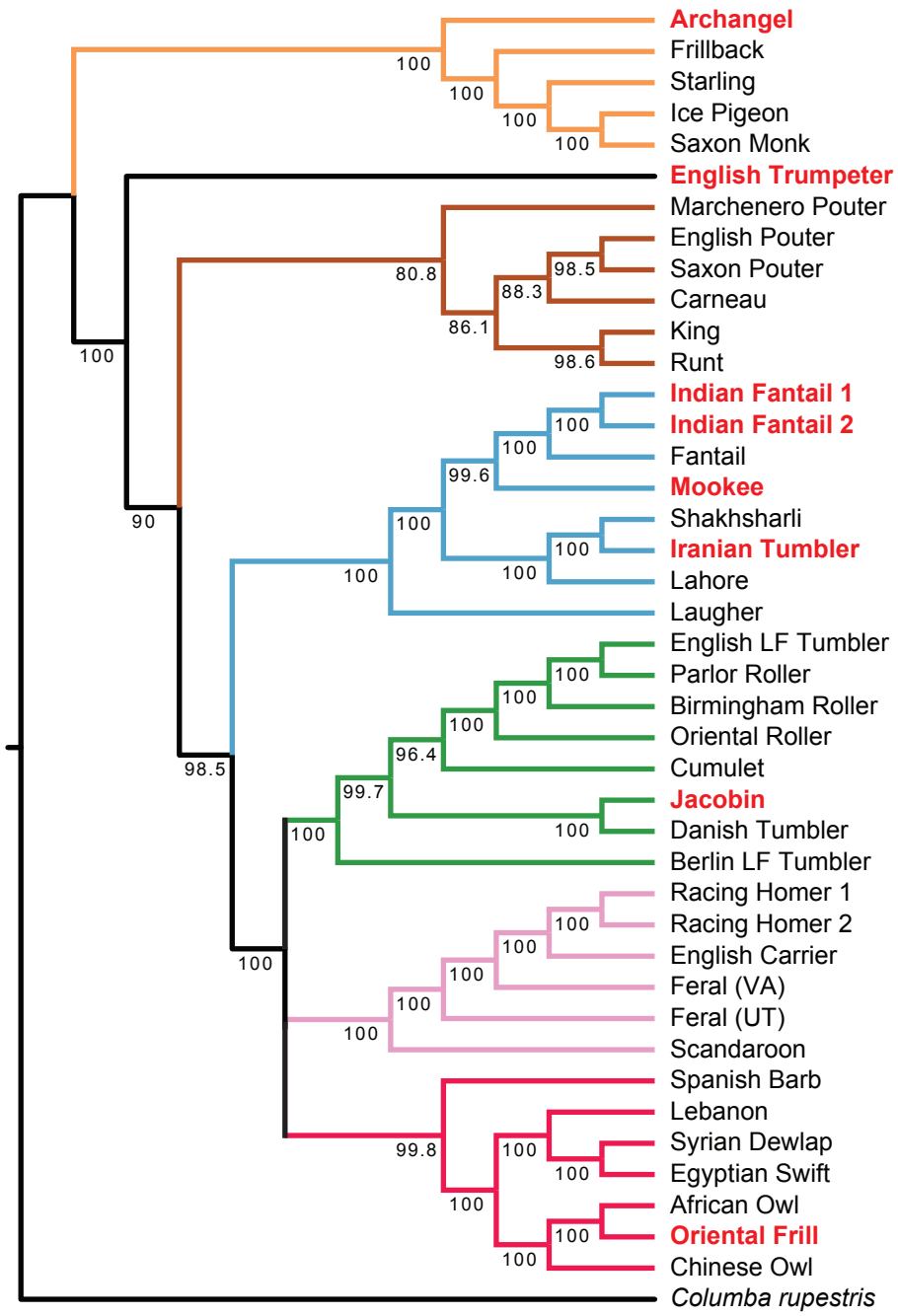

Fig. S16.

Neighbor-joining tree of domestic and feral *Columba livia* and sister species *C. rupestris* based on genotypes from 1.48 million SNP loci. This diagram emphasizes the topology of the tree and branch lengths are not to scale. Percent bootstrap support (>50%, based on 1000 iterations) is indicated on branches. Breeds with head crests are indicated with bold, red lettering.

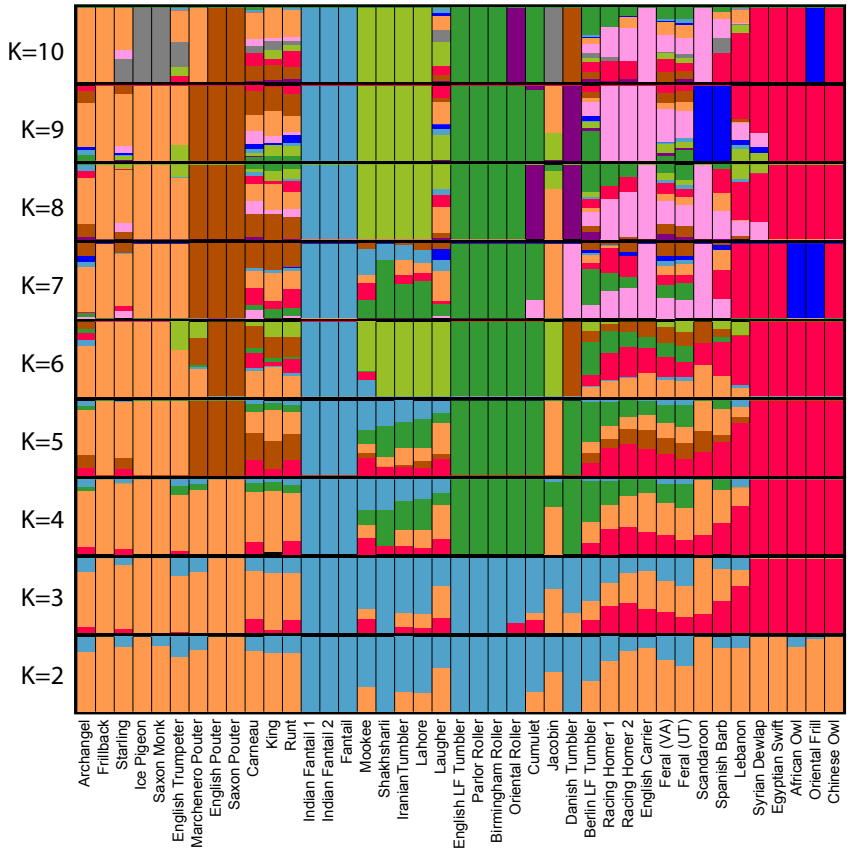

Fig. S17.

ADMIXTURE plot for rock pigeon genomes (excluding the outgroup *C. rupestris*). For this analysis, 3950 SNP loci with MAF>0.10 were included to examine genetic structure within the rock pigeon only. CV error data suggest that that K=1 is the most likely number of populations (see Fig. S18); however, higher K values are biologically informative about allelic similarity among breeds as well (for example, patterns of population membership at K=6-8 are similar to groupings in the tree in Fig. 1). Several breeds were inconsistent in their cluster assignments, including very ancient breeds (laugher, cumulet, Jacobin, Spanish barb, runt) and recent hybrids (English trumpeter, Carneau, king, Berlin long-face tumbler, racing homer). The modern racing homer was derived from the cumulet, owl, carrier, and other breeds approximately 200 years ago, and this recent admixture is evident at K≥2.

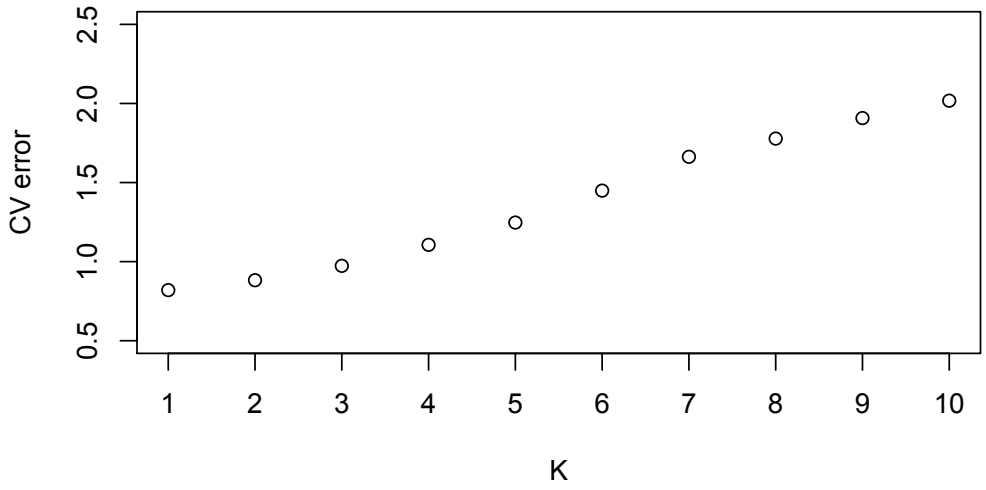

Fig. S18.

Plot of cross-validation (CV) errors in ADMIXTURE for each value of K between 1 and 10 in an analysis of 41 *C. livia* genomes. 3950 SNP loci with MAF > 0.10 were included. Lower CV errors indicate a better model fit. Thus, our data imply a best fit at K=1, or a single population.

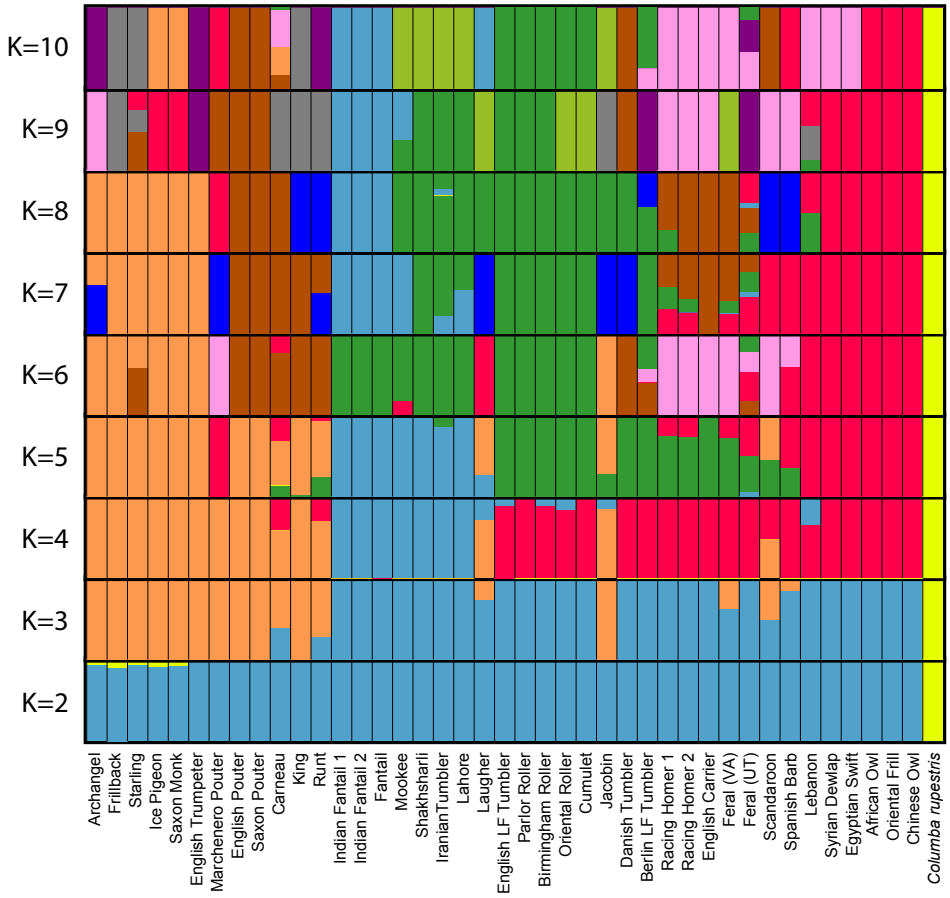

Fig. S19.

ADMIXTURE plot indicating proportion of membership of each bird in each of K putative ancestral populations for K=2 to K=10. Dataset includes the reference genome and all 41 resequenced *Columba* genomes and 10,026 SNP sites. CV error data suggest that that K=1 is the most likely number of populations (see Fig. S20). At K=2 and higher, the outgroup *C. rupestris* is distinct from the *C. livia* breeds.

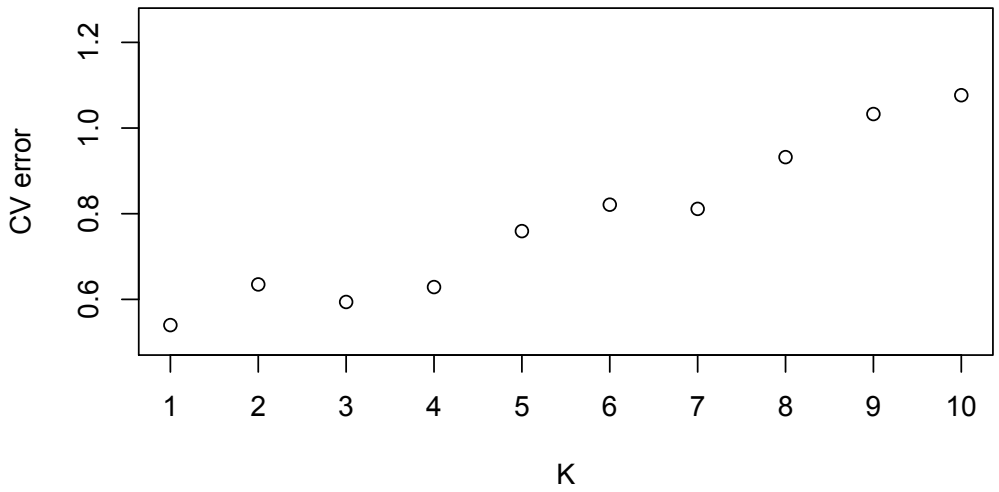

Fig. S20.

Plot of cross-validation (CV) errors in ADMIXTURE for each value of K between 1 and 10 in an analysis of all 42 *Columba* genomes. Lower CV errors indicate a better model fit. Thus, our data imply a best fit at K=1, or a single population.

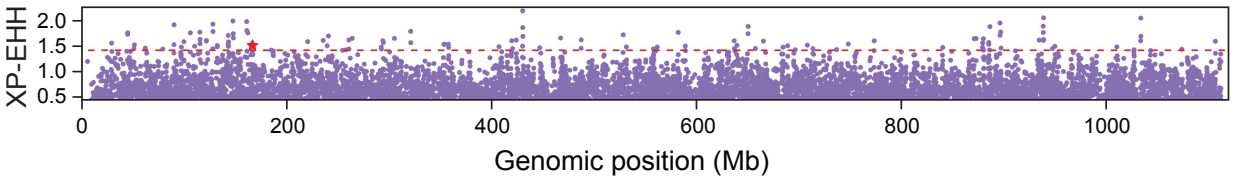

Fig. S21.

Genome-wide cross-population extended haplotype homozygosity (XP-EHH, unstandardized). The window of highest F_{ST} (Fig. 2B) corresponds to a position in the top 1% of XP-EHH scores (red star), suggesting positive selection in crested birds.

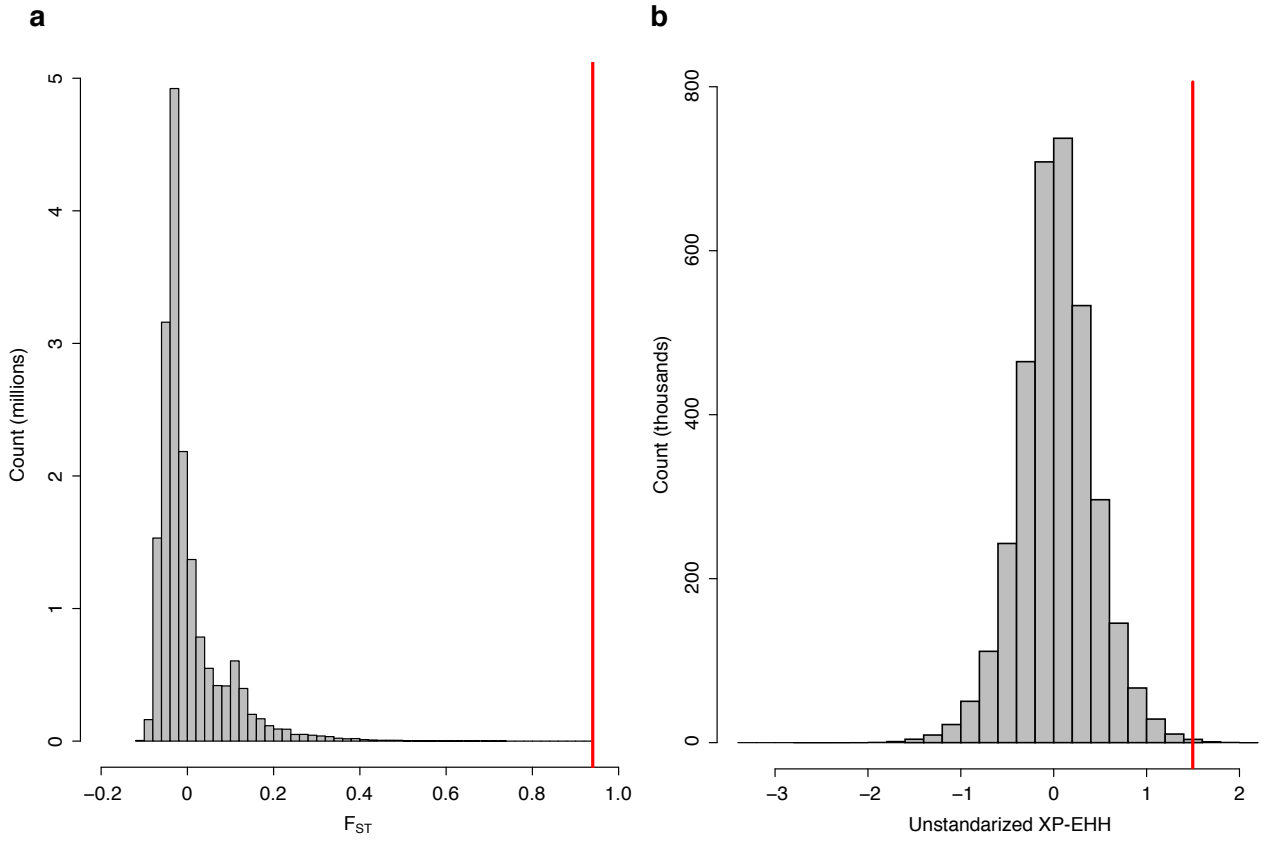

Fig. S22.

Distribution of F_{ST} (**a**) and XP-EHH (**b**) statistics in the comparison between genomes of crested and non-crested pigeons. Red lines indicate scores at the *cr* locus (*EphB2*) on scaffold 612.

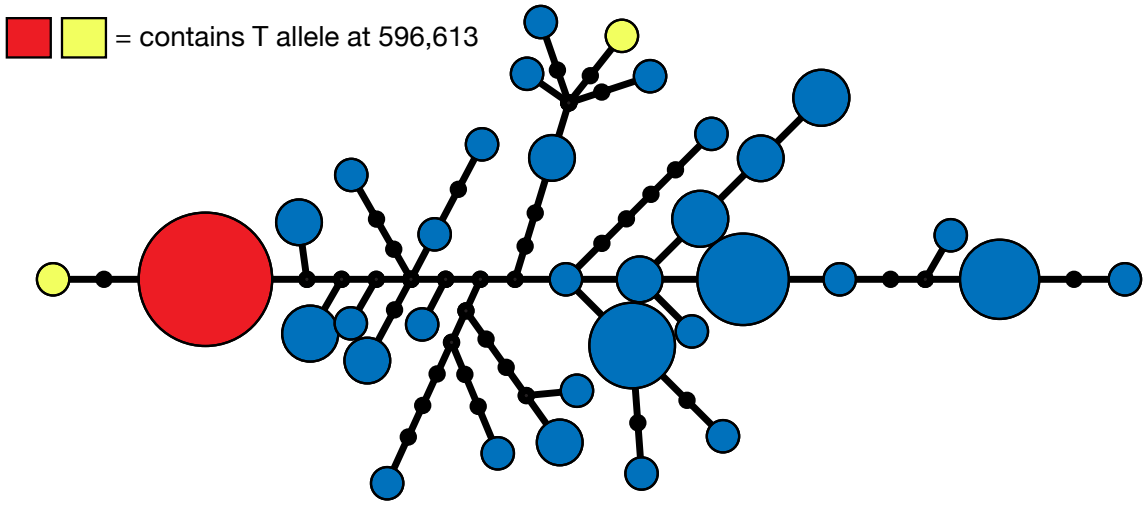

Fig. S23.

Haplotype network diagram of a 27.4-kb interval around the *cr* locus on scaffold 612. All crested birds in the resequencing set were homozygous for a 27.4-kb haplotype (red), and two uncrested birds were heterozygous for haplotypes containing the T allele at scaffold 612:596,613 (yellow). Haplotypes in uncrested birds without the T allele are shown in blue. Sizes of circles are proportional to the number of chromosomes containing a haplotype, and line segments (separated by dots) represent single nucleotide changes. All haplotypes with the T allele share an 11-kb haplotype (see Fig. 2), and the apparent divergence of the yellow haplotype at the top of the diagram is due to recombination with another haplotype.

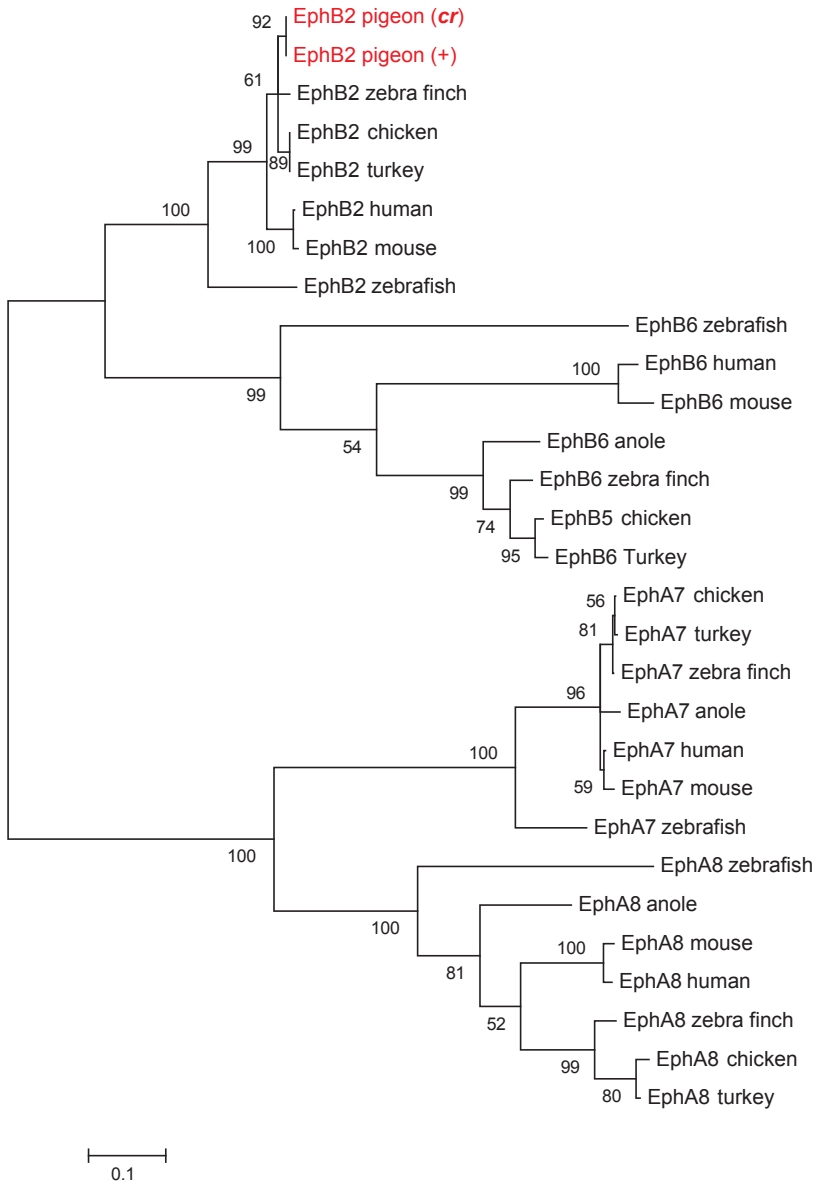

Fig. S24.

Unrooted maximum likelihood tree of vertebrate Eph receptor protein sequences. Pigeon *EphB2* alleles are more closely related to *EphB2* orthologs of other vertebrates than to other *EphB* or *EphA* genes. Tree was generated in MEGA 5 with the JTT matrix-based model (*71, 72*) using annotated Eph receptor amino acid sequences from Ensembl and UCSC genome browsers. The percentage of replicate trees in which amino acid sequences clustered together in the bootstrap test (500 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. All positions containing gaps and missing data were eliminated. The final dataset includes 290 positions.

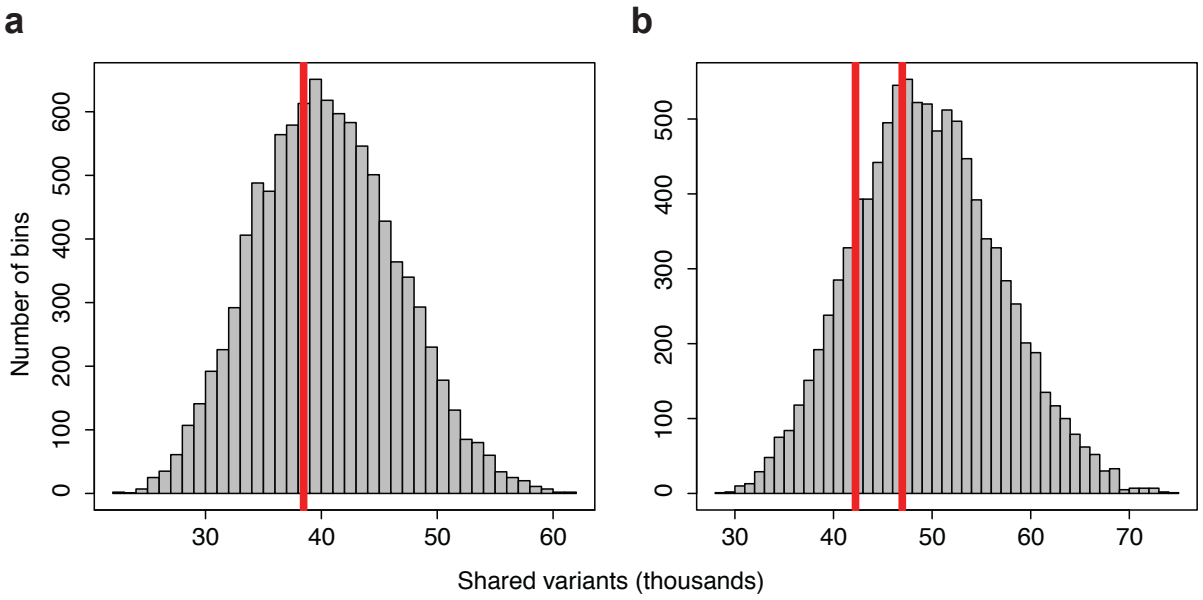

Fig. S25.

Numbers of shared variants in 10,000 bins of 7 and 8 random genomes. **a**, Shared variants in bins of 8 genomes, the number of crested genomes in the resequencing set. The number of variants shared by the 8 crested birds (red line) lay near the peak of the normal distribution. b, Shared variants in bins of 7 genomes. Two Indian fantails are included in the set of 8 resequenced crested birds. Since these two birds are closely related (Figs. 1, S17, S18), we also used bins of 7 instead of 8 to assess the number of variants shared among the 7 crested *breeds*. Red lines indicate positions of the two 7-bird bins that contain one Indian fantail and crested birds from 6 other breeds. The numbers of shared variants in the bin of 8 crested birds and the two bins containing one Indian fantail and the other 6 crested breeds lay within the normal distribution, indicating that the group of crested breeds is not highly structured.

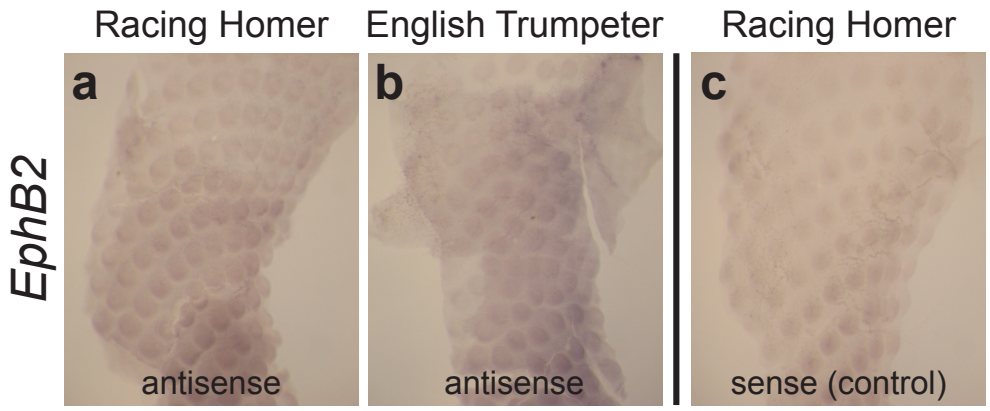

Fig. S26.

Expression of *EphB2* mRNA in the neck and occipital skin of pigeon embryos as detected by whole-mount in situ hybridization. Unlike *EphA4* (see main text Fig. 3), *EphB2* is expressed weakly and is not obviously polarized in the feather placodes of stage 36 embryos of racing homer (**a**, uncrested) or English trumpeter (**b**, crested) pigeon breeds. Signal from *EphB2* antisense probe is only slightly elevated above background, as indicated by sense control (**c**).

Fig. S27.
Male Danish tumbler pigeon used for the reference genome sequence.

Table S1.

Statistics of raw data of pigeon genome sequencing. Coverage calculation was based on the estimated genome size of 1.3Gb.

		Raw data		After filtering and error correction	
Insert Size	Read Length (bp)	Total Data (Gb)	Sequence coverage (X)	Total Data (Gb)	Sequence coverage (X)
200bp	100	29.11	22.39	24.25	18.65
500bp	100	31.94	24.57	23.64	18.18
800bp	100	32.97	25.36	20.51	15.78
2kb	50	14.36	11.05	8.5	6.54
5kb	50	45.08	3.47	2.65	2.04
10kb	50	7.59	5.84	0.99	0.76
20kb	50	6.8	5.23	1.03	0.79
Total		127.27	97.9	81.57	62.75

Table S2.

Statistics of RNA-seq data. Read mapping was done by Tophat (*33*), using parameters "-r 20 --mate-std-dev 10 -m 2 -I 100000".

Sample	#Total reads	#Reads mapped to genome	Mapped rate (%)
Danish Tumbler (heart)	26,128,246	17,850,929	68.32
Danish Tumbler (liver)	46,640,568	34,125,091	73.17
Oriental Frill (heart)	18,609,914	13,135,374	70.58
Oriental Frill (liver)	31,205,996	24,505,498	78.53
Racing Homer (heart)	23,241,351	17,319,740	74.52
Racing Homer (liver)	35,047,903	26,843,926	76.59

Table S3.

Genome size estimation. Data from 3 short-insert libraries (200 bp, 500 bp, 800 bp) were used to estimate the genome size according to the formula, G=kmer_num/kmer_depth.

genome	Kmer length	#kmer	Peak depth	Estimated genome size
pigeon	17	27,351,030,104	21	1,302,430,004

Table S4.

Statistics of the assembled genome. Note that sequences shorter than 100 bp were not included in the statistics.

	Contig		Scaffold	
	Size (bp)	Number	Size (bp)	Number
N90	5,460	51,170	617,714	394
N80	9,609	36,379	1,135,308	263
N70	13,675	26,914	1,624,766	181
N60	17,804	19,932	2,320,313	124
N50	22,406	14,473	3,148,738	82
Longest	250,040		25,666,195	
Total Size	1,090,726,554		1,111,581,692	
Total Number (>100 bp)		143,123		38,878
Total Number (>2 kb)		71,982		2,190

Table S5.

Assembly assessment with EST data. EST data from *Columba livia* were downloaded from the NCBI EST database.

Dataset	Number	Total length (bp)	Covered by assembly	with >90% sequence in one scaffold		with >50% sequence in one scaffold	
				Number	Percent	Number	Percent
All	2,108	614,321	87.86%	1,524	72.30	1,743	82.69
>100bp	2,082	612,127	88.04%	1,511	72.57	1,724	82.81
>200bp	1,755	561,555	89.40%	1,297	73.90	1,472	83.87
>500bp	58	32,604	89.66%	29	50.00	36	62.07

Table S6.

Comparison of 4 avian assemblies.

Genome Feature	Chicken	Zebra finch	Turkey	Pigeon
N50 contig length	36kb	39kb	12.6kb	22kb
N50 scaffold length	7Mb	10Mb	1.5Mb	3.1Mb
Assembled bases	1.06Gb	1.2Gb	931M	1.11Gb
GC (%)	41.5	41.3	40.5	41.5

Table S7.

Repeats predicted in the assembly. The overlaps between repeats were excluded before the calculation of the total.

Type	Repeat Size	% of genome
Proteinmask	42,558,810	3.828671
Repeatmasker	41,235,770	3.709648
Trf	20,769,448	1.868459
Denovo	67,844,898	6.103456
Total	97,039,882	8.729892

Table S8.

General statistics of predicted protein-coding genes.

Gene set		Number	Average transcript length (bp)	Average CDS length (bp)	Average exons per gene	Average exon length (bp)	Average intron length (bp)
De novo	Augustus	24156	18145.26	1156.19	6.67	173.21	2993.70
	Genscan	37395	21756.10	1323.55	8.01	165.31	2916.22
Homolog	G.gallus	10835	21775.19	1527.21	9.65	158.25	2340.63
	H.sapiens	7712	26389.97	1768.23	10.82	163.45	2507.76
	T.guttata	12894	20246.15	1447.93	9.12	158.81	2315.74
Final gene set		17300	18364.87	1404.2	8.47	165.87	2271.79

Table S9.

CEGMA assessment for the pigeon gene set, compared with the chicken gene set.

	Pigeon (#gene)	Chicken (#gene)
Identified CEGMA genes	197	191
Overlap with pigeon gene set more than 80% in CDS level	166	167
Overlap with pigeon gene set more than 50% in CDS level	187	185

Table S10.

Statistics of functional annotation.

		#Gene	Percent (%)
Total		17300	-
Annotated	Swiss-Prot	12841	74.23
	KEGG	8870	51.27
	InterPro	13602	78.62
	GO	11319	65.43
Unannotated		1895	10.95

Table S11.

Non-coding RNA genes in the assembly.

Type		Copy	Average length (bp)	Total length (bp)	% of genome
miRNA		173	84.54	14,626	0.001316
tRNA		188	75.86	14,262	0.001283
rRNA	**rRNA**	119	87.74	10,441	0.000939
	18S	6	95.83	575	0.000052
	28S	18	154.33	2,778	0.00025
	5.8S	1	155.00	155	0.000014
	5S	94	73.76	6,933	0.000624
snRNA	**snRNA**	184	114.57	21,080	0.001896
	CD-box	100	89.45	8,945	0.000805
	HACA-box	54	142.31	7,685	0.000691
	splicing	22	134.09	2,950	0.000265

Table S12.

GO terms enriched in pigeon gene predictions that are not annotated in other birds. (MF, molecular function; BP, biological process.)

GO ID	GO Term	Class	Level	P value
GO:0008907	integrase activity	MF	3	3.770E-02
GO:0018149	peptide cross-linking	BP	6	3.770E-02

Table S13.

IPR terms enriched in pigeon gene predictions that are not annotated in other birds.

IPR ID	IPR Title	P value
IPR008160	Collagen triple helix repeat	6.981E-17
IPR018957	Zinc finger, C3HC4 RING-type	1.480E-08
IPR003596	Immunoglobulin V-set, subgroup	3.413E-08
IPR013106	Immunoglobulin V-set	5.662E-07
IPR007110	Immunoglobulin-like	1.996E-05
IPR001841	Zinc finger, RING-type	3.153E-05
IPR011004	Trimeric LpxA-like	1.719E-04
IPR001037	Integrase, C-terminal, retroviral	1.990E-04
IPR003302	Cornifin (SPRR)	1.990E-04
IPR008936	Rho GTPase activation protein	2.015E-04
IPR016133	Insect antifreeze protein	6.415E-04
IPR000198	Rho GTPase-activating protein domain	7.853E-04
IPR013164	Cadherin, N-terminal	1.207E-03
IPR012337	Ribonuclease H-like	1.689E-02
IPR013787	S100/CaBP-9k-type, calcium binding, subdomain	2.292E-02
IPR013649	Integrin alpha-2	2.751E-02
IPR001101	Plectin repeat	4.475E-02
IPR001584	Integrase, catalytic core	4.475E-02
IPR002717	MOZ/SAS-like protein	4.475E-02

Table S14.

KEGG pathways enriched in pigeon gene predictions that are not annotated in other birds.

Map ID	Map Title	P value
map00230	Purine metabolism	6.942E-49
map03020	RNA polymerase	7.836E-40
map00240	Pyrimidine metabolism	3.702E-26

Table S15.

Gene families under expansion or contraction in pigeon lineage. The functions were assigned based on the best hits to the SwissProt database.

Pigeon	Zebra finch	Turkey	Chicken	Lizard	Expansion or contraction	Putative function
12	4	4	5	10	expansion	Type II keratin
7	3	1	2	1	expansion	Lactosylceramide 4-alpha-galactosyltransferase
4	11	13	14	0	contraction	PHD finger protein 7
14	18	20	24	84	contraction	Protocadherin

Table S16.

Classification of type II keratins in four avian genomes, based on SwissProt annotation.

	Pigeon	Zebra finch	Turkey	Chicken
Type II keratin, cytoskeletal 75	7	3	1	4
Type II keratin, cytoskeletal 6A	1	0	1	0
Type II keratin, cytoskeletal 79	1	0	0	0
Type II keratin, cytoskeletal 5	1	0	1	0
Type II keratin, cytoskeletal 1	1	0	0	0
Type II keratin, cytoskeletal cochleal	1	1	1	1
Total	12	4	4	5

Table S17.

GO terms enriched in putatively lost genes in pigeon lineage.

GO ID	GO Term	Class	Level	P-value
GO:0005126	cytokine receptor binding	MF	5	4.770E-10
GO:0050909	sensory perception of taste	BP	7	5.382E-07
GO:0008534	oxidized purine base lesion DNA N-glycosylase activity	MF	6	7.026E-07
GO:0007631	feeding behavior	BP	4	7.026E-07
GO:0006952	defense response	BP	4	8.359E-07
GO:0007218	neuropeptide signaling pathway	BP	6	8.359E-07
GO:0005801	cis-Golgi network	CC	5	4.453E-06
GO:0006950	response to stress	BP	3	4.507E-06
GO:0005576	extracellular region	CC	2	5.729E-06
GO:0005136	interleukin-4 receptor binding	MF	6	6.984E-06
GO:0004568	chitinase activity	MF	6	1.067E-05
GO:0006032	chitin catabolic process	BP	7	1.067E-05
GO:0003721	telomeric template RNA reverse transcriptase activity	MF	8	8.009E-05
GO:0016798	hydrolase activity, acting on glycosyl bonds	MF	4	8.243E-05
GO:0006284	base-excision repair	BP	7	1.205E-04
GO:0016813	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amidines	MF	5	1.205E-04
GO:0006888	ER to Golgi vesicle-mediated transport	BP	5	2.432E-04
GO:0003684	damaged DNA binding	MF	5	2.877E-04
GO:0008061	chitin binding	MF	5	3.147E-04
GO:0006289	nucleotide-excision repair	BP	7	4.145E-04
GO:0004045	aminoacyl-tRNA hydrolase activity	MF	6	4.698E-04
GO:0045596	negative regulation of cell differentiation	BP	4	9.056E-04
GO:0050896	response to stimulus	BP	2	9.056E-04
GO:0006414	translational elongation	BP	6	6.237E-03
GO:0005044	scavenger receptor activity	MF	6	1.069E-02
GO:0006259	DNA metabolic process	BP	5	2.324E-02

GO:0007186	G-protein coupled receptor protein signaling pathway	BP	5	2.570E-02
GO:0006278	RNA-dependent DNA replication	BP	7	2.667E-02
GO:0016455	RNA polymerase II transcription mediator activity	MF	5	3.289E-02
GO:0016592	mediator complex	CC	4	3.289E-02
GO:0006357	regulation of transcription from RNA polymerase II promoter	BP	7	3.623E-02
GO:0008033	tRNA processing	BP	7	3.683E-02
GO:0008083	growth factor activity	MF	5	3.683E-02
GO:0051258	protein polymerization	BP	7	4.233E-02

Table S18.

IPR domains enriched in putatively lost genes in pigeon lineage.

IPR ID	IPR Title	P-value
IPR000471	Interferon alpha/beta/delta	2.628E-15
IPR009079	Four-helical cytokine-like, core	9.456E-15
IPR022409	PKD/Chitinase domain	1.432E-08
IPR007960	Mammalian taste receptor	1.024E-07
IPR000601	PKD domain	1.024E-07
IPR000874	Bombesin/neuromedin-B/ranatensin peptide family	1.324E-07
IPR001704	Prepro-orexin	1.324E-07
IPR003566	T-cell surface glycoprotein CD5	1.324E-07
IPR007233	Sybindin-like protein	1.324E-07
IPR012904	8-oxoguanine DNA glycosylase, N-terminal	1.324E-07
IPR002354	Interleukin-4	2.193E-06
IPR006035	Ureohydrolase	2.193E-06
IPR001223	Glycoside hydrolase, family 18, catalytic domain	3.283E-06
IPR011583	Chitinase II	3.283E-06
IPR003265	HhH-GPD domain	4.587E-06
IPR011257	DNA glycosylase	4.587E-06
IPR000369	Potassium channel, voltage-dependent, beta subunit, KCNE	9.213E-06
IPR012294	Transcription factor TFIID, C-terminal/DNA glycosylase, N-terminal	1.589E-05
IPR003038	Defender against death DAD protein	2.052E-05
IPR003545	Telomere reverse transcriptase	2.052E-05
IPR019403	Mediator complex, subunit Med19, metazoa	2.052E-05
IPR021891	Telomerase ribonucleoprotein complex - RNA-binding domain	2.052E-05
IPR022773	Siva	2.052E-05
IPR008160	Collagen triple helix repeat	2.828E-05
IPR002347	Glucose/ribitol dehydrogenase	6.677E-05
IPR002198	Short-chain dehydrogenase/reductase SDR	7.486E-05
IPR002557	Chitin binding domain	9.728E-05
IPR002759	Ribonuclease P/MRP protein subunit	9.728E-05

IPR002833	Peptidyl-tRNA hydrolase, PTH2	9.728E-05
IPR001813	Ribosomal protein 60S	1.514E-04
IPR019391	Storkhead-box protein, winged-helix domain	1.514E-04
IPR008717	Noggin	3.068E-04
IPR011012	Longin-like	6.321E-04
IPR001190	Speract/scavenger receptor	1.634E-03
IPR017448	Speract/scavenger receptor-related	1.878E-03
IPR001859	Ribosomal protein P2	4.266E-03
IPR003226	Metal-dependent protein hydrolase	4.266E-03
IPR005651	Uncharacterised protein family UPF0434/Trm112	4.266E-03
IPR019605	Misato Segment II, myosin-like	4.266E-03
IPR000477	Reverse transcriptase	4.871E-03
IPR017853	Glycoside hydrolase, catalytic core	5.169E-03
IPR002035	von Willebrand factor, type A	7.444E-03
IPR003979	Tropoelastin	7.726E-03
IPR003008	Tubulin/FtsZ, GTPase domain	8.973E-03
IPR003129	Laminin G, thrombospondin-type, N-terminal	9.175E-03
IPR001846	von Willebrand factor, type D domain	9.783E-03
IPR001325	Interleukin-4/interleukin-13	1.059E-02
IPR001254	Peptidase S1/S6, chymotrypsin/Hap	1.405E-02
IPR009003	Peptidase cysteine/serine, trypsin-like	1.446E-02
IPR008795	Prominin	2.309E-02

Table S19.

KEGG pathways enriched in putatively lost genes in pigeon lineage.

Map ID	Map Title	P-value
map00510	N-Glycan biosynthesis	7.541E-04
map04742	Taste transduction	7.541E-04
map00072	Synthesis and degradation of ketone bodies	1.223E-03
map00040	Pentose and glucuronate interconversions	2.890E-03
map00520	Amino sugar and nucleotide sugar metabolism	2.890E-03
map04350	TGF-beta signaling pathway	2.890E-03
map00330	Arginine and proline metabolism	2.890E-03
map04622	RIG-I-like receptor signaling pathway	3.912E-03
map03410	Base excision repair	4.557E-03
map00650	Butanoate metabolism	7.232E-03
map04640	Hematopoietic cell lineage	4.813E-02

Table S20.

Putative pseudogenes identified in pigeon. In "Type" column "F" indicates frameshift and "S" indicates premature stop codon; putative functions were assigned by BLASTing the proteins of zebra finch against SwissProt database.

Seq name	Start	End	Type	Homolog in zebra finch	Putative function
scaffold53	45328	46490	F	ENSTGUP00000002855	5-hydroxytryptamine receptor 1A
scaffold240	38891	55369	F	ENSTGUP00000016958	A disintegrin and metalloproteinase with thrombospondin motifs 14
scaffold730	54737	125833	S	ENSTGUP00000005592	ALK tyrosine kinase receptor
scaffold111	3943548	3980555	F	ENSTGUP00000008006	Alpha-1,6-mannosylglycoprotein 6-beta-N-acetylglucosaminyltransferase B
scaffold23	935009	958005	F	ENSTGUP00000010806	Ankyrin repeat domain-containing protein 29
scaffold16	12011433	12088795	F	ENSTGUP00000004825	Anoctamin-3
scaffold94	2132410	2149196	F	ENSTGUP00000018000	Argininosuccinate lyase
scaffold72	1113381	1115963	F	ENSTGUP00000006818	Aryl-hydrocarbon-interacting protein-like 1
scaffold332	263242	318074	F	ENSTGUP00000009007	Aryl hydrocarbon receptor repressor
scaffold26	1242875	1332710	F	ENSTGUP00000003439	Astrotactin-2
scaffold707	450425	571621	S	ENSTGUP00000008137	ATP-binding cassette sub-family A member 13
scaffold272	314086	460677	F	ENSTGUP00000005485	BMP-binding endothelial regulator protein
scaffold232	2435000	2442089	F	ENSTGUP00000009740	Brachyury protein
scaffold391	260763	265428	F	ENSTGUP00000017617	Brain-specific angiogenesis inhibitor 2
scaffold14	178819	180260	F	ENSTGUP00000005572	Brain-specific homeobox/POU domain protein 3
scaffold577	92982	94774	F	ENSTGUP00000003022	BTB/POZ domain-containing protein 17
scaffold67	1930727	2004054	F	ENSTGUP00000001995	Cadherin-9
scaffold60	5283434	5318155	F	ENSTGUP00000011814	Caprin-2
scaffold551	110581	116426	F	ENSTGUP00000008945	Cell division cycle-associated protein 7
scaffold102	5602968	5669075	F	ENSTGUP00000013113	Collagen alpha-1 (IX) chain
scaffold347	986011	1110505	F	ENSTGUP00000010486	Contactin-4
scaffold347	473455	552656	F	ENSTGUP00000010416	Contactin-6
scaffold391	442400	569052	F	ENSTGUP00000001799	CUB and sushi domain-containing protein 1
scaffold1	6781652	6812564	F	ENSTGUP00000012141	Cyclic nucleotide-gated cation channel beta-3
scaffold38	4613745	4617830	F	ENSTGUP00000008416	Cyclic nucleotide-gated channel rod photoreceptor subunit alpha
scaffold194	518579	519941	F	ENSTGUP00000001087	Cytokine receptor-like factor 1
scaffold34	3650819	3659757	S	ENSTGUP00000015661	Cytosolic phospholipase A2 epsilon
scaffold232	4759144	4765687	F	ENSTGUP00000010098	Delta-like protein 1
scaffold73	177527	183529	F	ENSTGUP00000000959	Delta-type opioid receptor
scaffold394	230174	236397	F	ENSTGUP00000012018	Dihydropyrimidinase-related protein 4

scaffold128	2875006	2964895	F	ENSTGUP00000010818	DNA-binding protein SATB2
scaffold196	2506214	2552640	F	ENSTGUP00000006561	Doublecortin domain-containing protein 2
scaffold77	805734	886292	F	ENSTGUP00000000216	Down syndrome cell adhesion molecule-like protein 1
scaffold209	2443295	2497377	F	ENSTGUP00000009472	Dynein heavy chain 3, axonemal
scaffold97	383916	513582	F	ENSTGUP00000007744	Dynein heavy chain 5, axonemal
scaffold133	469779	480210	F	ENSTGUP00000000801	E3 ubiquitin-protein ligase UHRF1
scaffold9	1089235	1138025	F	ENSTGUP00000001397	ELAV-like protein 2
scaffold111	3206137	3224294	F	ENSTGUP00000008382	Envoplakin
scaffold391	1689031	1711474	F	ENSTGUP00000001574	Ephrin type-A receptor 10
scaffold218	2950041	2979523	F	ENSTGUP00000013330	Estrogen receptor beta
scaffold577	31420	33059	F	ENSTGUP00000015065	Fascin-2
scaffold1246	237364	237706	F	ENSTGUP00000004424	Feather keratin 2
scaffold837	15669	15974	S	ENSTGUP00000017121	Feather keratin 2
scaffold156	18388	18678	S	ENSTGUP00000014178	Feather keratin Cos1-1/Cos1-3/Cos2-1
scaffold534	29367	29659	F	ENSTGUP00000018103	Feather keratin Cos1-1/Cos1-3/Cos2-1
scaffold216	6673871	6738878	S	ENSTGUP00000012834	Fer-1-like protein 6
scaffold18	644059	671351	S	ENSTGUP00000008341	FERM and PDZ domain-containing protein 2
scaffold466	7040	35410	S	ENSTGUP00000013723	Frizzled-3
scaffold140	511386	651332	F	ENSTGUP00000011271	Gamma-1-syntrophin
scaffold59	5782872	5784360	F	ENSTGUP00000011655	Gap junction alpha-3 protein
scaffold111	1511805	1520852	F	ENSTGUP00000009145	Glutamate [NMDA] receptor subunit epsilon-3
scaffold3	2687804	2927380	F	ENSTGUP00000003214	Glutamate receptor delta-2 subunit
scaffold277	2033463	2089915	F	ENSTGUP00000008650	Glutamate receptor-interacting protein 2
scaffold101	6133930	6213352	F	ENSTGUP00000013858	Glutamate receptor, ionotropic kainate 1
scaffold506	3370159	3371517	F	ENSTGUP00000014651	Gonadotropin-releasing hormone II receptor
scaffold444	634672	648755	F	ENSTGUP00000010512	GRB2-related adapter protein 2
scaffold16	3247944	3289446	F	ENSTGUP00000009045	Harmonin
scaffold215	945389	987563	F	ENSTGUP00000004131	High affinity cAMP-specific and IBMX-insensitive 3',5'-cyclic phosphodiesterase 8B
scaffold56	4029944	4070302	F	ENSTGUP00000010590	Homeobox protein aristaless-like 4
scaffold264	259693	263080	F	ENSTGUP00000003980	Homeobox protein SIX2
scaffold194	178941	181045	F	ENSTGUP00000014409	Hyaluronan and proteoglycan link protein 4
scaffold644	11400	14073	F	ENSTGUP00000017647	Insulin receptor-related protein
scaffold644	4894	11106	F	ENSTGUP00000017650	Insulin receptor-related protein
scaffold111	859332	880125	F	ENSTGUP00000008671	Integrin beta-4
scaffold415	1021407	1025421	F	ENSTGUP00000008764	Iroquois-class homeodomain protein irx-2
scaffold748	53800	56769	F	ENSTGUP00000002436	Keratin-like protein KRT222
scaffold487	170774	203064	F	ENSTGUP00000005027	Kinesin-like protein KIF15
scaffold176	1599190	1620885	F	ENSTGUP00000001231	Kinesin-like protein KIF21A
scaffold232	3417955	3456388	F	ENSTGUP00000009939	Kinesin-like protein KIF25

scaffold179	1133008	1146449	F	ENSTGUP00000004052	Leishmanolysin-like peptidase
scaffold9	2521066	2522894	F	ENSTGUP00000001446	Leucine-rich repeat and immunoglobulin-like domain-containing nogo receptor-interacting protein 2
scaffold7	16570045	16600480	F	ENSTGUP00000010668	Leucine-rich repeat-containing protein 7
scaffold7	16605281	16645225	F	ENSTGUP00000010664	Leucine-rich repeat-containing protein 7
scaffold250	893558	902750	F	ENSTGUP00000014037	Leukemia NUP98 fusion partner 1
scaffold7	17700659	17741714	F	ENSTGUP00000017593	LIM homeobox transcription factor 1-alpha
scaffold176	592634	625095	F	ENSTGUP00000001421	Liprin-alpha-2
scaffold1150	11463	17570	F	ENSTGUP00000014265	Myosin heavy chain, skeletal muscle, adult
scaffold20	2129046	2138699	S	ENSTGUP00000007528	Myosin-Ih
scaffold814	398584	507636	F	ENSTGUP00000001192	Myosin IIIA
scaffold265	2414191	2415095	F	ENSTGUP00000017047	Neurogenic differentiation factor 2
scaffold18	279786	280902	F	ENSTGUP00000008157	Neuropeptide Y receptor type 4
scaffold184	426954	456615	F	ENSTGUP00000013544	Ninein
scaffold94	324053	334628	F	ENSTGUP00000004305	P2X purinoceptor 1
scaffold454	1063266	1068135	F	ENSTGUP00000000933	Patatin-like phospholipase domain-containing protein 1
scaffold642	2596155	2596912	F	ENSTGUP00000016048	PHD finger protein 19
scaffold531	36026	145255	F	ENSTGUP00000007971	Phosphotidylinositol phosphatase PTPRQ
scaffold31	5724976	5739739	S	ENSTGUP00000010970	PI-PLC X domain-containing protein 1
scaffold642	2499461	2533419	S	ENSTGUP00000006676	Potassium channel subfamily T member 1
scaffold81	362009	363086	F	ENSTGUP00000004824	Potassium voltage-gated channel subfamily D member 2
scaffold538	35268	45241	F	ENSTGUP00000003541	Potassium voltage-gated channel subfamily G member 3
scaffold102	24350817	24352251	F	ENSTGUP00000013517	Potassium voltage-gated channel subfamily S member 3
scaffold38	15821423	15874948	F	ENSTGUP00000010064	Prominin-1-A
scaffold52	160598	161899	S	ENSTGUP00000017844	Pro-Pol polyprotein
scaffold625	1321252	1326040	F	ENSTGUP00000005239	Protein bassoon
scaffold146	1061437	1078178	F	ENSTGUP00000007260	Protein FAM83G
scaffold102	21210037	21268902	F	ENSTGUP00000013472	Protein GREB1
scaffold366	2440469	2633022	F	ENSTGUP00000002597	Protein piccolo
scaffold68	7887843	7969779	F	ENSTGUP00000011710	Protein unc-80 homolog
scaffold96	6109	22383	F	ENSTGUP00000005189	Protein Wnt-2
scaffold79	14434675	14436751	F	ENSTGUP00000001749	Protocadherin-10
scaffold17	11199015	11202590	F	ENSTGUP00000012815	Protocadherin-8
scaffold599	2381086	2474045	S	ENSTGUP00000013246	Proton-coupled amino acid transporter 4
scaffold383	1893302	1947343	F	ENSTGUP00000012116	Protor-1
scaffold1398	97670	103937	F	ENSTGUP00000003796	Rac GTPase-activating protein 1
scaffold775	92513	362730	F	ENSTGUP00000012537	Regulating synaptic membrane exocytosis protein 2

scaffold83	13471	15660	F	ENSTGUP00000000137	Retinal homeobox protein Rx1
scaffold7	2268125	2339194	F	ENSTGUP00000005983	Retinal-specific ATP-binding cassette transporter
scaffold34	7763640	7768799	F	ENSTGUP00000011874	Retinol dehydrogenase 12
scaffold362	1106455	1109934	F	ENSTGUP00000010860	Rhodopsin
scaffold102	20266489	20271297	F	ENSTGUP00000013430	Ribonucleoside-diphosphate reductase subunit M2
scaffold1374	61859	62940	F	ENSTGUP00000003799	RNA-binding protein MEX3A
scaffold101	15699493	15699781	F	ENSTGUP00000013940	Roundabout homolog 2
scaffold589	835372	990644	F	ENSTGUP00000013641	Runt-related transcription factor 2
scaffold34	9257529	9456426	F	ENSTGUP00000012081	Ryanodine receptor 3
scaffold94	1318739	1325164	F	ENSTGUP00000004678	Scavenger receptor cysteine-rich domain-containing group B protein
scaffold421	604452	609051	S	ENSTGUP00000002141	Serpin B4
scaffold79	6283354	6350042	F	ENSTGUP00000001250	Short transient receptor potential channel 7
scaffold16	6042061	6080197	F	ENSTGUP00000008559	Signal peptide, CUB and EGF-like domain-containing protein 2
scaffold219	707593	709512	F	ENSTGUP00000011587	SLIT and NTRK-like protein 5
scaffold627	336497	339003	F	ENSTGUP00000012970	SLIT and NTRK-like protein 6
scaffold32	4393984	4456791	F	ENSTGUP00000007683	Sodium channel protein type 1 subunit alpha
scaffold265	514646	548036	F	ENSTGUP00000003375	Sodium channel protein type 2 subunit alpha
scaffold32	4048043	4108186	F	ENSTGUP00000007354	Sodium channel protein type 2 subunit alpha
scaffold32	4524109	4559382	F	ENSTGUP00000007765	Sodium channel protein type 2 subunit alpha
scaffold60	2341383	2417808	F	ENSTGUP00000011594	Stabilin-2
scaffold20	797776	809621	F	ENSTGUP00000008222	Sushi domain-containing protein 2
scaffold31	13538580	13639250	F	ENSTGUP00000013766	Syntaxin-binding protein 5-like
scaffold102	16266990	16310938	F	ENSTGUP00000013346	Thyroid peroxidase
scaffold427	139698	316535	F	ENSTGUP00000007563	Thyrotropin-releasing hormone-degrading ectoenzyme
scaffold64	977357	1035340	F	ENSTGUP00000007506	Tolloid-like protein 2
scaffold363	1311767	1312758	F	ENSTGUP00000011952	Trace amine-associated receptor 1
scaffold725	213034	249428	F	ENSTGUP00000003679	Transient receptor potential cation channel subfamily M member 8
scaffold148	207822	258675	F	ENSTGUP00000000839	Transmembrane channel-like protein 2
scaffold974	46425	64294	F	ENSTGUP00000011000	Transmembrane protease, serine 6
scaffold77	1495726	1508435	F	ENSTGUP00000000412	Tripartite motif-containing protein 29
scaffold873	138748	155693	F	ENSTGUP00000003891	Tripartite motif-containing protein 71
scaffold34	19784160	19790557	F	ENSTGUP00000012663	tRNA wybutosine-synthesizing protein 2/3/4
scaffold599	604986	655114	F	ENSTGUP00000013283	Tyrosinase
scaffold75	25742	37680	F	ENSTGUP00000015371	Ubiquitin-associated and SH3 domain-containing protein A
scaffold220	1606156	1607116	F	ENSTGUP00000013710	Uncharacterized protein C12orf53 homolog

scaffold873	274315	294467	F	ENSTGUP00000003918	Uncharacterized protein C3orf77
scaffold277	1671707	1709132	F	ENSTGUP00000008500	Urocanate hydratase
scaffold440	230917	617963	F	ENSTGUP00000002977	Usherin
scaffold176	1637994	1666541	S	ENSTGUP00000001207	Voltage-dependent L-type calcium channel subunit alpha-1S
scaffold1281	81092	254843	F	ENSTGUP00000003146	Voltage-dependent N-type calcium channel subunit alpha-1B
scaffold87	6780837	6852339	F	ENSTGUP00000017636	Voltage-dependent R-type calcium channel subunit alpha-1E
scaffold444	518589	523070	F	ENSTGUP00000010514	Voltage-dependent T-type calcium channel subunit alpha-1I
scaffold102	20709936	20729400	S	ENSTGUP00000013443	V-type proton ATPase subunit C 2
scaffold1	6941443	6960735	F	ENSTGUP00000012100	V-type proton ATPase subunit d 2
scaffold832	294993	303476	F	ENSTGUP00000011968	Wee1-like protein kinase
scaffold647	708096	712973	F	ENSTGUP00000005942	Wiskott-Aldrich syndrome protein family member 3
scaffold111	5096295	5096381	S	ENSTGUP00000007788	Unknown function
scaffold1320	64495	66032	F	ENSTGUP00000009143	Unknown function
scaffold166	483495	483598	F	ENSTGUP00000005831	Unknown function
scaffold244	2086729	2086854	S	ENSTGUP00000010883	Unknown function

Table S21.

GO enrichment of putative pseudogenes in pigeon.

GO ID	GO Term	Class	Level	P-value
GO:0008324	cation transmembrane transporter activity	MF	6	1.235E-04
GO:0005886	plasma membrane	CC	4	1.235E-04
GO:0032991	macromolecular complex	CC	2	1.537E-04
GO:0005262	calcium channel activity	MF	8	1.537E-04
GO:0055085	transmembrane transport	BP	3	2.583E-04
GO:0003774	motor activity	MF	8	2.842E-04
GO:0015672	monovalent inorganic cation transport	BP	6	2.903E-04
GO:0006816	calcium ion transport	BP	8	3.426E-04
GO:0034220	ion transmembrane transport	BP	4	3.624E-04
GO:0016020	membrane	CC	3	3.812E-04
GO:0005272	sodium channel activity	MF	8	4.848E-04
GO:0005887	integral to plasma membrane	CC	6	4.978E-04
GO:0005891	voltage-gated calcium channel complex	CC	5	9.931E-04
GO:0003777	microtubule motor activity	MF	9	3.749E-03
GO:0005245	voltage-gated calcium channel activity	MF	9	4.147E-03
GO:0051925	regulation of calcium ion transport via voltage-gated calcium channel activity	BP	7	4.147E-03
GO:0007155	cell adhesion	BP	3	4.984E-03
GO:0005856	cytoskeleton	CC	5	5.122E-03
GO:0007018	microtubule-based movement	BP	4	6.323E-03
GO:0044430	cytoskeletal part	CC	4	7.068E-03
GO:0006810	transport	BP	3	7.068E-03
GO:0004970	ionotropic glutamate receptor activity	MF	7	7.068E-03
GO:0005234	extracellular-glutamate-gated ion channel activity	MF	10	7.068E-03
GO:0006814	sodium ion transport	BP	7	8.218E-03
GO:0030247	polysaccharide binding	MF	4	1.144E-02
GO:0016021	integral to membrane	CC	5	1.867E-02
GO:0044425	membrane part	CC	3	2.281E-02
GO:0030288	outer membrane-bounded periplasmic space	CC	4	3.309E-02
GO:0005540	hyaluronic acid binding	MF	6	3.845E-02
GO:0007275	multicellular organismal development	BP	3	4.349E-02
GO:0070588	calcium ion transmembrane transport	BP	5	4.349E-02
GO:0030286	dynein complex	CC	5	4.349E-02
GO:0005515	protein binding	MF	3	4.473E-02
GO:0004871	signal transducer activity	MF	3	4.473E-02
GO:0015276	ligand-gated ion channel activity	MF	7	4.927E-02

Table S22 **IPR enrichment of putative pseudogenes in pigeon.**

IPR ID	IPR Title	P-value
IPR010526	Sodium ion transport-associated	1.346E-04
IPR003961	Fibronectin, type III	1.179E-03
IPR014873	Voltage-dependent calcium channel, alpha-1 subunit, IQ domain	1.179E-03
IPR001696	Voltage gated sodium channel, alpha subunit	1.639E-03
IPR008957	Fibronectin type III domain	2.017E-03
IPR002077	Voltage-dependent calcium channel, alpha-1 subunit	2.771E-03
IPR001320	Ionotropic glutamate receptor	1.857E-02
IPR019594	Glutamate receptor, L-glutamate/glycine-binding	1.967E-02
IPR003968	Potassium channel, voltage dependent, Kv	2.827E-02
IPR000859	CUB	2.844E-02
IPR003971	Potassium channel, voltage dependent, Kv9	4.402E-02

Table S23.

KEGG pathway enrichment of putative pseudogenes in pigeon.

Map ID	Map Title	P-value
04080	Neuroactive ligand-receptor interaction	7.488E-03
04020	Calcium signaling pathway	2.963E-02

Table S24.

GO terms enriched in Neoaves-specific genes. (CC, cellular component; MF, molecular function; BP, biological process.)

GO ID	GO Term	Class	Level	P value
GO:0007040	lysosome organization	BP	7	1.677E-06
GO:0004348	glucosylceramidase activity	MF	6	1.485E-05
GO:0047961	glycine N-acyltransferase activity	MF	8	6.177E-05
GO:0051015	actin filament binding	MF	6	6.177E-05
GO:0005764	lysosome	CC	7	1.118E-04
GO:0016810	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	MF	4	2.252E-04
GO:0006665	sphingolipid metabolic process	BP	6	8.289E-04
GO:0016811	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides	MF	5	1.385E-03
GO:0004523	ribonuclease H activity	MF	9	1.640E-03
GO:0008108	UDP-glucose:hexose-1-phosphate uridylyltransferase activity	MF	7	1.725E-03
GO:0000247	C-8 sterol isomerase activity	MF	6	1.725E-03
GO:0006696	ergosterol biosynthetic process	BP	6	1.725E-03
GO:0044444	cytoplasmic part	CC	4	2.400E-03
GO:0003725	double-stranded RNA binding	MF	5	7.147E-03
GO:0003676	nucleic acid binding	MF	3	7.736E-03
GO:0004109	coproporphyrinogen oxidase activity	MF	6	7.741E-03
GO:0016814	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in cyclic amidines	MF	5	7.741E-03
GO:0030674	protein binding, bridging	MF	4	7.741E-03
GO:0006955	immune response	BP	3	1.068E-02
GO:0004045	aminoacyl-tRNA hydrolase activity	MF	6	1.068E-02
GO:0005737	cytoplasm	CC	4	1.352E-02
GO:0005739	mitochondrion	CC	5	1.352E-02
GO:0017176	phosphatidylinositol N-acetylglucosaminyltransferase activity	MF	7	1.356E-02
GO:0006012	galactose metabolic process	BP	7	1.356E-02
GO:0009968	negative regulation of signal transduction	BP	4	1.356E-02
GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds	MF	5	1.514E-02
GO:0005125	cytokine activity	MF	5	2.810E-02
GO:0010467	gene expression	BP	4	2.993E-02
GO:0006779	porphyrin biosynthetic process	BP	6	3.215E-02
GO:0019068	virion assembly	BP	4	3.569E-02
GO:0003677	DNA binding	MF	4	3.696E-02

Table S25.

IPR terms enriched in Neoaves-specific genes.

IPR ID	IPR Title	P value
IPR003308	Integrase, N-terminal zinc-binding domain	2.123E-08
IPR001139	Glycoside hydrolase, family 30	1.475E-06
IPR003350	Homeodomain protein CUT	9.634E-06
IPR001584	Integrase, catalytic core	3.742E-05
IPR012858	Dendritic cell-specific transmembrane protein-like	3.742E-05
IPR006846	Ribosomal protein S30	3.742E-05
IPR009829	Protein of unknown function DUF1395	3.742E-05
IPR015938	Glycine N-acyltransferase, N-terminal	3.742E-05
IPR022768	Fascin domain	3.742E-05
IPR016187	C-type lectin fold	2.708E-04
IPR008999	Actin cross-linking	3.064E-04
IPR010982	Lambda repressor-like, DNA-binding	5.205E-04
IPR008063	Fas receptor	5.205E-04
IPR001304	C-type lectin	5.618E-04
IPR013158	APOBEC-like, N-terminal	7.011E-04
IPR002156	Ribonuclease H domain	7.011E-04
IPR012337	Ribonuclease H-like	7.011E-04
IPR010625	CHCH	7.011E-04
IPR000118	Granulin	7.011E-04
IPR000940	Methyltransferase, NNMT/PNMT/TEMT	7.011E-04
IPR001328	Peptidyl-tRNA hydrolase	7.011E-04
IPR002036	Uncharacterised protein family UPF0054, metalloprotease YbeY, predicted	7.011E-04
IPR005341	Protein Transporter, Pam16	7.011E-04
IPR005849	Galactose-1-phosphate uridyl transferase, N-terminal	7.011E-04
IPR005850	Galactose-1-phosphate uridyl transferase, C-terminal	7.011E-04
IPR006716	ERG2/sigma1 receptor-like	7.011E-04
IPR006849	IKI3	7.011E-04
IPR007128	Nnf1	7.011E-04
IPR007635	Tis11B-like protein, N-terminal	7.011E-04
IPR008657	Jumping translocation breakpoint	7.011E-04
IPR008806	RNA polymerase III Rpc82, C -terminal	7.011E-04
IPR009125	DAPIT	7.011E-04
IPR009450	Phosphatidylinositol N-acetylglucosaminyltransferase	7.011E-04
IPR009787	Protein of unknown function DUF1352	7.011E-04
IPR010342	Protein of unknown function DUF938	7.011E-04
IPR010681	Plethodontid receptivity factor PRF	7.011E-04
IPR010723	HemN, C-terminal	7.011E-04
IPR012574	Mitochondrial proteolipid	7.011E-04
IPR012918	RTP801, C-terminal	7.011E-04
IPR013197	RNA polymerase III subunit RPC82-related, helix-turn-helix	7.011E-04
IPR013549	Domain of unknown function DUF1731, C-terminal	7.011E-04

IPR013652	Glycine N-acyltransferase, C-terminal	7.011E-04
IPR018881	Uncharacterised protein family UPF0565	7.011E-04
IPR019095	Mediator complex, subunit Med18, metazoa/fungi	7.011E-04
IPR008160	Collagen triple helix repeat	1.604E-03
IPR007248	Mpv17/PMP22	1.916E-03
IPR009865	Proacrosin binding sp32	1.916E-03
IPR019522	Phosphoinositide 3-kinase 1B, gamma adapter, p101 subunit	1.916E-03
IPR017853	Glycoside hydrolase, catalytic core	3.101E-03
IPR001356	Homeobox	3.158E-03
IPR000235	Ribosomal protein S7	3.158E-03
IPR002772	Glycoside hydrolase, family 3, C-terminal	3.158E-03
IPR003573	Interleukin-6/G-CSF/MGF	3.158E-03
IPR007741	Ribosomal protein/NADH dehydrogenase domain	3.158E-03
IPR009626	Uncharacterised protein family UPF0258	3.158E-03
IPR009764	Ovarian carcinoma immunoreactive antigen	3.158E-03
IPR013093	ATPase, AAA-2	3.158E-03
IPR019489	Clp ATPase, C-terminal	3.158E-03
IPR001159	Double-stranded RNA-binding	3.213E-03
IPR006630	RNA-binding protein Lupus La	3.684E-03
IPR008962	PapD-like	4.312E-03
IPR008253	Marvel	4.574E-03
IPR014730	Electron transfer flavoprotein, alpha/beta-subunit, N-terminal	4.814E-03
IPR009079	Four-helical cytokine-like, core	5.821E-03
IPR005437	Gamma-aminobutyric-acid A receptor, gamma subunit	6.742E-03
IPR007904	APOBEC-like, C-terminal	6.742E-03
IPR021673	C-terminal domain of RIG-I	6.742E-03
IPR010844	Occludin/RNA polymerase II elongation factor, ELL domain	9.229E-03
IPR004877	Cytochrome b561, eukaryote	1.187E-02
IPR009851	Modifier of rudimentary, Modr	1.187E-02
IPR009057	Homeodomain-like	1.565E-02
IPR021128	MARVEL-like domain	1.799E-02
IPR006593	Cytochrome b561/ferric reductase transmembrane	1.801E-02
IPR001368	TNFR/CD27/30/40/95 cysteine-rich region	1.916E-02
IPR003036	Core shell protein Gag P30	2.099E-02
IPR003165	Stem cell self-renewal protein Piwi	2.099E-02
IPR000120	Amidase	2.467E-02
IPR001270	Chaperonin ClpA/B	2.733E-02
IPR006638	Elongator protein 3/MiaB/NifB	2.733E-02
IPR007593	Interferon-induced transmembrane protein	2.733E-02
IPR010926	Myosin tail 2	2.733E-02
IPR011146	Histidine triad-like motif	2.733E-02
IPR000181	Formylmethionine deformylase	2.733E-02
IPR000892	Ribosomal protein S26e	2.733E-02
IPR006996	Dynamitin subunit 2	2.733E-02
IPR009445	Protein of unknown function DUF1077, TMEM85	2.733E-02
IPR015216	SANT associated	2.733E-02

IPR015362	Exon junction complex, Pym	2.733E-02
IPR018902	Uncharacterised protein family UPF0573/UPF0605	2.733E-02
IPR019351	Protein of unknown function DUF2039	2.733E-02
IPR020546	ATPase, F1 complex, delta/epsilon subunit, N-terminal	2.733E-02
IPR020547	ATPase, F1 complex, delta/epsilon subunit, C-terminal	2.733E-02
IPR021148	Protein of unknown function DUF579	2.733E-02
IPR022702	DNA (cytosine-5)-methyltransferase 1, replication foci domain	2.733E-02
IPR022730	DAZ associated protein 2	2.733E-02
IPR003115	ParB-like nuclease	2.733E-02
IPR007197	Radical SAM	3.051E-02
IPR001279	Beta-lactamase-like	4.213E-02
IPR004156	Organic anion transporter polypeptide OATP	4.626E-02
IPR002035	von Willebrand factor, type A	4.927E-02
IPR015373	Interferon alpha/beta receptor, beta chain	4.972E-02
IPR010515	Collagenase NC10/endostatin	4.972E-02
IPR020977	Beta-casein-like	4.972E-02

Table S26.

KEGG pathways enriched in Neoaves-specific genes.

Map ID	Map Title	P value
map04512	ECM-receptor interaction	1.758E-04
map00791	Atrazine degradation	1.758E-04
map00511	Other glycan degradation	6.854E-03
map03020	RNA polymerase	8.276E-03
map04623	Cytosolic DNA-sensing pathway	2.361E-02
map00230	Purine metabolism	2.361E-02
map04510	Focal adhesion	3.218E-02
map03010	Ribosome	4.144E-02

Table S27.

MHC B locus of chicken aligned to pigeon assembly.

Chicken MHC locus B Region				Pigeon				
Seq name	Start	End	Seq len	Seq name	Start	End	+/-	Seq len
AB268588	25	234	241833	scaffold2135	1554	1769	-	2932
AB268588	18881	19294	241833	scaffold335	1124193	1124649	-	3333588
AB268588	19556	19656	241833	scaffold218	1097607	1097707	-	4344174
AB268588	20530	22050	241833	scaffold1271	116686	117370	+	180513
AB268588	23743	23908	241833	scaffold3891	211	376	+	733
AB268588	26023	35031	241833	scaffold1043	4777	26100	+	36866
AB268588	37149	37635	241833	scaffold1679	5742	6158	+	31163
AB268588	41923	43328	241833	scaffold306	147143	148078	+	5637556
AB268588	44359	64482	241833	scaffold1061	6786	31767	+	32087
AB268588	44670	45123	241833	scaffold445	1825303	1825665	-	2813049
AB268588	59429	59614	241833	scaffold38	14155392	14155569	-	19163802
AB268588	59615	59732	241833	scaffold121	1602344	1602473	-	2822310
AB268588	63890	64158	241833	scaffold172	742893	743166	-	7630735
AB268588	66399	66746	241833	scaffold2773	1466	1808	+	2734
AB268588	68540	68781	241833	scaffold2451	11408	11656	-	18113
AB268588	76322	76656	241833	scaffold1679	14793	15062	-	31163
AB268588	76883	88040	241833	scaffold2451	65	17464	+	18113
AB268588	80868	81025	241833	scaffold534	100691	100801	-	193724
AB268588	85197	85335	241833	scaffold1679	5743	5881	-	31163
AB268588	88041	89568	241833	scaffold679	157761	161245	+	187166
AB268588	97204	97383	241833	scaffold111	496683	496880	-	5910799
AB268588	98984	99173	241833	scaffold314	188883	189100	+	3454423
AB268588	99174	99379	241833	scaffold17	2851508	2851713	-	18212151
AB268588	99409	99570	241833	scaffold2459	79365	79530	+	121524
AB268588	99571	99906	241833	scaffold1938	1804	2338	-	9131
AB268588	105954	108327	241833	scaffold1060	4681	10159	+	47963
AB268588	114066	114250	241833	scaffold1060	15816	16078	+	47963
AB268588	116771	117046	241833	scaffold486	988487	988756	+	3153651
AB268588	121272	121706	241833	scaffold1060	25068	25487	+	47963
AB268588	144258	144442	241833	scaffold4426	615	842	-	1359
AB268588	144443	145525	241833	scaffold4680	60	1328	-	2495
AB268588	145590	145865	241833	C16435426	66	368	-	373
AB268588	148409	148530	241833	scaffold2825	9296	9423	-	9425
AB268588	149326	149623	241833	scaffold3458	2	262	-	266
AB268588	151364	152466	241833	scaffold4680	60	1348	+	2495
AB268588	154310	157761	241833	scaffold417	12400	29538	+	208115
AB268588	167553	167704	241833	scaffold363	1942872	1943023	+	4238849
AB268588	170482	170880	241833	scaffold2691	89	474	-	866

AB268588	172589	173260	241833	C16605824	2	661	+	671
AB268588	175266	176029	241833	scaffold1621	101	911	-	2723
AB268588	186436	187106	241833	C16605824	2	661	-	671
AB268588	192750	193131	241833	scaffold136	174729	175090	-	190033
AB268588	194335	195513	241833	scaffold1339	13494	13942	+	28095
AB268588	202834	202978	241833	scaffold4793	2281	2425	+	2856
AB268588	203734	205205	241833	scaffold171	2706568	2708114	+	2848140
AB268588	207727	208063	241833	scaffold4677	644	975	-	980
AB268588	208771	209033	241833	scaffold1872	2830	3066	-	11578
AB268588	210063	210205	241833	scaffold145	159858	160000	+	195293
AB268588	212023	212551	241833	scaffold3167	470	1019	-	1041
AB268588	222320	227432	241833	scaffold853	2441	10636	-	12383
AB268588	227913	228013	241833	scaffold161	4097273	4097373	+	11859676
AB268588	235574	241429	241833	scaffold335	3315356	3330267	+	3333588

Table S28.

Three-epoch demographic model for *C. livia.*

	Epoch 1 population size	Epoch 2 generations	Epoch 2 population size	Epoch 3 generations	Epoch 3 population size	TMRCA (years)
Point estimate	95,079	1,496,541	760,597	1	90	1,650,636
Lower 95% CI	16,499	986,564	742,949	1	1	1,611,061
Higher 95% CI	276,107	1,692,590	782,576	90	6,839	1,730,866

Supplementary References

26. R. Li *et al.*, The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311 (Jan 21, 2010).
27. R. Li *et al.*, SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966 (2009).
28. G. Benson, Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573 (Jan 15, 1999).
29. R. A. Dalloul *et al.*, Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. *PLoS Biol* **8**, (2010).
30. L. Hillier, W. Miller, E. Birney, W. Warren, R. Hardison, Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695 (Dec 9, 2004).
31. W. C. Warren *et al.*, The genome of a songbird. *Nature* **464**, 757 (Apr 1, 2010).
32. E. Birney, M. Clamp, R. Durbin, GeneWise and Genomewise. *Genome Res* **14**, 988 (May, 2004).
33. C. Trapnell, L. Pachter, S. L. Salzberg, TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105 (May 1, 2009).
34. C. Trapnell *et al.*, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**, 511 (May, 2010).
35. M. Stanke, S. Waack, Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**, ii215 (Oct, 2003).
36. C. Burge, S. Karlin, Prediction of complete gene structures in human genomic DNA. *Journal of molecular biology* **268**, 78 (Apr 25, 1997).
37. C. G. Elsik *et al.*, Creating a honey bee consensus gene set. *Genome Biol* **8**, R13 (2007).
38. R. Apweiler *et al.*, UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32**, D115 (Jan 1, 2004).
39. R. Apweiler *et al.*, The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res* **29**, 37 (Jan 1, 2001).
40. M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27 (Jan 1, 2000).
41. T. M. Lowe, S. R. Eddy, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955 (Mar 1, 1997).
42. E. P. Nawrocki, D. L. Kolbe, S. R. Eddy, Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335 (May 15, 2009).
43. S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, S. R. Eddy, Rfam: an RNA family database. *Nucleic Acids Res* **31**, 439 (Jan 1, 2003).
44. H. Li *et al.*, TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res* **34**, D572 (Jan 1, 2006).
45. D. W. Huang, B. T. Sherman, R. A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1 (Jan, 2009).

46. T. De Bie, N. Cristianini, J. P. Demuth, M. W. Hahn, CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269 (May 15, 2006).
47. S. Guindon, F. Delsuc, J. F. Dufayard, O. Gascuel, Estimating maximum likelihood phylogenies with PhyML. *Methods in molecular biology* **537**, 113 (2009).
48. A. Morgulis *et al.*, Database indexing for production MegaBLAST searches. *Bioinformatics* **24**, 1757 (Aug 15, 2008).
49. R Development Core Team. (R Foundation for Statistical Computing, Vienna, Austria, 2008).
50. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289 (Jan 22, 2004).
51. S. Purcell *et al.*, PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559 (Sep, 2007).
52. N. A. Rosenberg, DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes* **4**, 137 (2004).
53. A. R. Rogers, C. Huff, Linkage disequilibrium between loci with unknown phase. *Genetics* **182**, 839 (Jul, 2009).
54. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *Journal of molecular biology* **215**, 403 (Oct 5, 1990).
55. D. Posada, K. A. Crandall, MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817 (1998).
56. M. A. Pacheco *et al.*, Evolution of modern birds revealed by mitogenomics: timing the radiation and origin of major orders. *Mol Biol Evol* **28**, 1927 (Jun, 2011).
57. Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586 (Aug, 2007).
58. K. Nam *et al.*, Molecular evolution of genes in avian genomes. *Genome Biol* **11**, R68 (2010).
59. R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, C. D. Bustamante, Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* **5**, e1000695 (Oct, 2009).
60. R. R. Hudson, Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337 (Feb, 2002).
61. B. S. Weir, C. C. Cockerham, Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**, 1358 (1984).
62. 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 (Nov, 2007).
63. M. Clement, D. Posada, K. A. Crandall, TCS: a computer program to estimate gene genealogies. *Mol Ecol* **9**, 1657 (Oct, 2000).
64. J. D. Thompson, D. G. Higgins, T. J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673 (Nov 11, 1994).
65. L. L. Abler *et al.*, A high throughput in situ hybridization method to characterize mRNA expression patterns in the fetal mouse lower urogenital tract. *JoVE* (2011).
66. A. Fazl, The art of training pigeons in the East [annotated translation from Ain-i-Akbari, 1590]. *The Zoologist (London)* **12**, 167 (1888).

67. W. B. Tegetmeier, *Pigeons: Their Structure, Varieties, Habits, and Management*. (George Routledge and Sons, London, 1868), pp. 178.
68. National Pigeon Association, *2010 National Pigeon Association Book of Standards*. (Purebred Pigeon Publishing, Goodlettsville, TN, 2010).
69. K. Pelak *et al.*, The characterization of twenty sequenced human genomes. *PLoS Genet* **6**, (Sep, 2010).
70. W. Kent, BLAT - the BLAST-like alignment tool. *Genome Res* **12**, 656 (2002).
71. D. T. Jones, W. R. Taylor, J. M. Thornton, The rapid generation of mutation data matrices from protein sequences. *CABIOS* **8**, 275 (Jun, 1992).
72. K. Tamura *et al.*, MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**, 2731 (Oct, 2011).