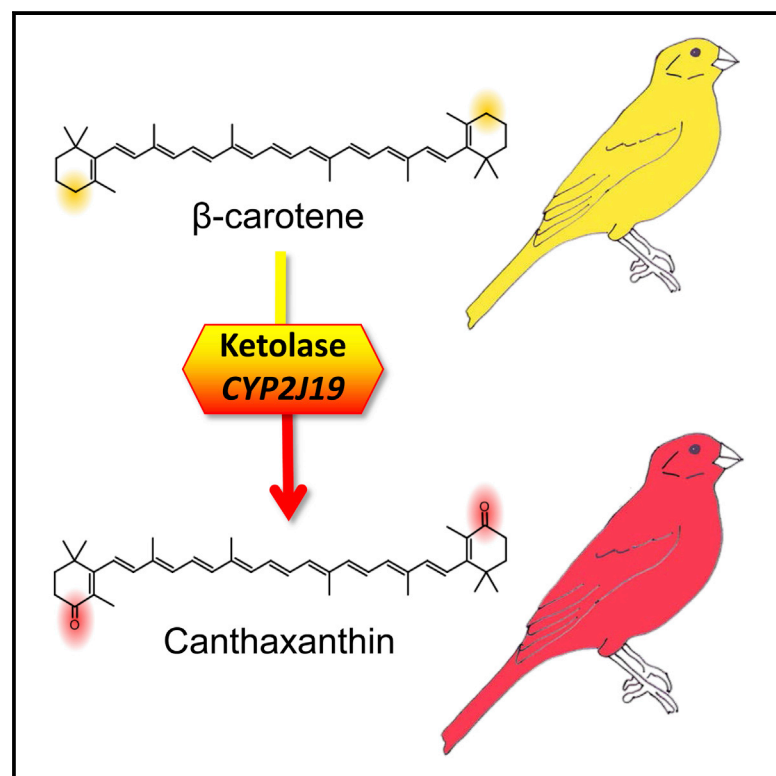


# Current Biology

## Genetic Basis for Red Coloration in Birds

### Graphical Abstract



### Authors

Ricardo J. Lopes, James D. Johnson, Matthew B. Toomey, ..., Geoffrey E. Hill, Joseph C. Corbo, Miguel Carneiro

### Correspondence

hillgee@auburn.edu (G.E.H.),  
jcorbo@pathology.wustl.edu (J.C.C.),  
miguel.carneiro@cibio.up.pt (M.C.)

### In Brief

To produce red coloration of bills and feathers, birds convert yellow dietary carotenoids to red ketocarotenoids via the action of an unknown enzyme. Lopes et al. use whole-genome sequencing of yellow and red canaries to implicate *CYP2J19* as the ketolase that catalyzes this conversion.

### Highlights

- Two genomic regions are required for red ketocarotenoid-based coloration in canaries
- The first region contains *CYP2J19*, a cytochrome P450 family member
- *CYP2J19* is implicated as the ketolase required for ketocarotenoid formation in birds
- The second region maps to a cluster of genes involved in epidermal differentiation

# Genetic Basis for Red Coloration in Birds

Ricardo J. Lopes,<sup>1,9</sup> James D. Johnson,<sup>2,9</sup> Matthew B. Toomey,<sup>3,9</sup> Mafalda S. Ferreira,<sup>1</sup> Pedro M. Araujo,<sup>1,4</sup> José Melo-Ferreira,<sup>1,5</sup> Leif Andersson,<sup>6,7,8</sup> Geoffrey E. Hill,<sup>2,\*</sup> Joseph C. Corbo,<sup>3,\*</sup> and Miguel Carneiro<sup>1,5,\*</sup>

<sup>1</sup>CIBIO/InBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Campus Agrário de Vairão, Universidade do Porto, 4485-661 Vairão, Portugal

<sup>2</sup>Department of Biological Sciences, Auburn University, Auburn, AL 36849, USA

<sup>3</sup>Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110, USA

<sup>4</sup>Marine and Environmental Sciences Centre (MARE), Department of Life Sciences, University of Coimbra, 3004-517 Coimbra, Portugal

<sup>5</sup>Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre s/n., 4169-007 Porto, Portugal

<sup>6</sup>Science of Life Laboratory Uppsala, Department of Medical Biochemistry and Microbiology, Uppsala University, 75123 Uppsala, Sweden

<sup>7</sup>Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, 75007 Uppsala, Sweden

<sup>8</sup>Department of Veterinary Integrative Biosciences, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX 77845, USA

<sup>9</sup>Co-first author

\*Correspondence: [hillgee@auburn.edu](mailto:hillgee@auburn.edu) (G.E.H.), [jcorbo@pathology.wustl.edu](mailto:jcorbo@pathology.wustl.edu) (J.C.C.), [miguel.carneiro@cibio.up.pt](mailto:miguel.carneiro@cibio.up.pt) (M.C.)

<http://dx.doi.org/10.1016/j.cub.2016.03.076>

## SUMMARY

The yellow and red feather pigmentation of many bird species [1] plays pivotal roles in social signaling and mate choice [2, 3]. To produce red pigments, birds ingest yellow carotenoids and endogenously convert them into red ketocarotenoids via an oxidation reaction catalyzed by a previously unknown ketolase [4–6]. We investigated the genetic basis for red coloration in birds using whole-genome sequencing of red siskins (*Spinus cucullata*), common canaries (*Serinus canaria*), and “red factor” canaries, which are the hybrid product of crossing red siskins with common canaries [7]. We identified two genomic regions introgressed from red siskins into red factor canaries that are required for red coloration. One of these regions contains a gene encoding a cytochrome P450 enzyme, *CYP2J19*. Transcriptome analysis demonstrates that *CYP2J19* is significantly upregulated in the skin and liver of red factor canaries, strongly implicating *CYP2J19* as the ketolase that mediates red coloration in birds. Interestingly, a second introgressed region required for red feathers resides within the epidermal differentiation complex, a cluster of genes involved in development of the integument. Lastly, we present evidence that *CYP2J19* is involved in ketocarotenoid formation in the retina. The discovery of the carotenoid ketolase has important implications for understanding sensory function and signaling mediated by carotenoid pigmentation.

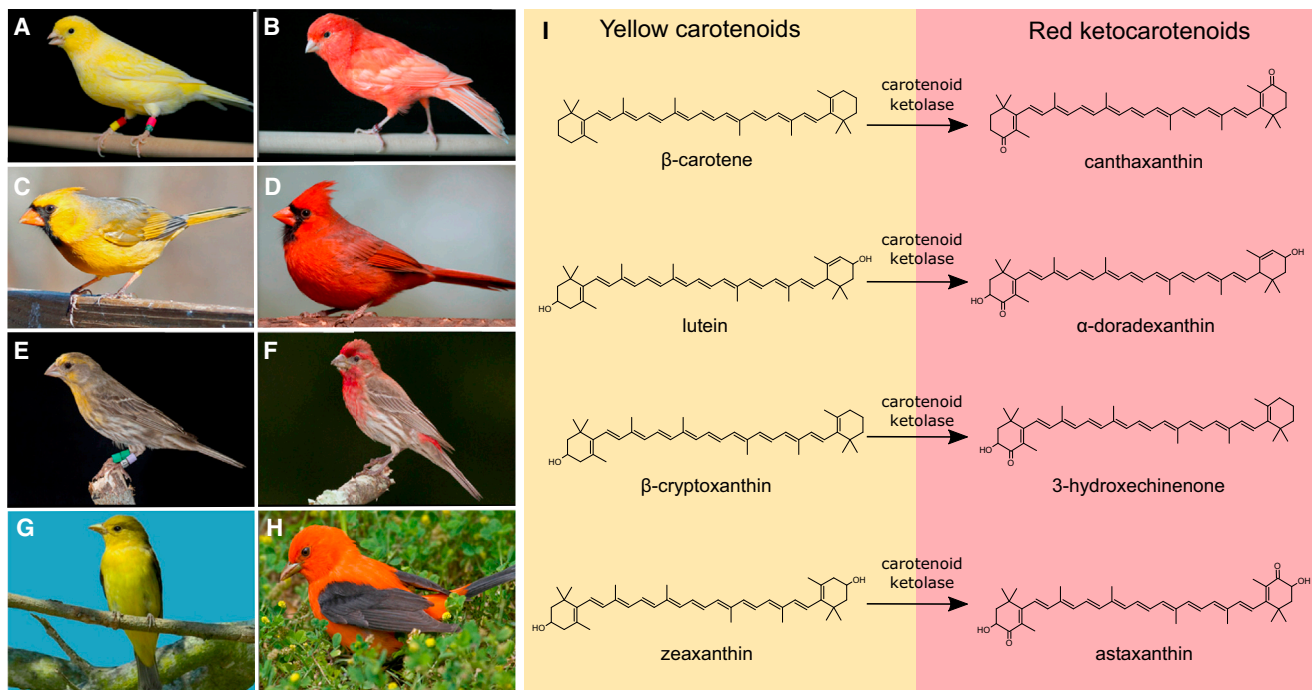
## RESULTS AND DISCUSSION

To identify the genetic basis of red coloration in birds (Figure 1), we took advantage of the unique genetic history of “red factor” canaries. Starting in the 1920s, bird fanciers crossed yellow

common canaries with the red siskin, a South American bird with red ketocarotenoid-pigmented feathers [7]. Hybrid offspring were then backcrossed with common canaries over multiple generations to create the world’s first red factor canary (Figure 2A). Given this genetic history, we reasoned that the genome of red factor canaries would contain regions responsible for red coloration introgressed from red siskins onto a background of common canary DNA. To identify these introgressed regions, we performed whole-genome sequencing of pooled DNA samples from red factor canaries, common canaries (both domestic and wild), and red siskins (Table S1). We generated a total of ~1.5 billion sequence reads that were mapped to the canary reference genome, resulting in an average effective coverage of 19.3× per pool (Table S1).

To detect signatures of genetic differentiation between red factor and common canaries, we measured the fixation index ( $F_{ST}$ ), a metric for summarizing allele frequency differences between populations [12]. We averaged  $F_{ST}$  values across the genome using a sliding-window approach and found that the average level of genetic differentiation was low ( $F_{ST} = 0.079$ ) (Figure 2B), permitting us to detect regions of heightened differentiation indicative of positive selection. The strongest signals of selection in our sliding-window analysis were restricted to two genomic regions (Figure 2B): one located on scaffold NW\_007931131, homologous to zebra finch chromosome 8 (~24,000,000–25,600,000 bp), and the other located on scaffold NW\_007931203, homologous to zebra finch chromosome 25 (~700,000–900,000 bp). All windows above the 99.9<sup>th</sup> percentile of the empirical distribution ( $F_{ST} \geq 0.45$ ) map to these two regions.

Next, we searched for consistent differences in allele frequencies of individual SNPs between two distinct breeds of red factor canaries and five breeds of common canaries. Using a Cochran-Mantel-Haenszel (CMH) test [13], we evaluated 9,414,439 SNPs and found that 15,681 SNPs (0.17%) were significantly associated with red coloration after Bonferroni correction. Importantly, 10,216 of the significant SNPs (65.1%) and all of the top 100 SNPs ( $p \leq 2.43 \times 10^{-18}$ ) localized to the same two genomic regions revealed by the  $F_{ST}$  analysis (Figure 2C).



**Figure 1. Red Feather Coloration Is Mediated by Carotenoid Ketolation**

(A) Yellow common canaries (*Serinus canaria*) lack red ketocarotenoids in their feathers.

(B) Red factor canaries have ketocarotenoid-pigmented plumage. This breed is the product of hybridization between the yellow common canary and the red siskin (*Spinus cucullata*) [7].

(C and D) Rare mutant northern cardinal males (*Cardinalis cardinalis*; C) lack the red ketocarotenoid-containing feathers worn by wild-type males of the species (D) [8].

(E and F) Male house finches (*Haemorrhous mexicanus*) have feather coloration ranging from yellow (E) to red (F). Plumage redness is proportional to the abundance of ketocarotenoids [9, 10].

(G) Nonbreeding male scarlet tanagers (*Piranga olivacea*) have yellow plumage.

(H) Breeding males grow red ketocarotenoid-based plumage [5, 11].

(I) Examples of the metabolic conversions used by birds to produce ketocarotenoids from yellow dietary precursors via the action of a carotenoid ketolase.

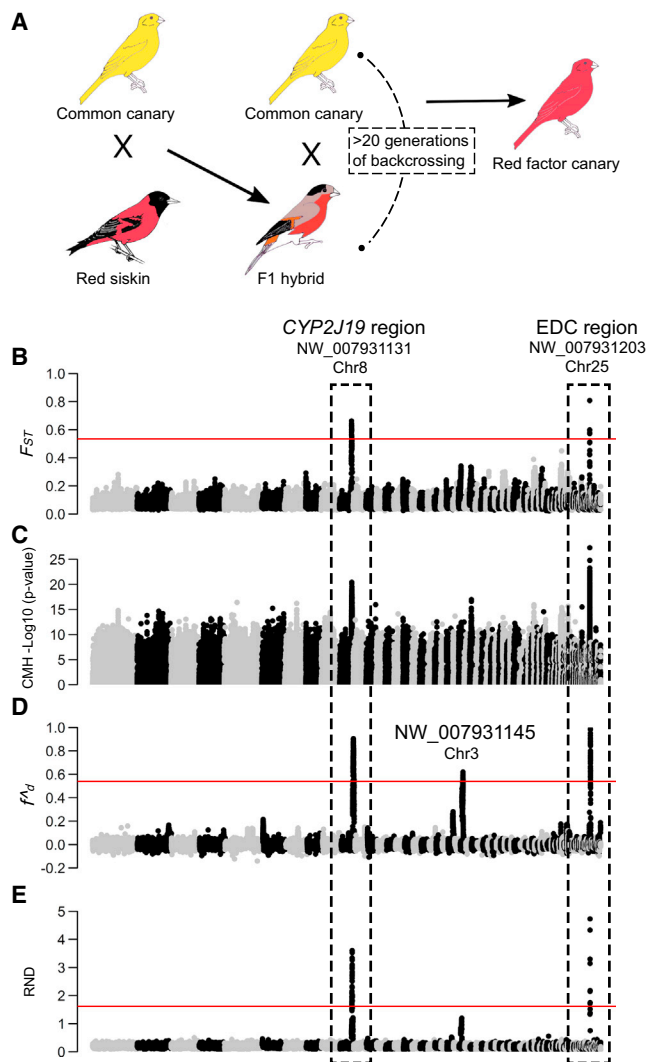
Photo credits: Rebecca J. Koch (A and B), Jim McCormac (C), and Geoffrey E. Hill (D–H).

To locate genomic segments of red siskin origin across the red factor canary genome, we used summary statistics that enabled us to quantify levels of introgression. We started by comparing the genomes of non-red canaries to that of the red siskin. We found that the two species are well differentiated (average nucleotide divergence = 1.77%) and the genomes are well sorted, with 99.4% of all the possible 20 kb windows in the genome displaying at least one diagnostic mutation. This sharp differentiation means that introgressed segments in the red factor genome should be unambiguously identifiable.

We then computed a statistic ( $f^{\wedge}_d$ ) that measures the fraction of the genome shared through introgression [14]. This statistic varies between 0 (no introgression) and 1 (complete replacement). When averaged across the entire genome,  $f^{\wedge}_d$  was close to 0 ( $f^{\wedge}_d = 0.006$ ), suggesting that the overall genetic contribution of red siskin to the red factor canary genome is small, which is consistent with historical records indicating that many generations of backcrossing canary-siskin hybrids to common canaries were necessary to both fix the red trait and improve hybrid fertility [7]. However, the sliding-window analysis identified several segments of the genome with elevated  $f^{\wedge}_d$  values

(Figure 2D), indicative of introgression of red siskin haplotypes in specific genomic regions. The two strongest signals of introgression overlapped the same two top regions in the  $F_{ST}$  analysis and CMH test (NW\_007931131 and NW\_007931203). A third outlier region emerged from this analysis located on scaffold NW\_007931145, which is homologous to zebra finch chromosome 3 (~24,100,000–26,950,000 bp).

The relative node depth statistic (RND) was also calculated between red factor and non-red canaries. RND is a measure of genetic divergence that controls for mutation rate variation, thus allowing us to distinguish between low mutation rate and introgression as the cause of sequence similarity [15]. This analysis pinpointed the same outlier regions, corroborating the previous findings from the  $f^{\wedge}_d$  statistic (Figure 2E). Overall, the substantial overlap between differentiation and introgression statistics indicates that the outlier regions identified here are strong candidates for the genomic regions mediating red coloration in canaries. Furthermore, the fact that at least two genomic regions are implicated in red coloration in canaries (see below) is consistent with the genetic architecture of this trait, which is known to be governed by more than one locus [16].



**Figure 2. The Origin of Red Canaries and Genome-wide Scans for Directional Selection and Introgression**

(A) Red factor canaries were created by crossing common canaries with red siskins. Hybrids were backcrossed with common canaries for many generations to improve the fertility of the line and to remove all siskin characteristics except red coloration. The result is a phenotypically normal canary but with the capacity to produce red ketocarotenoids from yellow dietary carotenoids. (B)  $F_{ST}$  scan across the genome between red and non-red breeds using whole-genome sequencing data (see Table S1). Each dot represents  $F_{ST}$  averaged over 20 Kb windows and iterated in steps of 10 Kb across each scaffold. The 99.9th percentile of the empirical distribution is shown as a red horizontal line. (C)  $-\log_{10}$  values of the CMH statistic for every polymorphic SNP through pairwise comparison of allele frequencies between red and non-red canaries. (D)  $f_{Ad}$  values summarized in non-overlapping windows of 100 SNPs. The 99.9th percentile of the empirical distribution is shown as a red horizontal line. (E) RND values between red and non-red canaries summarized in non-overlapping windows of 10,000 positions (both polymorphic and monomorphic). The 99.9th percentile of the empirical distribution is shown as a red horizontal line. The different scaffolds are presented on the x axis in the same order as they appear in the canary reference genome assembly.

To conduct fine-mapping of these outlier regions, we identified a set of variants that unambiguously distinguish the red siskin and non-red-canary genomes. We then genotyped these vari-

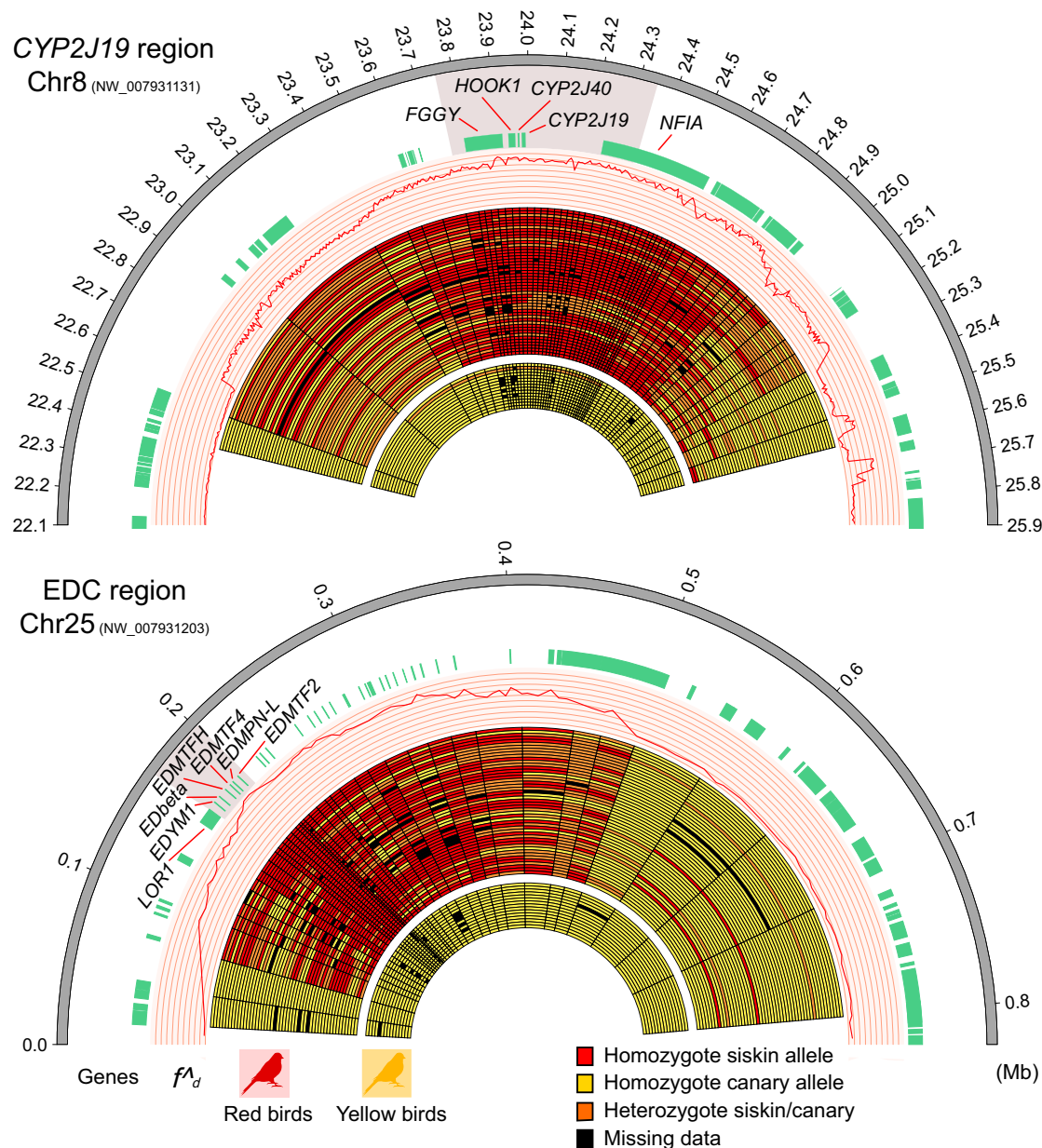
ants in a larger cohort of birds representing a broader range of breeds. For the regions homologous to zebra finch chromosomes 8 (NW\_007931131) and 25 (NW\_007931203), we were able to substantially reduce the size of the candidate intervals (Figure 3). For NW\_007931131, all red birds ( $n = 49$ ) carried at least one haplotype of siskin origin over a 631 kb segment (positions 23,728,812–24,359,539 bp). For NW\_007931203, all red birds were homozygous for the red siskin haplotype encompassing 34 kb (positions 169,696–203,507 bp). In the region homologous to zebra finch chromosome 3 (scaffold NW\_007931145), many red birds did not carry any haplotype of red siskin origin, making it unlikely that any gene in this region is strictly necessary for red coloration (Figure S1). Together with the whole-genome sequencing results (Figure 2), these data indicate that NW\_007931145 contains haplotypes that were introgressed from red siskin but that are segregating at moderate frequencies in red factor canaries (Figure 2B).

Two additional observations from these genotyping results and from animals generated by crossing yellow and red factor canaries provide information about the genetic architecture of red coloration. First, we identified a bird displaying yellow plumage that was heterozygous for the NW\_007931131 region but homozygous for the common canary allele in the NW\_007931203 region. Second, individuals that are heterozygous for both candidate regions do not express red, but rather a yellow/orange coloration. We can thus conclude that both regions are necessary to produce red coloration. The NW\_007931131 region appears to act in a dominant fashion such that a single copy of the red siskin allele is sufficient, whereas homozygosity for the siskin allele in the NW\_007931203 region is necessary to express red feather coloration.

We next investigated the gene content of the two introgressed regions that are necessary for red coloration. Using the published annotation of the canary reference genome, together with an in-depth annotation analysis derived from de novo transcriptome assemblies obtained separately for red and yellow canaries, we identified five genes within the interval on scaffold NW\_007931131: *FGGY*, *HOOK1*, *CYP2J19*, *CYP2J40*, and *NFIA*. Within the interval on scaffold NW\_007931203, we identified six genes—*EDMY1*, *EDbeta*, *EDMTFH*, *EDMTF4*, *EDMPN-L*, and *EDMTF2*—plus the promoter region of one gene (*LOR1*) (detailed in Table S2). The latter interval maps to the epidermal differentiation complex (EDC), an extended block of genes involved in integumentary development [18]. We also utilized whole-genome sequencing data to detect structural variants (i.e., deletions, duplications, inversions, or translocations) that differ between the common canary and red siskin haplotypes (Table S3). We identified several candidate variants, but none overlapped annotated genes, suggesting that the gene content and synteny of both regions is likely to be preserved between the two divergent haplotypes. However, without high-quality genome assemblies for both species, we cannot absolutely exclude the existence of some minor structural differences.

The skin and liver are the two most important anatomical sites for conversion of yellow dietary carotenoids into red ketocarotenoids for plumage coloration in birds [19]. Thus, to further pinpoint candidate causal genes in the introgressed



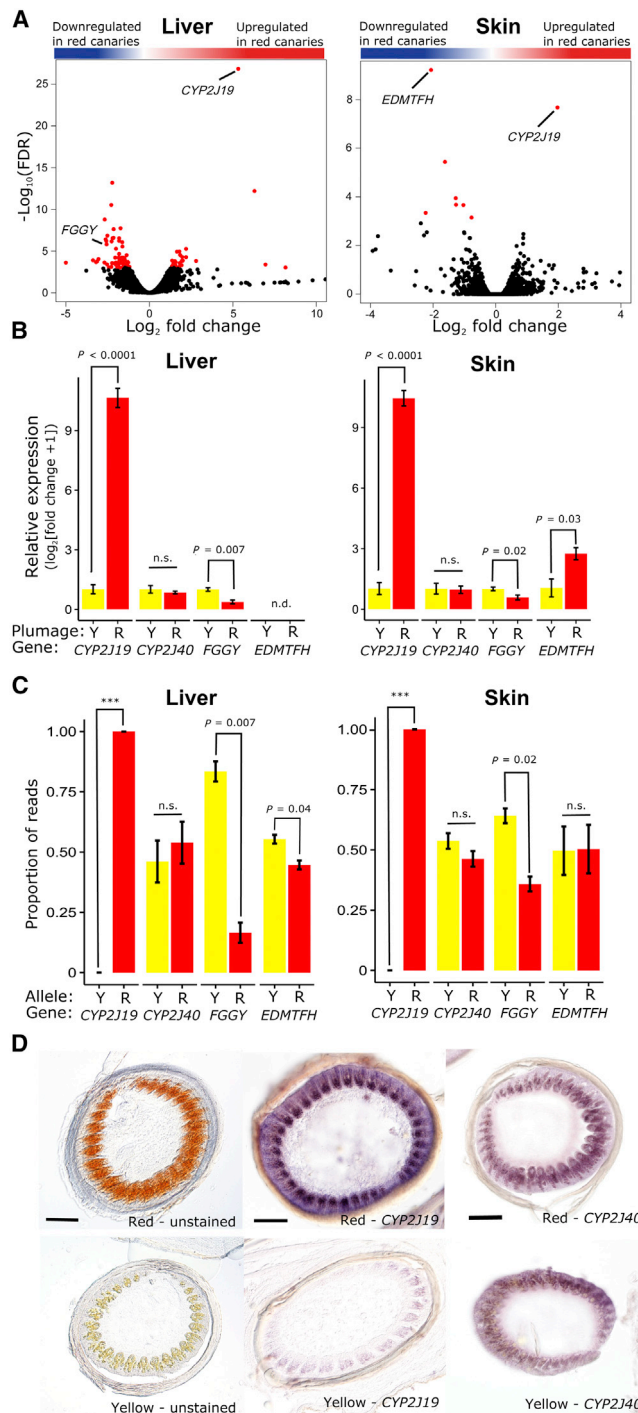


**Figure 3. Fine Mapping of Introgressed Segments from Red Siskin Origin in the Red Canary Genome**

The outermost semi-circle represents the genomic coordinate in megabases. The next semi-circle (from the outside inward) shows the location of genes from the canary genome annotation (green boxes). In the next semi-circle,  $f^d$  values are shown (solid red line). The innermost semi-circles represent the genotyping results for SNPs found to be diagnostic between red siskin and common canaries in 49 red canaries and 15 non-red canaries, as indicated by red and yellow canary silhouettes, respectively. Each row represents one individual, and individuals appear in the same order on both Chr8 and Chr25 graphs. Red, yellow, and orange squares indicate positions homozygous for the red siskin allele, homozygous for the yellow canary allele, and heterozygous for both alleles, respectively. Missing data are represented by black boxes. Light-gray highlighting indicates the longest continuous regions where all red individuals carry at least one copy of the red siskin haplotype. Only the names of genes within these regions are shown (see Table S2). The circular plot was generated using Circos [17]. See also Figure S1.

regions, we analyzed differential gene expression between red and yellow canaries in both adult skin (plucked 10 days prior to induce feather regeneration) and liver by RNA sequencing (RNA-seq; Table S4). We observed nine genes in the skin and 102 genes in the liver that were differentially expressed (false discovery rate [FDR] = 0.1%) (Figure 4A). Within the two candi-

date regions associated with red coloration, we detected significant differential expression of three genes in liver and/or skin: *CYP2J19*, *FGGY*, and *EDMTF2* (Figure 4A). To corroborate the RNA-seq results, we conducted qPCR. We found *CYP2J19* to be expressed at more than 1,000-fold higher levels in both skin and liver of red canaries compared to yellow canaries,



**Figure 4. Levels and Patterns of Gene Expression in Red and Yellow Canaries**

(A) Volcano plot of statistical significance after FDR correction (y axis) against  $\log_2$  fold change in expression (x axis) for both liver and skin using RNA-seq (see Table S4). Significant genes are depicted as red dots, and those that also overlap the candidate regions are labeled.

(B) qPCR analysis of the transcript levels of four genes (*CYP2J19*, *CYP2J40*, *FGGY*, and *EDMTFH*) measured in the liver and skin of three red (R) and three yellow (Y) canaries. Expression is presented as the mean  $\pm$  SD of fold change relative to the mean expression level in the respective tissue of the yellow canaries.

whereas *FGGY* had 2- to 3-fold lower expression in red birds compared to yellow (Figure 4B). In contrast, we found no detectable expression of *EDMTFH* in liver and slightly higher levels in the skin of red birds compared to yellow birds (Figure 4B). This latter result conflicts with the results from RNA-seq, which showed *EDMTFH* to be downregulated in red skin. Upon reexamination of the RNA-seq data, we discovered that the apparent differential expression observed for *EDMTFH* was due to a read-mapping artifact, in which red-skin-derived *EDMTFH* reads mapped less efficiently to the common canary transcriptome assembly. Thus, we conclude that *CYP2J19* expression is markedly elevated in the skin and liver of red birds relative to yellow birds, but that *FGGY* and *EDMTFH* levels are only moderately different between the two.

To further characterize these expression differences, we crossed red factor canaries to common canaries and measured allele-specific expression in the offspring. Strikingly, we found that 100% of *CYP2J19* transcripts were derived from the red allele in both skin and liver in all three offspring examined (Figure 4C). Additionally, we found preferential expression of the yellow allele of *FGGY* in both skin (~83% of normalized reads deriving from the yellow allele) and liver (~64% of normalized reads deriving from the yellow allele). In contrast, we found similar levels of expression of yellow and red alleles of *EDMTFH* in skin and liver (although *EDMTFH* was undetectable in liver by qPCR, it was possible to amplify trace amounts of the transcript by PCR for allelic analysis). Thus, the most notable finding from these studies was the occurrence of extreme allelic imbalance favoring expression of the red-skin-derived allele of *CYP2J19* in both skin and liver.

Unlike *EDMTFH*, which encodes a structural protein with no known enzymatic function [18], both *FGGY* and *CYP2J19* are predicted to encode enzymes and therefore have the potential to mediate the formation of red ketocarotenoids from yellow precursors. Because skin and liver are the two most likely sites of feather ketocarotenoid production in birds [19], a carotenoid ketolase should be upregulated in red bird skin and/or liver relative to yellow birds. Expression of *FGGY* is significantly lower in the red factor canary than in the yellow canary, the opposite of what one would expect for a ketolase. Furthermore, *FGGY* encodes a protein with homology to a family of kinases that phosphorylate carbohydrate substrates [20], a function seemingly

(C) Quantitation by allele-specific RNA-seq of four genes (*CYP2J19*, *CYP2J40*, *FGGY*, and *EDMTFH*) measured in the liver and skin of three F1 red  $\times$  yellow hybrid canaries. Allele expression is presented as the mean  $\pm$  SD of the proportion of red and yellow allele read counts for each gene. Significance was determined with a one-sample Student's *t* test with the null hypothesis of  $\mu = 0.5$ . \*\*\* indicates that 100% of the reads were from the red allele.

(D) Unstained sections and in situ hybridization of regenerating feather follicles (10 days post-pluck) of red and yellow canaries. Reddish-orange ketocarotenoid pigmentation is evident in the developing barb ridges of red canary feather follicles, whereas fainter yellow carotenoid-based pigmentation is evident in the yellow feather (leftmost panels). In situ hybridization probes for *CYP2J19* and *CYP2J40* also localize to the barb ridge of the developing feather. *CYP2J19* expression is markedly elevated in the red canary feather follicles compared to the yellow (middle panels), whereas comparable levels of *CYP2J40* expression are seen in red and yellow feathers (rightmost panels). Scale bars, 100  $\mu$ m.

See also Figures S2 and S3.

unrelated to the lipid oxidation required of a ketolase. Although we cannot exclude a role for FGGY in the production of red feathers, these considerations indicate that it is not an ideal candidate for the ketolase.

*CYP2J19*, in contrast, is markedly elevated in the red factor canary and is a member of a large superfamily of cytochrome P450 oxygenases that act on a range of small-molecule substrates, including carotenoids [21–23]. Members of this family are known to mediate ketolation in the 4 position of the  $\beta$ -ion-one ring of retinoids in mammals [24]. Furthermore, a cytochrome P450 family member, *crtR*, is required for astaxanthin production from  $\beta$ -carotene in the yeast *Xanthophyllomyces dendrorhous* [25].

If skin is a site of ketocarotenoid production in red canaries and *CYP2J19* encodes the carotenoid ketolase, then we expect higher level of *CYP2J19* expression in the regenerating feather follicles of red birds compared to yellow birds. To investigate this possibility, we analyzed the expression of *CYP2J19* in skin by in situ hybridization. We found strong expression of *CYP2J19* in the regenerating feather follicle of red factor canaries, whereas transcripts were present at much lower levels in the skin of yellow birds (Figure 4D). The pattern of *CYP2J19* expression correlates precisely with the distribution of ketocarotenoids in the regenerating feather (Figure 4D). For comparison, we also analyzed the expression of *CYP2J40* and found comparable levels in the regenerating feather follicles of red and yellow birds (Figure 4D). Overall, these findings make *CYP2J19* a strong candidate for the avian carotenoid ketolase.

Because ketocarotenoids are also present in the avian retina, we investigated whether *CYP2J19* might be expressed in this tissue in a pattern consistent with a role in carotenoid ketolation. The red single cone in the avian retina utilizes a ketocarotenoid, astaxanthin, as an intracellular spectral filter to enhance color discrimination [26–28]. Astaxanthin is thought to be synthesized locally within the red cone from a yellow dietary precursor, zeaxanthin (Figures 1 and S2A). We found that *CYP2J19* transcript levels in the chicken retina correlate well with the levels of astaxanthin over development (Figures S2B, S2D, and S2E). Furthermore, in situ hybridization indicates that *CYP2J19* expression is restricted to a subset of cells in the outer nuclear layer, consistent with expression in red single cones (Figure S2C). This correlation is compatible with the hypothesis that *CYP2J19* mediates production of astaxanthin in the developing red cone.

Interestingly, yellow canaries have levels of *CYP2J19* expression in the retina similar to those found in the retinas of red birds (Figure S3). This finding suggests that the expression differences observed between yellow and red birds in the skin and liver are due to *cis*-regulatory differences between the two *CYP2J19* alleles. The vast majority of bird species possess ketocarotenoid-based oil droplets in their retinas and thus presumably express *CYP2J19* in this tissue. However, only a subset of bird species deposit these red pigments in their skin and feathers. Thus, we hypothesize that nearly all bird species have the latent capacity to produce ketocarotenoid-based red feather coloration, but that *cis*-regulatory changes at the *CYP2J19* locus are required for the gain of expression in the skin and/or liver that leads to the emergence of this trait in selected species.

Homozygosity for red siskin alleles in the EDC region is also necessary for red feather production in canaries. If *CYP2J19* encodes the ketolase, then what is the function of the EDC region? We did not identify major differences in gene expression within the EDC region between red and yellow birds, suggesting that coding variants might be involved. To address this point, we scanned all six candidate genes in the EDC region for coding variants. We identified 11 nonsynonymous variants in three genes (*EDMTFH* [12], *EDMPN-L* [1], and *EDMTF2* [2]) (Table S5). In addition, we evaluated the five candidate genes in the *CYP2J19* region for coding variants and found another 24 nonsynonymous mutations in three genes (*HOOK1* [7], *CYP2J40* [7], and *CYP2J19* [14]). We did not identify frameshift or STOP loss/gain mutations in any of the genes. Based on sequence conservation, we predict that several amino acid changes for which the derived state is present in red canaries might alter protein function (Table S5), but functional assays will be required to determine what role, if any, these coding variants play in red feather coloration.

The enzyme that converts yellow carotenoids into red ketocarotenoids in birds has long been sought [5, 7, 29, 30]. Here, we present *CYP2J19* as a strong candidate for the carotenoid ketolase. Genetic and expression analysis in canary and chicken support *CYP2J19* as the ketolase, or a component thereof. In a co-submitted paper, *CYP2J19* was also identified as the gene responsible for red coloration in the bill and legs of zebra finches [31], suggesting that it serves as the ketolase in multiple tissues across diverse groups of birds. Carotenoid coloration is widely accepted as a condition-dependent signal of individual quality that is assessed in mate choice and other social interactions [32, 33]. The discovery of the ketolase gene in birds presents unprecedented opportunities for investigating the signal content of carotenoid coloration [34]. The oxidation potential required by P450 enzymes suggests that production of red pigments via *CYP2J19* will be sensitive to the organism-wide oxidative state, potentially explaining why red carotenoid coloration is so consistently linked to individual quality [35, 36].

## ACCESSION NUMBERS

The accession number for the whole-genome resequencing and RNA-seq read data reported in this paper is GenBank: SRP065487. The accession number for the SNP genotyping data and de novo transcriptome assemblies reported in this paper is Dryad: <http://dx.doi.org/10.5061/dryad.sm12c>.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, three figures, and five tables and can be found with this article online at <http://dx.doi.org/10.1016/j.cub.2016.03.076>.

## AUTHOR CONTRIBUTIONS

M.C., J.C.C., and G.E.H. conceived and coordinated the study with input from R.J.L. and L.A. M.C., J.D.J., M.S.F., and J.M.-F. conducted the bioinformatic analyses of the whole-genome resequencing and RNA-seq datasets. M.C., R.J.L., and P.M.A. maintained the canaries and performed the controlled crosses. M.B.T. and J.C.C. designed and performed qPCR, in situ hybridization, allele-specific expression experiments, and retinal analyses. M.C., J.C.C., and G.E.H. wrote the paper with input from other authors.



## ACKNOWLEDGMENTS

The work was supported by Fundação para a Ciência e Tecnologia (FCT) through POPH-QREN funds from the European Social Fund and Portuguese MCTES (FCT Investigator grants to M.C. [IF/00283/2014/CP1256/CT0012] and J.M.-F. [IF/00033/2014] and a post-doc fellowship to R.J.L. [SFRH/BPD/84141/2012]); by a research fellowship (PD/BD/108131/2015) attributed to MSF in the scope of the Biodiversity, Genetics, and Evolution (BIODIV) PhD program at CIBIO/InBIO and University of Porto; by the projects “Genomics and Evolutionary Biology” and “Genomics Applied to Genetic Resources” co-financed by North Portugal Regional Operational Programme 2007/2013 (ON.2—O Novo Norte) under the National Strategic Reference Framework (NSRF) and the European Regional Development Fund (ERDF); and by an EU FP7 REGPOT grant (CIBIO-New-Gen) (286431). G.E.H. and J.D.J. were funded by NSF-IOS (1243207) and by the Office of the Vice President for Research at Auburn University. J.C.C. and M.B.T. were funded by a Human Frontier Science Program grant and grants from the NIH (R01EY024958 and R01EY026672). M.B.T. was additionally funded by fellowships from the National Science Foundation (award no. 1202776), NIH (EY013360), and McDonnell Center for Cellular and Molecular Neurobiology at Washington University. We thank N. Rafati for help with expression analysis, S. Afonso for help with multiple laboratory procedures, Jess Hoisington-López at the Center for Genome Sciences and Systems Biology for sequencing advice and expertise, D. Hughes for advice on statistical analysis, and S. Shen for careful reading of the manuscript. We thank four anonymous reviewers for their constructive comments on the manuscript. This study was conducted following the recommendations of Guide for the Care and Use of Laboratory Animals of the NIH and following Portuguese and European regulations for the maintenance of live birds in captivity (Federation of European Laboratory Animal Science Associations, FELASA). All of the animals were handled according to a protocol approved by the Animal Studies Committee at Washington University in St. Louis (ASC protocol no. 20140072). Wild canary samples were collected under the auspices of a local collecting license (Cabildo de La Palma, Consejería de Medio Ambiente, Transportes y Seguridad y Emergencias).

Received: October 2, 2015

Revised: February 8, 2016

Accepted: March 24, 2016

Published: May 19, 2016

## REFERENCES

- Stradi, R. (1998). *The Colour of Flight: Carotenoids in Bird Plumages* (Solei Gruppo Editoriale Informatico).
- Hill, G.E. (2006). Female choice for ornamental coloration. In *Bird Coloration, Volume 2: Function and Evolution*, G.E. Hill, and K.J. McGraw, eds. (Harvard University Press).
- Pryke, S.R., Lawes, M.J., and Andersson, S. (2011). Agonistic carotenoid signalling in male red-collared widowbirds: aggression related to the colour signal of both the territory owner and model intruder. *Anim. Behav.* 62, 695–704.
- Goodwin, T. (1984). *Animals, Volume 2: The Biochemistry of Carotenoids* (Chapman and Hall).
- Brush, A.H. (1990). Metabolism of carotenoid pigments in birds. *FASEB J.* 4, 2969–2977.
- Koch, R.E., Hill, G.E., and McGraw, K.J. (2016). Effects of diet on plumage coloration and carotenoid deposition in red and yellow domestic canaries (*Serinus canaria*). *Wilson J. Ornithol.* <http://dx.doi.org/10.1676/15-153.1>.
- Birkhead, T. (2003). *A Brand-New Bird: How Two Amateur Scientists Created the First Genetically Engineered Animal* (Basic Books).
- McGraw, K.J., Hill, G.E., and Parker, R.S. (2003). Carotenoid pigments in a mutant cardinal: implications for the genetic and enzymatic control mechanisms of carotenoid metabolism in birds. *Condor* 105, 587–592.
- Inouye, C.Y., Hill, G.E., Montgomerie, R., and Stradi, R.D. (2001). Carotenoid pigments in male house finch plumage in relation to age, subspecies, and ornamental coloration. *Auk* 118, 900–915.
- McGraw, K.J., Nolan, P.M., and Crino, O.L. (2006). Carotenoid accumulation strategies for becoming a colourful House Finch: analyses of plasma and liver pigments in wild moulting birds. *Funct. Ecol.* 20, 678–688.
- Hudon, J. (1991). Unusual carotenoid use by the Western Tanager (*Piranga ludoviciana*) and its evolutionary implications. *Can. J. Zool.* 69, 2311–2320.
- Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* 38, 1358–1370.
- McDonald, J.H. (2009). *Handbook of Biological Statistics, Second Edition* (Sparky House Publishing).
- Martin, S.H., Davey, J.W., and Jiggins, C.D. (2015). Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* 32, 244–257.
- Feder, J.L., Xie, X., Rull, J., Velez, S., Forbes, A., Leung, B., Dambroski, H., Filchak, K.E., and Aluja, M. (2005). Mayr, Dobzhansky, and Bush and the complexities of sympatric speciation in *Rhagoletis*. *Proc. Natl. Acad. Sci. USA* 102 (Suppl 1), 6573–6580.
- Perez-Beato, O. (2008). *Fundamentals of Color Genetics in Canaries: Reproduction and Control* (RoseDog Books).
- Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645.
- Strasser, B., Miltz, V., Hermann, M., Rice, R.H., Eigenheer, R.A., Alibardi, L., Tschachler, E., and Eckhart, L. (2014). Evolutionary origin and diversification of epidermal barrier proteins in amniotes. *Mol. Biol. Evol.* 31, 3194–3205.
- del Val, E., Senar, J.C., Garrido-Fernández, J., Jarén, M., Borràs, A., Cabrera, J., and Negro, J.J. (2009). The liver but not the skin is the site for conversion of a red carotenoid in a passerine bird. *Naturwissenschaften* 96, 797–801.
- Zhang, Y., Zagnitko, O., Rodionova, I., Osterman, A., and Godzik, A. (2011). The FGGY carbohydrate kinase family: insights into the evolution of functional specificities. *PLoS Comput. Biol.* 7, e1002318.
- Coon, M.J. (2005). Cytochrome P450: nature’s most versatile biological catalyst. *Annu. Rev. Pharmacol. Toxicol.* 45, 1–25.
- Guengerich, F.P., and Cheng, Q. (2011). Orphans in the human cytochrome P450 superfamily: approaches to discovering functions and relevance in pharmacology. *Pharmacol. Rev.* 63, 684–699.
- Nelson, D.R., Koymans, L., Kamataki, T., Stegeman, J.J., Feyereisen, R., Waxman, D.J., Waterman, M.R., Gotoh, O., Coon, M.J., Estabrook, R.W., et al. (1996). P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics* 6, 1–42.
- Roos, T.C., Jugert, F.K., Merk, H.F., and Bickers, D.R. (1998). Retinoid metabolism in the skin. *Pharmacol. Rev.* 50, 315–333.
- Alcaino, J., Barahona, S., Carmona, M., Lozano, C., Marcoleta, A., Niklitschek, M., Sepúlveda, D., Baeza, M., and Cifuentes, V. (2008). Cloning of the cytochrome p450 reductase (crtR) gene and its involvement in the astaxanthin biosynthesis of *Xanthophyllomyces dendrorhous*. *BMC Microbiol.* 8, 169.
- Vorobyev, M. (2003). Coloured oil droplets enhance colour discrimination. *Proc. Biol. Sci.* 270, 1255–1261.
- Toomey, M.B., Collins, A.M., Frederiksen, R., Cornwall, M.C., Timlin, J.A., and Corbo, J.C. (2015). A complex carotenoid palette tunes avian colour vision. *J. R. Soc. Interface* 12, 20150563.
- Goldsmith, T.H., Collins, J.S., and Licht, S. (1984). The cone oil droplets of avian retinas. *Vision Res.* 24, 1661–1671.
- Wald, G., and Zussman, H. (1937). Carotenoids of the chicken retina. *Nature* 140, 197.
- Davies, B.H. (1985). Carotenoid metabolism in animals: a biochemist’s view. *Pure Appl. Chem.* 57, 679–684.



31. Mundy, N.I., Stapley, J., Bennison, C., Tucker, R., Twyman, H., Kim, K.-W., Burke, T., Birkhead, T.R., Andersson, S., and Slate, J. (2016). Red carotenoid coloration in the zebra finch is controlled by a cytochrome P450 gene cluster. *Curr. Biol.* Published online May 19, 2016. <http://dx.doi.org/10.1016/j.cub.2016.04.047>.
32. Hill, G.E. (2002). *A Red Bird in a Brown Bag: The Function and Evolution of Colorful Plumage in the House Finch* (Oxford University Press).
33. Svensson, P.A., and Wong, B.B.M. (2011). Carotenoid-based signals in behavioural ecology: a review. *Behaviour* **148**, 131–189.
34. Hill, G.E. (2006). Environmental regulation of ornamental coloration. In *Bird Coloration, Volume 1: Mechanisms and Measurements*, G.E. Hill, and K.J. McGraw, eds. (Harvard University Press).
35. Hill, G.E., and Johnson, J.D. (2012). The vitamin A-redox hypothesis: a biochemical basis for honest signaling via carotenoid pigmentation. *Am. Nat.* **180**, E127–E150.
36. Johnson, J.D., and Hill, G.E. (2013). Is carotenoid ornamentation linked to the inner mitochondria membrane potential? A hypothesis for the maintenance of signal honesty. *Biochimie* **95**, 436–444.

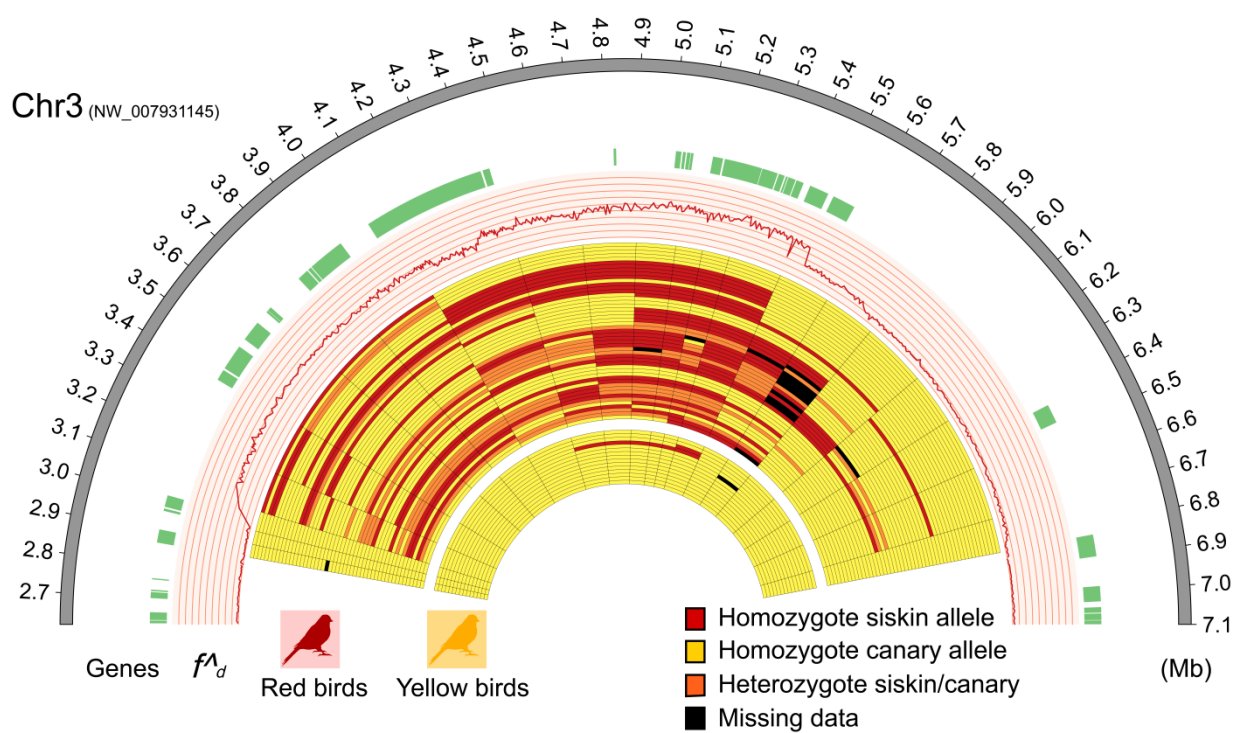
**Current Biology, Volume 26**

## **Supplemental Information**

### **Genetic Basis for Red Coloration in Birds**

**Ricardo J. Lopes, James D. Johnson, Matthew B. Toomey, Mafalda S. Ferreira, Pedro M. Araujo, José Melo-Ferreira, Leif Andersson, Geoffrey E. Hill, Joseph C. Corbo, and Miguel Carneiro**

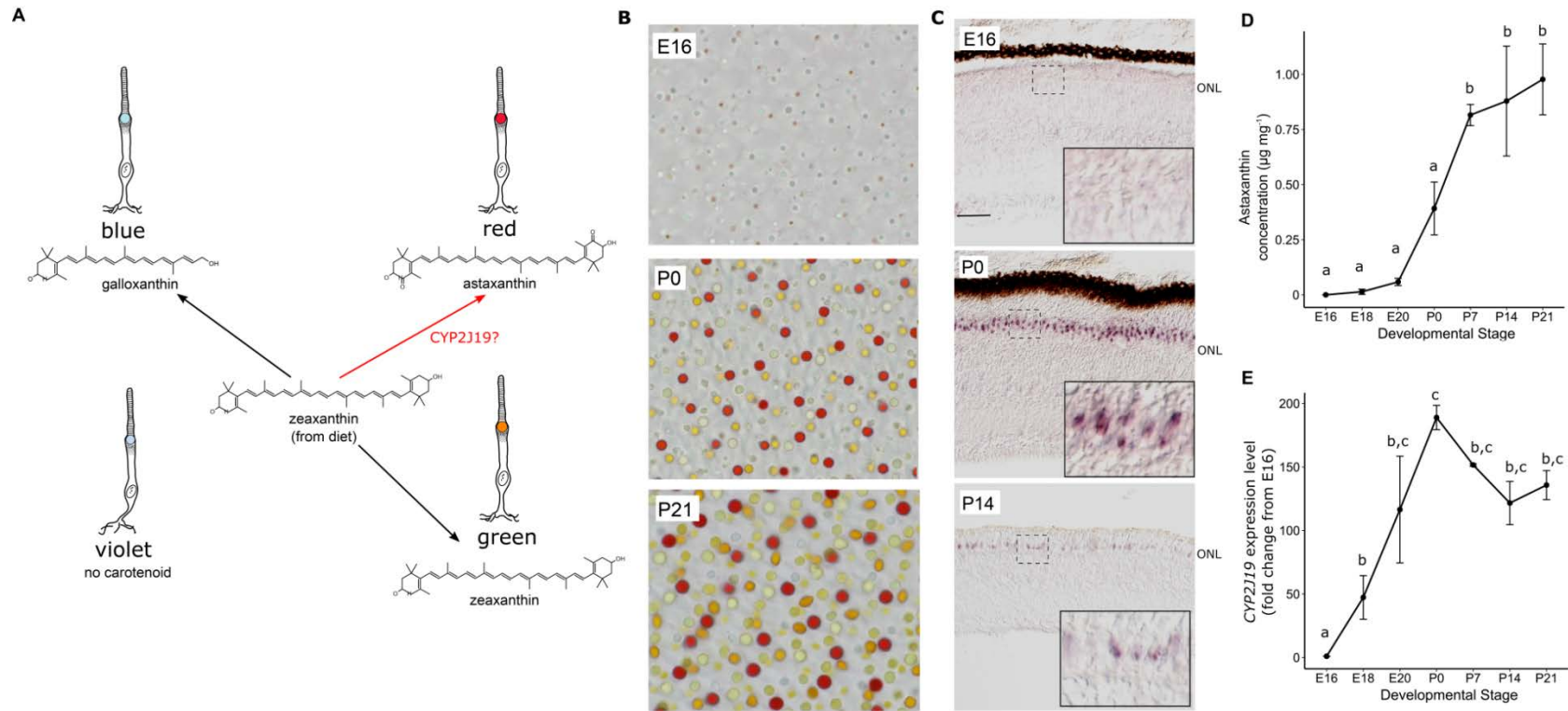
**FIGURE S1**



**Figure S1. Related to Figure 3. Fine scale mapping of a third outlier segment from the introgression analysis. See Figure 3 legend for details.**

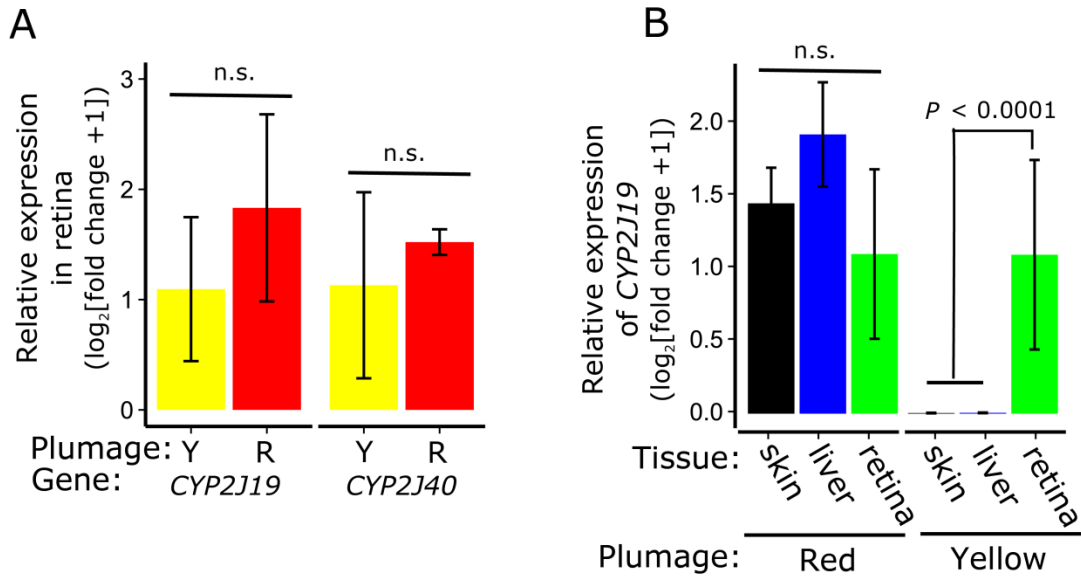


**FIGURE S2**



**Figure S2. Related to Figure 4. *CYP2J19* expression correlates with astaxanthin levels in the developing chicken retina. (A)** The single cone photoreceptors of the avian retina contain oil droplets pigmented with the indicated carotenoids, all derived from zeaxanthin deposited in the yolk or derived from the diet [S1]. The red single cone contains a red oil droplet pigmented with the ketocarotenoid, astaxanthin. We hypothesize that *CYP2J19* converts zeaxanthin into astaxanthin in the red single cone. **(B)** Brightfield images of a flatmounted retina (photoreceptor side up) showing how the size and pigmentation of cone oil droplets increase rapidly over the course of development from 16 days of incubation (E16), through hatch day (P0), up to 21 days post-hatch (P21) in the chicken (*Gallus gallus*). **(C)** *In situ* hybridization of the developing chicken retina with a probe targeting *CYP2J19*. Vertical sections of stained retinas are shown. High magnification images of the dotted regions are shown as insets. The signal is first detectable at hatch day (P0) and is localized to a subset of photoreceptors in the outer nuclear layer (ONL). Scale bar = 50  $\mu$ m. **(D)** The concentration of astaxanthin in whole retina extracts from chicken (*Gallus gallus*) increases significantly over the course of development ( $F_{6,12} = 28.6$ ,  $P < 0.0001$ ). Concentration is presented as mean  $\pm$  standard deviation  $\mu$ g of astaxanthin per mg of protein and points denoted with different letters are significantly different (Tukey's posthoc test  $P < 0.02$ ). **(E)** Quantitative PCR indicates that *CYP2J19* expression in the chicken retina increases significantly across development ( $F_{6,12} = 45.3$ ,  $P < 0.0001$ ). Values are reported as the mean  $\pm$  standard deviation of the fold change relative to the earliest time point (E16), and points that do not share a letter in common are significantly different (Tukey's posthoc test  $P < 0.02$ ).

**FIGURE S3**



**Figure S3. Related to Figure 4. *CYP2J19* expression in yellow canaries is restricted to the retina.** (A) There is no significant difference (two-sample t-test all:  $t \geq -1.42$ ,  $P \geq 0.317$ ) in the transcript levels of *CYP2J19* and *CYP2J40* in the retinas of yellow ( $n = 3$ ) and red ( $n = 3$ ) canaries as measured by qPCR. Values are reported as the mean  $\pm$  standard deviation of the fold change relative to the mean transcript levels in yellow canaries. (B) There is no significant difference ( $F_{2,6} = 3.04$ ,  $P = 0.201$ ) in the transcript levels of *CYP2J19* among the tissues of red canaries. In yellow canaries, transcript levels of *CYP2J19* were significantly higher ( $F_{2,6} = 135.4$ ,  $P < 0.0001$ ) in the retina as compared to other tissues. Values are reported as the mean  $\pm$  standard deviation of the fold change relative to the mean transcript levels in retina.

**Table S1. Related to Figure 2. Whole genome resequencing details and read mapping statistics.**

Breed/Population	Species	Plumage coloration	Number of individuals pooled	Number of reads <sup>a</sup>	Percentage of reads mapping <sup>b</sup>	Percentage of reads MQ $\geq 20$ <sup>c</sup>	Percentage of positions $\geq 1$ reads <sup>d</sup>	Mean depth of coverage <sup>e</sup>
Lipochrome red	<i>Serinus canaria</i>	Red	16	202,038,581	99.24 (92.31)	88.42	96.02	16.9X
Black red	<i>Serinus canaria</i>	Red	12	275,063,528	99.16 (92.71)	87.61	96.25	23.0X
White recessive	<i>Serinus canaria</i>	White	12	234,684,768	97.21 (90.48)	82.34	96.28	17.4X
Gibber italicus	<i>Serinus canaria</i>	Yellow	12	219,183,670	99.08 (91.87)	86.98	96.03	18.2X
Wild	<i>Serinus canaria</i>	Yellow	39	266,623,880	97.48 (92.96)	86.03	96.21	20.1X
Red siskin	<i>Spinus cucullata</i>	Red	12	276,281,300	94.69 (91.84)	82.96	93.70	19.9X

<sup>a</sup>After trimming using Trimmomatic.

<sup>b</sup>Percentage of properly paired reads is given in parentheses.

<sup>c</sup>Reads with a Phred score mapping quality (MQ) equal or higher than 20.

<sup>d</sup>Percentage of positions with at least one read mapping.

<sup>e</sup>Including positions with zero reads mapping.



**Table S2. Related to Figure 3. List of the genes located with candidate regions for red coloration.**

Gene	Scaffold	Location	Orientation	Description
<i>FGGY</i>	NW_007931131	23804313-23921611	+	FGGY Carbohydrate Kinase Domain Containing
<i>HOOK1</i>	NW_007931131	23940664-23961203	+	Hook microtubule-tethering protein 1
<i>CYP2J40</i>	NW_007931131	23966246-23970926	-	Cytochrome P450 2J2-like
<i>CYP2J19</i>	NW_007931131	23974438-23983434	-	Cytochrome P450 2J2-like
<i>NFIA</i>	NW_007931131	24234130-24577671	+	Nuclear Factor I/A
<i>LORI*</i>	NW_007931203	156,540-167,023	-	Loricrin-like
<i>EDYM1</i>	NW_007931203	177,527-177,973	-	Epidermal Differentiation protein containing Y Motif 1
<i>EDbeta</i>	NW_007931203	182,976-183,346	+	Epidermal Differentiation protein beta
<i>EDMTFH</i>	NW_007931203	188,745-189,094	-	Epidermal Differentiation protein starting with MTF motif and rich in Histidine
<i>EDMTF4</i>	NW_007931203	192,347-192,710	+	Epidermal Differentiation protein starting with MTF motif 4
<i>EDMPN-L</i>	NW_007931203	194,523-194,732	-	Epidermal Differentiation protein containing a MPN sequence like
<i>EDMTF2</i>	NW_007931203	199,920-200,159	-	Epidermal Differentiation protein starting with MTF motif 2

\*The coding region of LOR1 is not included within the candidate region

**Table S3. List of structural variants reported within both outlier regions using three methods (*BreakDancer*, *DELly*, *LUMPY*). All variants were visually inspected using IGV (Integrative Genomics viewer) to ascertain the likelihood of being true variants.**

Scaffold	Position-1	Position-2	Type	Size	Method	True/False	Overlapping annotated genes	Differential mapping or read pair patterns between red breeds/siskin and common canaries
NW_007931130	23932595	23933151	DEL	522	<i>BreakDancer</i>	True	No	Yes
NW_007931131	24058510	24059243	INV	297	<i>BreakDancer; DELly</i>	True	No	Yes
NW_007931131	24154916	24161980	DEL	7064	<i>BreakDancer; DELly; LUMPY</i>	(False positive) <sup>1</sup>	-	No
NW_007931203	170832	171517	DUP	685	<i>DELly; LUMPY</i>	True	No	Yes
NW_007931203	171002	198985	INV	27983	<i>DELly; LUMPY</i>	(False positive) <sup>2</sup>	<i>EDMY1; Edbeta; EDMTFH; EDMTF4; EDMPN-L</i>	Yes
NW_007931203	179472	179721	DUP <sup>2</sup>	249	<i>DELly</i>	(False positive) <sup>3</sup>	-	No
NW_007931203	184471	347600	DEL <sup>3</sup>	162905	<i>DELly</i>	(False positive) <sup>4</sup>	-	Yes
NW_007931203	184689	347598	DUP <sup>3</sup>	162909	<i>LUMPY</i>	(False positive) <sup>5</sup>	-	Yes
NW_007931203	199456	207941	INV	8484	<i>DELly</i>	(False positive) <sup>6</sup>	-	No

Inferred variants of the same type with overlapping boundaries and inferred by multiple methods were collapsed and the most extreme 5' and 3' ends are given.

<sup>1</sup>Likely a false positive due to possible misassembly and highly repetitive region.

<sup>2</sup>False positive confirmed by PCR and Sanger sequencing of 1 red siskin and 1 red canary. The obtained PCR product was similar in size and sequence to the canary reference genome sequence.

<sup>3</sup>Likely a false positive due to possible misassembly or highly repetitive region. Many reads have their paired read mapping to various small scaffolds.

<sup>4</sup>The large variant inferred is a false positive. Although there is differences in read pair patterns between red breeds/siskin and common canaries in both coordinates (5' and 3'), read depth does not support either a deletion or a duplication.

<sup>5</sup>Likely false positive due to misassembly. Several reads in both coordinates have their paired read mapping on the other coordinate. However, this pattern was observed both in red breeds/siskin and common canaries.

**Table S4. Related to Figure 4. Number of RNA-seq reads for each sample, phenotype and tissue type before and after filtering.**

Tissue	Phenotype	Sample Name	No. Reads raw data	No. Reads after filtering
Liver	Red	red36L	35,255,036	34,459,236
		red37L	50,097,120	48,962,080
		red38L	30,386,458	29,645,156
		red39L	44,560,912	43,607,448
		red40L	58,816,138	57,490,160
	Yellow	yel3L	46,936,126	45,888,670
		yel4L	57,880,326	56,470,040
		yel5L	73,415,842	71,750,786
Skin	Red	red36S	44,077,712	42,808,884
		red37S	44,027,624	42,708,234
		red38S	26,433,622	25,659,742
		red39S	30,264,906	29,391,528
		red40S	28,954,682	28,081,226
	Yellow	yel3S	38,191,944	37,101,102
		yel4S	26,874,364	26,042,484
		yel5S	51,621,108	50,171,384
Total		851634728	829746964	



**Table S5. Nonsynonymous mutations detected within the two candidate regions and their potential functional impact inferred using SIFT.**

Scaffold	Gene	Position genome (bp)	Reference base	Alternative base	Aminoacid and mRNA mutation details	Number Species position	Derived allele in red		Derived allele in yellow	
							Effect	Score	Effect	Score
NW_007931131	<i>HOOK1</i>	23,948,028	C	T	p.Thr111Ile/c.332C>T	45	Potentially Functional	0.00	-	-
NW_007931131	<i>HOOK1</i>	23,950,021	C	G	p.Leu160Val/c.478C>G	45	-	-	Potentially Functional	0.03
NW_007931131	<i>HOOK1</i>	23,950,034	C	T	p.Ser164Phe/c.491C>T	45	Potentially Functional	0.00	-	-
NW_007931131	<i>HOOK1</i>	23,950,692	G	A	p.Arg187Gln/c.560G>A	45	Potentially Functional*	0.03	Potentially Functional*	0.03
NW_007931131	<i>HOOK1</i>	23,953,400	A	G	p.Asp230Gly/c.689A>G	46	Potentially Functional	0.00	-	-
NW_007931131	<i>HOOK1</i>	23,953,464	C	A	p.Asp251Glu/c.753C>A	46	-	-	Potentially Functional	0.00
NW_007931131	<i>HOOK1</i>	23,955,448	C	T	p.His348Tyr/c.1042C>T	46	Potentially Functional	0.00	-	-
NW_007931131	<i>CYP2J40</i>	23,966,575	G	T	p.Thr461Asn/c.1382C>A	37	-	-	Potentially Tolerated	0.51
NW_007931131	<i>CYP2J40</i>	23,967,380	A	G	p.Met393Thr/c.1178T>C	41	Potentially Functional	0.02	-	-
NW_007931131	<i>CYP2J40</i>	23,968,338	C	T	p.Asp283Asn/c.847G>A	41	-	-	Potentially Tolerated	0.18
NW_007931131	<i>CYP2J40</i>	23,968,674	C	T	p.Arg252Gln/c.755G>A	42	-	-	Potentially Tolerated	0.40
NW_007931131	<i>CYP2J40</i>	23,970,208	T	C	p.Asn70Asp/c.208A>G	42	Potentially Tolerated*	0.31	Potentially Tolerated*	0.36
NW_007931131	<i>CYP2J40</i>	23,970,223	T	C	p.Ile65Val/c.193A>G	40	Potentially Functional	0.02	-	-
NW_007931131	<i>CYP2J40</i>	23,970,754	G	A	p.Ala58Val/c.173C>T	36	Potentially Functional	0.00	-	-
NW_007931131	<i>CYP2J19</i>	23,974,442	G	T	p.Pro479His/c.1436C>A	41	Potentially Tolerated	0.72	-	-
NW_007931131	<i>CYP2J19</i>	23,975,267	A	C	p.Asn386Lys/c.1158T>G	46	Does not result in an aminoacid change due to the next mutation.			
NW_007931131	<i>CYP2J19</i>	23,975,268	T	A	p.Asn386Ile/c.1157A>T	46	-	-	Potentially Tolerated	0.06
NW_007931131	<i>CYP2J19</i>	23,977,767	C	T	p.Asp245Asn/c.733G>A	46	Potentially Tolerated	0.13	-	-
NW_007931131	<i>CYP2J19</i>	23,977,865	G	A	p.Ser212Leu/c.635C>T	46	Potentially Tolerated	0.39	-	-
NW_007931131	<i>CYP2J19</i>	23,983,334	A	G	p.Val34Ala/c.101T>C	46	Potentially Functional	0.01	-	-
NW_007931131	<i>CYP2J19</i>	23,983,344	A	G	p.Phe31Leu/c.91T>C	48	-	-	Potentially Tolerated	0.21
NW_007931131	<i>CYP2J19</i>	23,983,385	T	C	p.Lys17Arg/c.50A>G	48	Potentially Tolerated	0.33	-	-
NW_007931131	<i>CYP2J19</i>	23,983,399	C	A	p.Met12Ile/c.36G>T	48	Potentially Functional*	0.01	Potentially Functional*	0.00
NW_007931131	<i>CYP2J19</i>	23,983,422	C	T	p.Val5Ile/c.13G>A	44	-	-	Potentially Tolerated	0.18
NW_007931203	<i>EDYM1</i>	177,669	T	G	p.Gln102Pro/c.305A>C	39	Potentially Functional	0.00	-	-
NW_007931203	<i>EDMTFH</i>	188,777	C	G	p.Leu106Phe/c.318G>C	6	Potentially Functional	0.00	-	-
NW_007931203	<i>EDMTFH</i>	188,832	C	T	p.Arg88Lys/c.263G>A	6	Potentially Functional	0.03	-	-
NW_007931203	<i>EDMTFH</i>	188,869	C	A	p.Val76Phe/c.226G>T	6	Potentially Tolerated	0.38	-	-
NW_007931203	<i>EDMTFH</i>	188,870	G	C	p.Ile75Met/c.225C>G	6	Potentially Functional	0.00	-	-
NW_007931203	<i>EDMTFH</i>	188,889	G	A	p.Ala69Val/c.206C>T	6	Potentially Functional	0.01	-	-
NW_007931203	<i>EDMTFH</i>	188,896	C	T	p.Val67Ile/c.199G>A	6	Potentially Functional	0.00	-	-
NW_007931203	<i>EDMTFH</i>	188,906	T	C	p.Ile63Met/c.189A>G	6	Potentially Tolerated	1.00	-	-
NW_007931203	<i>EDMTFH</i>	188,916	C	G	p.Ser60Thr/c.179G>C	6	Potentially Functional	0.00	-	-
NW_007931203	<i>EDMPN-L</i>	194,638	C	G	p.Ser32Thr/c.95G>C	31	Potentially Tolerated*	0.13	Potentially Tolerated*	0.13
NW_007931203	<i>EDMTF2</i>	200,102	C	T	p.Gly20Ser/c.58G>A	32	Potentially Tolerated	0.29	-	-
NW_007931203	<i>EDMTF2</i>	200,126	A	G	p.Ser12Pro/c.34T>C	32	Potentially Tolerated	0.19	-	-

\*Ancestral-Derived relationship could not be ascertained.

## SUPPLEMENTAL EXPERIMENTAL PROCEDURES

### Canary aviculture, sampling, and husbandry

Domestic canaries are descended from common canaries captured in the Canary Islands, and some lineages are centuries removed from the wild [S2]. Just as in dogs and other domestic animals, selective breeding created breeds of canaries with specific phenotypic traits, such as an absence of melanin pigmentation in feathers. A breed is a population of canaries defined by a specific phenotype. Breeds can show modest levels of genetic differentiation.

Given that pedigree and thus ancestry information for most canary birds is not as detailed as in other domesticated species such as dogs, in order to assure that the sampled individuals conform to the standards described for the breeds, we collected canary blood samples from aviaries of Portuguese licensed breeders with a known record of awards in national/international ornithological shows. Blood samples were obtained from selected live birds by brachial venipuncture with a sterile needle, following standard blood collection procedures. Blood was collected into a heparin-free capillary tube and immediately transferred into a vial with 96% ethanol. Similarly to canary breeds, DNA material from red siskins was obtained from local breeders following identical procedures. Samples from the Canary Islands were collected on La Palma Island from wild birds captured using mist-nets (Ecotone, Gdynia, Poland). Each bird was sampled following the same procedure described above, aged and sexed, and released at the site of capture. Samples were collected by RJL under the auspices of a local collecting license (Cabildo de La Palma, Consejería de Medio Ambiente, Transportes y Seguridad y Emergencias). All samples were collected by licensed authors of this study.

For RNA-seq analysis, birds were obtained from aviaries of Portuguese licensed breeders. They were kept under the same conditions in experimental facilities. Birds sacrificed for gene expression studies were held for 10-15 days following acquisition, and animal care complied with national and international regulations for the maintenance of live birds in captivity (FELASA, Federation of European Laboratory Animal Science Associations). Birds were euthanized following accepted practices as outlined by the IACUC (The Institutional Animal Care and Use Committee: American Association for Laboratory Animal Science, Memphis) and AVMSA (American Veterinary Medical Association). Each bird was first rendered unconscious and manual cervical dislocation was then performed. After sacrifice, tissues were collected by dissection. Mixed-breed canaries were obtained by crossing pure red factor and yellow birds. These animals were maintained and sacrificed as described above.

### Whole genome sequencing

Genome-wide polymorphism data and allele frequencies were obtained by means of whole genome sequencing of pools of individuals (summarized in Table S1). We sequenced four canary breeds (Lipochrome red, Black red, White recessive, and Gibber italicus), one wild population of canaries, and the red siskin. Genomic DNA was prepared from blood and combined in equimolar concentrations before library construction. The number of individuals incorporated in each pool varied from 12-39. Paired-end sequencing libraries for Illumina sequencing were then generated using the TruSeq DNA PCR-free Library Preparation Kit (Illumina, San Diego, CA) according to manufacturer protocols and sequenced using an Illumina HiSeq1500 with 2 x 100 bp reads. Whole genome sequencing data have been deposited in GenBank under the bioproject SRP065487.

Low quality bases and adaptor sequences were trimmed with the program Trimmomatic version 0.32 [S3] using the following parameter settings: *TRAILING* = 15, *SLIDINGWINDOW* = 4:20, and *MINLEN* = 30. Sequencing reads were then mapped to the canary reference genome assembly (SCA1; GenBank assembly accession: GCA\_000534875.1) [S4] with BWA-MEM [S5] using default settings, followed by local realignment performed using GATK [S6].

### Genetic differentiation and introgression statistics

Allele frequency differentiation was summarized across the genome using the fixation index ( $F_{ST}$ ) by means of the PoPoolation2 package [S7], which is designed to take into account sequencing errors and allele sampling biases associated with pool sequencing. Unequal sample sizes across positions associated with stochastic differences in depth of coverage are known to bias  $F_{ST}$  estimates. To correct for this effect, we used an unbiased estimator of  $F_{ST}$  [S8] as implemented in the PoPoolation2 package. Genetic differentiation across the genome was summarized using a sliding-window approach. Briefly, we averaged  $F_{ST}$  values over 20 kb windows with a 10 kb step across each scaffold, and at each new position,  $F_{ST}$  was calculated. We discarded indels, a 5 bp region surrounding each indel

site, and reads with mapping qualities < 20. We required a minimum coverage of 12X per position, a maximum coverage of three times the average coverage per pool, at least three reads supporting the minor allele in polymorphic sites, and a per base Phred quality score of 20 or higher. Windows with  $\leq 30\%$  of positions passing the quality filters described above were not considered. Final estimates were obtained by averaging values for all pairwise comparisons involving one breed carrying the red factor and breeds or wild birds not carrying the red factor. We defined candidate genomic regions under selection in red canaries as regions with window-based  $F_{ST}$  estimates above the 99.9<sup>th</sup> percentile of the empirical distribution and delimited their borders by extending them to each side as long as consecutive windows fell above 99<sup>th</sup> percentile. Three additional window sizes were attempted (10, 50, and 100 kb) but the results remained qualitatively unchanged; the top 50 windows with highest  $F_{ST}$  map to the same two genomic regions that appear as the strongest outliers in the original analysis (see Results).

To further explore consistent associations of individual alleles and red coloration in both breeds carrying the red factor we used the Cochran-Mantel-Haenszel (CMH) test [S9], also implemented in PoPoolation2. This test is a repeated test of independence, and in order not to violate the assumptions of the method, three replicates were performed in which each population was used only once:

- 1) red factor breed 1 vs yellow breed 1 + red factor breed 2 vs yellow breed 2;
- 2) red factor breed 1 vs yellow breed 2 + red factor breed 2 vs yellow breed 3;
- 3) red factor breed 1 vs yellow breed 3 + red factor breed 2 vs yellow breed 1.

$P$ -values for each position were then averaged across the different comparisons. Filtering criteria for polymorphic positions were similar to those described above for the  $F_{ST}$  analysis.

We utilized two summary statistics to identify regions in the red canary genome of potential red siskin origin: the fraction of allelic information shared through introgression ( $f^*d$ ) [S10] and the relative node depth (RND) [S11]. Prior to applying these statistics, we performed SNP calling using SAMtools v. 0.1.19-44428cd [S12] and VarScan2 [S13]. We used the following filtering criteria: 1) a minimum depth of 8 reads; 2) the allele with lower frequency was observed in at least four independent reads at a frequency  $\geq 0.01$ ; 3) a minimum Phred base quality of 20 at a position to count a read, and 4)  $p$ -value  $\leq 0.01$  for calling variants. Indels or SNPs located within 5 bp of an indel site were removed from the analysis. Allele frequencies for subsequent analysis were obtained from allele counts.

$f^*d$  is a modified version of a statistic originally developed to estimate the genome-wide fraction of admixture [S14]. This statistic considers three populations and one outgroup (O) with the following phylogenetic pattern (((P1, P2), P3), O). Briefly, it focuses on the relative abundance of two patterns referred to as ABBA and BABA, where A is the ancestral state and B the derived state. Under a neutral coalescent model with no gene flow, both ABBA and BABA sites are expected to be equally frequent but an excess of either ABBA or BABA patterns is expected when gene flow has occurred after the split of the two target populations (P1 and P2). This statistic quantifies introgression by comparing the observed difference in the number of ABBA and BABA patterns to that expected from a complete replacement of native alleles by introgressed ones, taking into account the fact that gene flow may occur in both directions. In our case, P1 stands for the non-red canaries, P2 stands for red canaries, and P3 for red siskin. We polarized the ancestral state of mutations using as outgroup the large ground finch (*Geospiza magnirostris*). To identify homologous regions of the genome between canaries and large ground finch, we aligned the canary [S4], zebra finch [S15], and large ground finch [S16] genomes using Mauve Genome Alignment Software [S17]. In order to weight each segregating site according to its fit to the ABBA or BABA pattern, we computed the statistic using allele frequencies instead of binary counts of fixed ABBA-BABA sites. Finally,  $f^*d$  values were summarized in non-overlapping windows of 100 SNPs for a median window size of 6,359 bp. Windows containing less than 50 variants, which can occur at the 3' end of scaffolds or when scaffolds are very small in size, were excluded from the analysis.

RND measures the relative divergence between populations across the genome while controlling for mutation rate variation. Values were calculated by dividing  $D_{xy}$  (the average pairwise differences between two groups) between red canaries and yellow canaries, divided by  $D_{xy}$  between red canaries and red siskin.  $D_{xy}$  values were calculated from minor allele frequency values for each SNP obtained from read counts. To avoid variation in the average genetic constraint among windows, we restricted the analysis to intronic and intergenic positions, since both are expected to be enriched for neutral evolving sites. We required 8 reads and a Phred quality score  $\geq 20$  both for polymorphic and monomorphic positions; otherwise, that position was coded as missing data. RND values were summarized in non-overlapping windows of 10,000 positions passing the quality filters (median size = 21,358 bp). As before, at the end of contigs or in contigs smaller than 10,000 bp we required at least 5,000 positions. The long divergence branch between canaries and red siskin when introgressed in red canaries should yield RND values well

above 1, while for non-introgressed regions of the genome, RND values are expected to lie between 0 and 1. For both introgression statistics, we searched for the most extreme regions as regions with estimates above the 99.9th percentile of the empirical distribution, and we delimited their borders by extending them to each side as long as consecutive windows fell above the 99<sup>th</sup> percentile.

### SNP genotyping

We genotyped 114 SNPs located in three independent genomic regions identified as outliers using differentiation and introgression statistics. These SNPs were chosen from the whole genome sequencing data and required to be fixed for diagnostic alleles between common breeds of canaries and red siskin. Given the nucleotide divergence between the canary and red siskin, regions surrounding each SNP were manually checked in order to avoid designing primers overlapping non-target SNPs or indels that could affect PCR efficiency. We genotyped 49 individuals from four breeds carrying the red factor (Lipochrome red,  $n = 35$ ; Black red,  $n = 3$ ; Portuguese Harlequin,  $n = 7$ ; Agata red,  $n = 4$ ), and 15 individuals from three breeds of yellow canary (Lipochrome yellow,  $n = 4$ ; Harz roller,  $n = 4$ ; Gibber italicus,  $n = 3$ ) and one wild population ( $n = 4$ ).

SNP genotyping was carried out using Sequenom's (San Diego, USA) iPLEX technology and detected on a Sequenom MassArray K2 platform available at the Instituto Gulbenkian de Ciência (Lisbon, Portugal). Sequenom's MassARRAY® Assay Design 3.0 software was used to design primers and establish the multiplex conditions. The resulting spectra and plots were manually inspected and software genotype calls were corrected whenever required. Individual genotypes are available online from the Dryad Digital Repository ([dx.doi.org/10.5061/dryad.sm12c](https://doi.org/10.5061/dryad.sm12c)).

### SNP/Indel functional annotation and detection of structural rearrangements

To functionally annotate SNP and indel variants we used the genetic variant annotation and effect prediction toolbox SnpEff [S18]. We focused strictly on SNP/indel variation that was found to be highly differentiated between common canaries and red birds (allele frequency difference  $\geq 0.80$ ), and each variant was manually inspected. We did not require full fixation of alleles due to potential sequencing errors. We focused specifically on variants of potential functional significance, such as nonsynonymous, frame-shift, STOP, and splice site mutations. Since the canary genome is not as comprehensively annotated as other genomes and contains many gaps, we further aligned homologous transcripts between common canaries and red factor canaries reconstructed from our RNA-seq data (see below). We then searched for potential additional variants of functional significance that may have been missed by the previous approach.

Within our candidate regions, we also screened for various types of structural variants, including deletions, insertions, duplications, inversions and translocations. The identification of structural variants was performed using three approaches that utilize different aspects of sequence data. First, we used Breakdancer [S19] with default parameters; a method that relies on paired-end alignment information (read pair orientation and insert size). Second, we used DELLY [S20] that also relies on paired-end alignment information but also on split-read alignments. We required a minimum Phred mapping quality of 10. Finally, we used LUMPY [S21] with default parameters, which utilizes a combination of multiple signals including paired-end alignment, split-read alignment, and read-depth information. Candidate structural variants were manually inspected and in one case evaluated by PCR followed by Sanger sequencing. Candidate variants were then intersected both with the canary bird genome annotation and with the additional genes detected using our RNA-seq data that mapped to the candidate regions.

### Impact of amino acid substitutions on protein function

The impact of amino acid substitutions on the function of proteins was inferred using the Aligned Sequences tool implemented in SIFT v1.03 [S22] available at the webserver: <http://sift.jcvi.org> [S23]. This method assumes that amino acid changes in positions conserved across an alignment of orthologous protein sequences will more likely affect protein function. Orthologs from 31-48 bird species for all seven candidate genes with amino acid changes between red and common yellow canaries (*HOOK1*, *CYP2J40*, *CYP2J19*, *EDYM1*, *EDMTF2*, *EDMPN-L* and *EDMTF2*) were identified within whole-genome shotgun contigs retrieved from NCBI's BLAST database. For the EDC region, we found that the syntenic location of genes as reported by Strasser et al. [S24] for *Gallus gallus* was highly conserved across all avian species. Having syntenic positions of individual EDC genes significantly facilitated the identification of individual genes. The absence of paralogs also contributed to positive gene identification. *EDMPN-L* was missing in *Gallus gallus* but was found in the majority of other species and was highly conserved. Orthologs with premature STOP codons were removed from the analysis. After aligning

translated protein sequences for each gene with ClustalW v2.0 [S25], followed by manual inspection of the alignments, we predicted the impact of the 35 nonsynonymous mutations between red and yellow birds found within candidate regions (see Results and Supplementary Table 6). For *EDMTFH*, we restricted the analysis to five passerine birds in addition to canaries, since the alignment was not satisfactory on deeper phylogenetic scales. We polarized the ancestral-derived residue state from the most closely related species to canaries and red siskin. Functional changes were assumed for normalized probabilities of tolerated change  $\leq 0.05$ .

### **RNA-seq analysis of canary skin and liver**

We generated RNA-seq data both for liver and skin of yellow (n = 3) and red factor canaries (n = 5) (Table S4). RNA-seq data have been deposited in GenBank under the bioproject SRP065487. To induce feather regeneration, we plucked small patches of feathers on the ventral side of the animal. We then allowed feathers to partially regrow for ~10 days prior to skin excision. The tissues were snap-frozen in liquid nitrogen immediately after dissection. RNA was extracted using Allprep RNeasy Mini kit (Qiagen Sciences Inc, Germantown, MD) and cDNA and Illumina libraries were prepared using the TruSeq RNA Sample Prep Kit v2 (Illumina) following standard procedures. RNA quality prior to sequencing was assessed using a BioAnalyzer 2200 system (Agilent, Foster City, CA). Libraries were then sequenced using 100 bp paired-end reads on an Illumina HiSeq1500.

### **Differential gene expression analysis in common and red factor canaries**

Reads were subjected to several quality controls prior to use, including trimming TruSeq Illumina adapters using Cutadapt version 1.7.1 [S26] and quality filtering with Trimmomatic version 0.32 using a 4 bp trimming sliding window approach establishing a 15 Phred quality score threshold (*SLIDINGWINDOW* = 4:15). Also using Trimmomatic, the first 13 bp bases of every read were cut (*LEADING* =13) due to a bias observed in nucleotide composition that is possibly due to the non-random hexamer primer activity of the primers used during library preparation [S27]. Finally, all reads  $\leq 30$  bp long (*MINLEN* = 30) were excluded along with their mate. Between each step, sample quality was assessed with FastQC version 0.10.1 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>).

Prior to gene expression analysis, reads were mapped to the canary genome reference annotation using Bowtie version 1.0.0 [S28] as implemented in RSEM version 1.2.19 [S29]. *RSEM* was run with the default parameters whereas Bowtie was used allocating 512 Mb of memory for the calculation of the best first alignment (--bowtie-chunkmbs). The relative abundances of genes were used in the subsequent analysis.

Differential expression analysis comparing liver and skin transcriptomes from yellow and red individuals was performed separately using edgeR version 3.10.2 [S30] in RStudio version 0.98.976 using R version 3.2.2 (R Core Team, 2015). The analysis consisted of an initial filtering step that excluded genes with  $<1$  count per million (CPM) mapped reads averaged across all biological replicates. Then, data normalization across libraries was performed using the trimmed mean of M-Values (TMM) method [S31] as implemented in edgeR. Next, the common and tagwise dispersion were calculated using a Cox-Reid adjusted profile likelihood. Finally, a Generalized Linear Model (GLM) was applied to accommodate the experimental design, linking the response variable (count data) to the predictor (color phenotype) with a log-linear model and modeling read counts with a negative binomial (NB) distribution [S32]. A likelihood ratio test was used to test for differential expression in the comparison made ("Yellow" vs. "Red") and a Benjamin-Hochberg multiple test correction was applied. A false discovery rate (FDR) of 0.1% was used as the significance threshold.

### **De novo transcriptome assembly and annotation**

Since the canary genome is not as comprehensively annotated as other available genomes and contains many gaps, to characterize the candidate genes' isoforms and improve the genome annotation, a *de novo* assembly for each color phenotype transcriptome was performed. RNA-seq reads were treated as described above. Two transcriptomes were then generated for the liver and skin samples of each group of red and yellow individuals using Trinity version 2.0.2 [S33]. The quality of each assembly was then assessed using TrinityStats.pl provided by the Trinity package.

In order to search for new genes and/or isoforms differing between common canaries and red factor canaries, which may have been missed by the annotation associated with the release of the canary reference genome, the resulting assemblies were mapped against the canary reference genome using GMAP [S34] with default parameters and Blastn as implemented in Blast+ version 2.2.29 [S35] requiring a maximum e-value of  $1e^{-20}$  and only reporting the best hit for each contig. We also reciprocally blasted the transcripts of both transcriptomes to look for

potential paralogs. Based on the RNA-seq data, we noticed that the annotation in scaffold NW\_007931203 in the canary genome was incomplete. Therefore, we manually curated the transcripts mapping to this region using as a template the detailed analysis of the epidermal differentiation complex (EDC) previously described for chickens [S24]. To identify putative genes, we combined blast searches with phylogenetic and syntenic information between both the canary and chicken genomes.

### **Sample collection from developing chicken retinas**

To examine the time course of astaxanthin accumulation and *CYP2J19* expression in the developing chicken retina, we acquired pathogen-free fertilized eggs from Charles River Laboratories (North Franklin, CT, USA). We incubated the eggs and reared chicks to specific developmental stages on a carotenoid-rich commercial diet (Start & Grow, Purina Mills, St. Louis, MO). We collected retinas from the developing embryos at 16, 18, 20 days of incubation and chicks at hatch day (P0), as well as 7, 14, and 21 days post-hatch. We euthanized the birds following an approved protocol (Washington University Animal Studies Committee protocol #20140072) with carbon dioxide asphyxiation and collected three biological samples of retina tissue for each analysis.

### **High performance liquid chromatography analysis of ketocarotenoids in the developing chicken retina**

Immediately following harvest, retinas were frozen at -80°C and stored for up to six months. For analysis, we thawed the retinas on ice, resuspended them in 500 µl of phosphate buffered saline (PBS), and homogenized them by brief sonication (10 sec and 20% amplitude, VC 505, Sonics & Materials, Inc. Newtown, CT). We determined the protein concentration in the homogenate with a bicinchoninic acid assay (cat. no. 23250, Thermo Scientific, Rockford, IL, USA). We extracted the ketocarotenoids from 500 µl of the homogenate three times with 500 µl of hexane:tert-butyl methyl ether 1:1 (vol:vol), combined the extracts from each sample, and dried them under nitrogen. We saponified the extracts with 0.02 M NaOH at room temperature, in the dark for four hours, following a published protocol [S36]. We extracted the carotenoids from the saponification solution with 2 ml of hexane:tert-butyl methyl ether 1:1 (vol:vol) and dried it under nitrogen. We resuspended this material in 200 µl of methanol:acetonitrile 1:1 (vol:vol) and injected 100 µl into an Agilent 1100 series HPLC equipped with a YMC carotenoid 5.0 µm column (4.6 mm × 250 mm, YMC Inc. Allentown, PA). The ketocarotenoids were eluted with a gradient mobile phase consisting of acetonitrile:methanol:dichloromethane (44:44:12) (vol:vol:vol) for 11 minutes, ramp up to acetonitrile:methanol:dichloromethane (35:35:30) from 11-21 minutes, then isocratic through 30 minutes. The column was heated to 30°C and the flow rate was 1.2 ml/min throughout the run. We monitored the samples at 480 nm and identified and quantified astaxanthin by comparison to an astaxanthin standard that was a gift of DSM Nutritional Products (Parsippany, NJ). Carotenoid levels were normalized to total protein content in each sample.

### **Quantitative real-time PCR analyses of gene expression in canary and chicken tissues**

Canary cDNA from skin, liver, and retina was generated as described above. Chicken retinas were harvested and placed directly in Trizol reagent (Life Technologies, Carlsbad, CA), frozen at -80°C and stored for up to six months prior to RNA extraction and analysis. We thawed the retina samples on ice and extracted RNA following the manufacturer's protocol (Trizol reagent, Life Technologies, Carlsbad, CA). We generated cDNA by reverse transcribing ~1 µg of RNA using Superscript IV reverse transcriptase (Life Technologies, Carlsbad, CA) following the manufacturer's protocols.

We designed the following primers to flank an exon/exon junction of the *CYP2J19*, *CYP2J40*, *FGGY*, *EDMTFH* and *GAPDH* genes in canary ('Sc') and *CYP2J19*, and *GAPDH* in chicken ('Gg'), respectively:

Sc_Cyp2J19_1F	AGTCCGAGGCAGGAGATTAT
Sc_Cyp2J19_1R	CGTTGTCTCTGTCCCAGTTAAG
Sc_Cyp2J40_2F	GAAGTGCAGAGGAAAGGCAA
Sc_Cyp2J40_2R	ATAGTGCCCTTTGGGATGTAG



Sc_FGGY_2F	TCTGGATGAACTTGCCCTTATC
Sc_FGGY_2R	GAACAGCGTGGTGATCTCAT
Sc_EDMTFH_2F	GAAGCTCTCGGGAAAGATCG
Sc_EDMTFH_2R	CAAACCTGAACGGTGAAGTTGAA
Sc_GAPDH_1F	TAGCCATTCTCCACCTTTG
Sc_GAPDH_1R	ACTCGGTTGCTGTATCCATATT
Gg_Cyp2J19_qPCR_F	CCTGAACACTTCCTGGAGAA
Gg_Cyp2J19_qPCR_R	CAGGGCAGTGAAGAAGATGAA
Gg_GAPDH_qPCR_F	GAGGGTAGTGAAGGCTGCTG
Gg_GAPDH_qPCR_R	TGGCTGTCACCATTGAAGTC

We determined primer efficiency with a 1:10 dilution series of cDNA performed in triplicate. Primers had efficiencies ranging from 87%-105% at the cycle threshold ( $C_t$ ) levels observed in the subsequent experiments. Each biological replicate, at each time point, was assayed in triplicate using Sybr® Green PCR master mix (Life Technologies, Carlsbad, CA) on an Applied Biosystems StepOne real time PCR system (Carlsbad, CA). We normalized the expression of all genes to *GAPDH* expression and compared expression of each gene between yellow and red canaries with two-sample *t*-tests.

### ***In situ* hybridization**

We cloned probe templates for *CYP2J19*, *CYP2J40* and *EDMTFH* from red canary skin cDNA and for *CYP2J19* from chicken retinal cDNA with the following primers for canary ('Sc') and chicken ('Gg'), respectively:

Sc_Cyp2J19_ISH_F	ATGCAGAGAACAGTACAA
Sc_Cyp2J19_ISH_R	CGGATGGCGGTCGCC
Sc_Cyp2J40_ISH_F	CCTGCTCATCGCAGACTACA
Sc_Cyp2J40_ISH_R	TCCCATCCTTCAGAAAATGC
Sc_EDMTFH_ISH_F	ATGAATTCCTTCAAGGATTGCTG
Sc_EDMTFH_ISH_R	TCACTGGGCTCTGTCCC
Gg_Cyp2J19_ISH_F	ATGGATTTTCGCTTTTGGCCC
Gg_Cyp2J19_ISH_R	GCAGCGAGGCAGGGC

We subcloned the PCR products into the BlueScript vector pBSK+ at the EcoRV restriction site and confirmed the sequences by Sanger sequencing. We used PCR to generate probe templates and followed established methods to generate digoxigenin-labeled probes from these templates [S37, S38]. All tissue was fixed overnight in 4% paraformaldehyde in 1x PBS and then embedded in Tissue-Tek OCT compound (Sakura, Torrance, CA). *In situ* hybridization was carried out on 12  $\mu$ m horizontal sections as previously described [S37]. Stained sections were imaged with an Olympus BX-51 microscope.

### Quantification of allelic imbalance in red x yellow F1 canaries by sequencing

To quantify the relative expression of red and yellow alleles of *CYP2J19*, *CYP2J40*, *FGGY* and *EDMTFH* in red x yellow F1 canaries, we harvested skin and liver from three individuals, extracted RNA, and generated cDNA as described above. We designed primers to amplify a single 302-399 bp amplicon from each transcript, each containing 4 to 14 SNPs between the yellow and red alleles:

CYP2J19_F	TTTCCAGGAACTGCTGCAC
CYP2J19_R	CTCCTGCTGCACTTTCTCCT
CYP2J40_F	AGCCGGGACTACATTGACAG
CYP2J40_R	CGCGTTGAAAGGGATAACAT
FGGY_F	TGGCAGTCAACTCTGAAGGA
FGGY_R	CAGGAGACAGGACGTGGTTT
EDMTFH_F	AGGCCAGCAATGACAACTTT
EDMTFH_R	GCTCTGTCCCCAACTGAAC

We amplified these fragments from cDNA using Phusion polymerase (NEB, Ipswich, MA) (35 cycles). We then normalized input DNA from each PCR amplicon to 67 ng and performed 3' adenylation, adapter ligation, and amplification (18 cycles) following published protocols [S37]. We pooled indexed library fragments and sequenced them on a MiSeq (2 × 250-bp reads, Illumina, San Diego, CA) at a loading concentration of 10 pM, yielding from 5444 to 236894 reads per amplicon. We aligned paired-end reads to the red allele of each gene using Bowtie 2 v2.1.0 [S37, S39]. For each transcript, we then counted the number of reads corresponding to red and yellow alleles to calculate the proportion of red alleles expressed in each individual and tissue. Finally, we performed a two-tailed one sample t-test (null hypothesis  $\mu = 0.5$ ) to test each transcript for biased expression in each tissue.

## SUPPLEMENTAL REFERENCES

- S1. Toomey, M.B., Collins, A.M., Frederiksen, R., Cornwall, M.C., Timlin, J.A., and Corbo, J.C. (2015). A complex carotenoid palette tunes avian colour vision. *J. Roy. Soc. Interface* *12*:20150563.
- S2. Birkhead, T. (2003). *A brand-new bird: how two amateur scientists created the first genetically engineered animal*, (New York: Basic Books New York).
- S3. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* *30*, 2114-2120.
- S4. Frankl-Vilches, C., Kuhl, H., Werber, M., Klages, S., Kerick, M., Bakker, A., de Oliveira, E.H., Reusch, C., Capuano, F., Vowinckel, J., et al. (2015). Using the canary genome to decipher the evolution of hormone-sensitive gene regulation in seasonal singing birds. *Genome Biol.* *16*, 19.
- S5. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* *25*, 1754-1760.
- S6. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* *20*, 1297-1303.
- S7. Kofler, R., Pandey, R.V., and Schlotterer, C. (2011). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* *27*, 3435-3436.
- S8. Karlsson, E.K., Baranowska, I., Wade, C.M., Salmon Hillbertz, N.H.C., Zody, M.C., Anderson, N., Biagi, T.M., Patterson, N., Pielberg, G.R., Kulbokas, E.J., et al. (2007). Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat. Genet.* *39*, 1321-1328.
- S9. McDonald, J.H. (2009). *Handbook of Biological Statistics* (2nd ed.). (Baltimore, Maryland: Sparky House Publishing).
- S10. Martin, S.H., Davey, J.W., and Jiggins, C.D. (2015). Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* *32*, 244-257.
- S11. Feder, J.L., Xie, X.F., Rull, J., Velez, S., Forbes, A., Leung, B., Dambroski, H., Filchak, K.E., and Aluja, M. (2005). Mayr, Dobzhansky, and Bush and the complexities of sympatric speciation in *Rhagoletis*. *PNAS* *102*, 6573-6580.
- S12. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* *25*, 2078-2079.
- S13. Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* *22*, 568-576.
- S14. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H., et al. (2010). A draft sequence of the Neandertal genome. *Science* *328*, 710-722.
- S15. Warren, W.C., Clayton, D.F., Ellegren, H., Arnold, A.P., Hillier, L.W., Kunstner, A., Searle, S., White, S., Vilella, A.J., Fairley, S., et al. (2010). The genome of a songbird. *Nature* *464*, 757-762.
- S16. Rands, C.M., Darling, A., Fujita, M., Kong, L., Webster, M.T., Clabaut, C., Emes, R.D., Heger, A., Meader, S., Hawkins, M.B., et al. (2013). Insights into the evolution of Darwin's finches from comparative analysis of the *Geospiza magnirostris* genome sequence. *BMC Genomics* *14*, 95.
- S17. Darling, A.C., Mau, B., Blattner, F.R., and Perna, N.T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* *14*, 1394-1403.
- S18. Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* *6*, 80-92.
- S19. Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* *6*, 677-681.
- S20. Rausch, T., Zichner, T., Schlattl, A., Stutz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* *28*, i333-i339.
- S21. Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* *15*, R84.
- S22. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* *4*, 1073-1081.
- S23. Sim, N.L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P.C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* *40*, W452-457.

- S24. Strasser, B., Mlitz, V., Hermann, M., Rice, R.H., Eigenheer, R.A., Alibardi, L., Tschachler, E., and Eckhart, L. (2014). Evolutionary origin and diversification of epidermal barrier proteins in amniotes. *Mol. Biol. Evol.* *31*, 3194-3205.
- S25. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* *23*, 2947-2948.
- S26. Martin, M.B. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* *17*.
- S27. Hansen, K.D., Brenner, S.E., and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* *38*, e131.
- S28. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* *10*, R25.
- S29. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* *12*, 323.
- S30. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* *26*, 139-140.
- S31. Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* *11*, R25.
- S32. McCarthy, D.J., Chen, Y., and Smyth, G.K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* *40*, 4288-4297.
- S33. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* *29*, 644-652.
- S34. Wu, T.D., and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* *21*, 1859-1875.
- S35. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* *10*, 421.
- S36. Toomey, M.B., and McGraw, K.J. (2007). Modified saponification and HPLC methods for analyzing carotenoids from the retina of quail: implications for its use as a nonprimate model species. *Invest. Ophthalm. Vis. Sci.* *48*, 3976-3982.
- S37. Enright, J.M., Lawrence, K.A., Hadzic, T., and Corbo, J.C. (2015). Transcriptome profiling of developing photoreceptor subtypes reveals candidate genes involved in avian photoreceptor diversification. *J. Comp. Neurol.* *523*, 649-668.
- S38. Trimarchi, J.M., Stadler, M.B., Roska, B., Billings, N., Sun, B., Bartch, B., and Cepko, C.L. (2007). Molecular heterogeneity of developing retinal ganglion and amacrine cells revealed through single cell gene expression profiling. *J. Comp. Neurol.* *502*, 1047-1065.
- S39. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* *9*, 357-359.