PROCEEDINGS B

rspb.royalsocietypublishing.org

Research

Cite this article: Toomey MB *et al*. 2018 A non-coding region near *Follistatin* controls head colour polymorphism in the Gouldian finch. *Proc. R. Soc. B* **285**: 20181788. http://dx.doi.org/10.1098/rspb.2018.1788

Received: 7 August 2018
Accepted: 5 September 2018

Subject Category:
Genetics and genomics

Subject Areas:
evolution, genetics, genomics

Keywords:
sympatric colour morphs, pleiotropy, balancing selection, *cis*-regulatory variation, avian coloration

Author for correspondence:
Miguel Carneiro
e-mail: miguel.carneiro@cibio.up.pt

†Co-first authors.

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.4233548.

THE ROYAL SOCIETY PUBLISHING

A non-coding region near *Follistatin* controls head colour polymorphism in the Gouldian finch

Matthew B. Toomey[1,†], Cristiana I. Marques[2,3,†], Pedro Andrade[2,3], Pedro M. Araújo[2,4], Stephen Sabatino[2], Małgorzata A. Gazda[2,3], Sandra Afonso[2], Ricardo J. Lopes[2], Joseph C. Corbo[1] and Miguel Carneiro[2,3]

[1]Department of Pathology and Immunology, Washington University School of Medicine, St Louis, MO 63110, USA
[2]CIBIO/InBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, 4485-661 Vairão, Portugal
[3]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, 4169-007 Porto, Portugal
[4]Centro de Ciências do Mar e do Ambiente, Departamento de Ciências da Vida, Universidade de Coimbra, 3004-517 Coimbra, Portugal

MBT, 0000-0001-9184-197X; PA, 0000-0003-2540-2471; RJL, 0000-0003-2193-5107; MC, 0000-0001-9882-7775

Discrete colour morphs coexisting within a single population are common in nature. In a broad range of organisms, sympatric colour morphs often display major differences in other traits, including morphology, physiology or behaviour. Despite the repeated occurrence of this phenomenon, our understanding of the genetics that underlie multi-trait differences and the factors that promote the long-term maintenance of phenotypic variability within a freely interbreeding population are incomplete. Here, we investigated the genetic basis of red and black head colour in the Gouldian finch (*Erythrura gouldiae*), a classic polymorphic system in which naturally occurring colour morphs also display differences in aggressivity and reproductive success. We show that the candidate locus is a small (approx. 70 kb) non-coding region mapping to the Z chromosome near the *Follistatin* (*FST*) gene. Unlike recent findings in other systems where phenotypic morphs are explained by large inversions containing hundreds of genes (so-called supergenes), we did not identify any structural rearrangements between the two haplotypes using linked-read sequencing technology. Nucleotide divergence between the red and black alleles was high when compared to the remainder of the Z chromosome, consistent with their maintenance as balanced polymorphisms over several million years. Our results illustrate how pleiotropic phenotypes can arise from simple genetic variation, probably regulatory in nature.

1. Introduction

Colour polymorphic species, where two or more genetically determined colour morphs occur within a single population, represent an ideal model system for investigating the processes that maintain genetic variation and drive speciation [1–4]. Often colour polymorphisms are linked to behavioural and physiological traits and have evolved and persist through sexual and natural selection acting on these traits in concert [1]. At the genomic level, such phenotypic correlations can arise through large-scale structural variation that brings coloration genes into linkage disequilibrium with genes mediating behaviour and physiology. For example, the white-throated sparrow (*Zonotrichia albicollis*) has two distinct colour morphs, white and tan, that use divergent reproductive strategies [5]. Underlying this polymorphism is a supergene, a greater than 100 Mb genomic inversion, that contains a suite of neuroendocrine and coloration genes linked via very low levels of recombination [6]. Similarly, the colour polymorphic and

© 2018 The Author(s) Published by the Royal Society. All rights reserved.

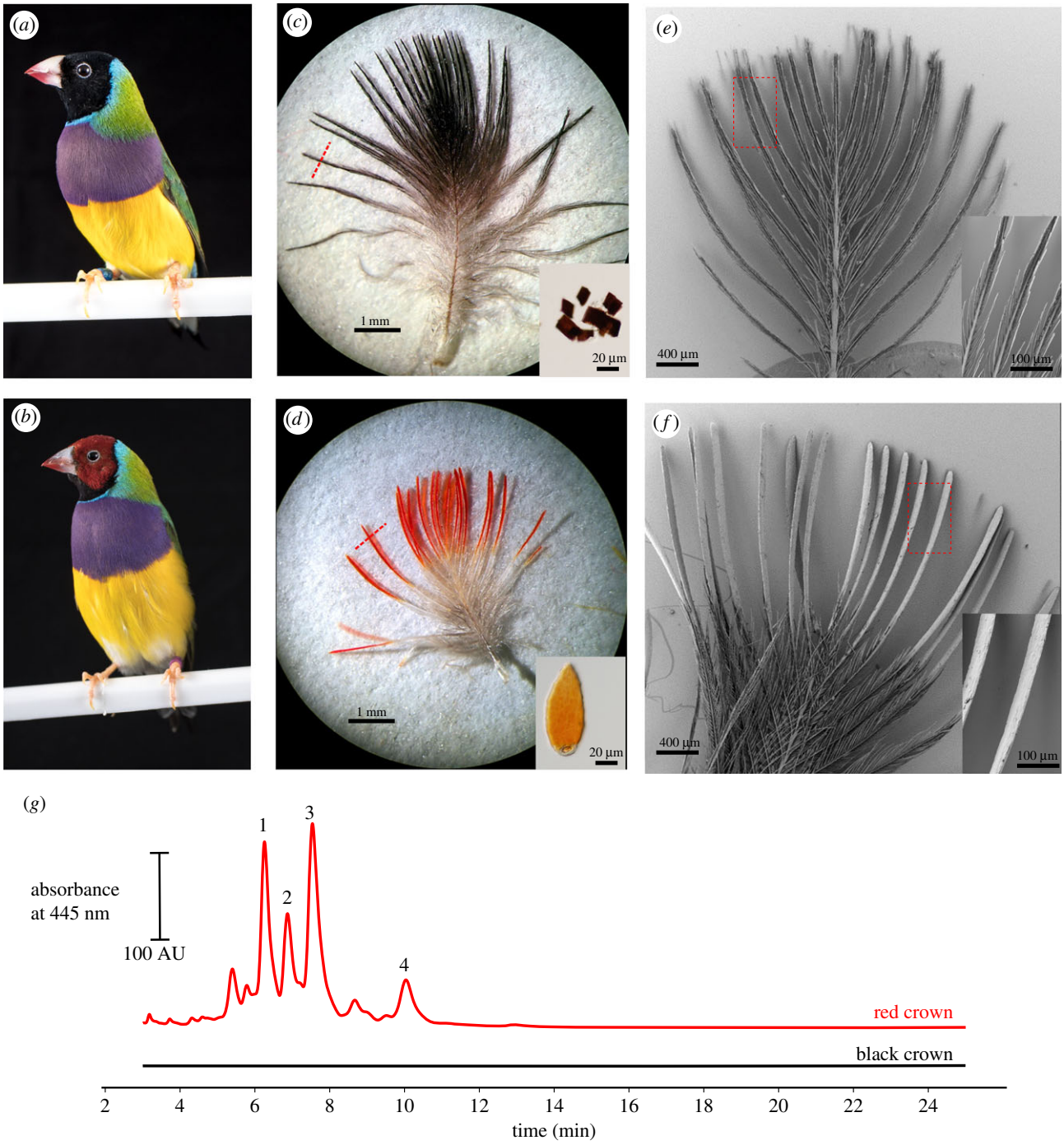

Figure 1. The head feathers of the black and red Gouldian finch morphs have distinct coloration, morphology and pigment content. Representative images of the black (*a*) and red (*b*) head-colour morphs of the Gouldian finch. (*c*,*d*) Light microscopy imaging of the individual head feathers of the black (*c*) and red (*d*) morphs. Inset within each image is a 12 µm cross-section through a single barb (approximate location in feather indicated by the dashed red line) imaged at 400× magnification. (*e*,*f*) Scanning electron micrographs of individual head feathers of the black (*e*) and red (*f*) morphs imaged at 23× magnification. Inset within each image is a 200× magnification detail of a region comparable to that indicated with the dashed red box. (*g*) Representative high performance liquid chromatography chromatograms of carotenoid extracts from head feathers of the black (black trace) and red (red trace) morphs. Carotenoids were not detected in the black morph. In the red morph, we observed high concentrations of what we have putatively identified as 1 and 2—papilioerythrinone isomers, 3—astaxanthin and 4—adonirubin. Full spectra of these individual peaks are presented in the electronic supplementary material, figure S1.

alternative-mating morphs of the lek-mating ruff (*Philomachus pugnax*) are controlled by a large genomic inversion [7,8]. Such large-scale variation in genomic architecture is not the only mechanism that can generate correlated phenotypes [1]. Simple genetic variation and differential gene expression within gene regulatory and endocrine pathways can also produce pleiotropic phenotypes. Such systems offer important opportunities for understanding the regulatory mechanisms that generate diversity in coloration, physiology and behaviour.

The Gouldian finch (*Erythrura gouldiae*) is a classic polymorphic system with an apparently simple genetic basis. The Gouldian finch is endemic to Australia and throughout its range two common colour morphs, black and red head colour, coexist at a ratio of approximately 7 : 3 (figure 1*a*,*b*) [9]. A third orange morph occurs at very low frequency (less than 0.1%). These head-colour morphs also have distinct physiological and behavioural phenotypes. Red morphs tend to be aggressive and dominant over black morph individuals in agonistic

encounters and have a heightened physiological response to social stress [10–12]. Red morphs substantially increase circulating sex steroid (testosterone) and stress hormone (corticosterone) levels in competitive social environments, while black morphs show no such response [11]. Within captive populations, Gouldian finches mate assortatively by head-colour morph, and some studies suggest the existence of incompatibilities between morphs that result in increased mortality of female offspring [13–16]. However, a recent study did not detect evidence of genetic incompatibilities in wild populations of the Gouldian finch [17]. The suite of morphological, behavioural and physiological traits associated with head-colour morph are determined by a sex-linked Mendelian locus where red (Z^R) is the dominant allele and black (Z^r) is recessive [18]. The orange head-colour morph is determined by an autosomal locus. Orange individuals are thus identical to red individuals at the *Red* locus, but express a different complement of carotenoids in their masks [19]. Recently, the *Red* locus was mapped to an approximately 7.2 cM interval on the Z chromosome using microsatellite markers [20]. However, the resolution of this mapping did not allow for the identification of the gene(s) and causal variants underlying this polymorphism.

To identify the causative locus that controls red and black head colour in the Gouldian finch, we generated a reference genome sequence for this species, resequenced the genomes of black and red morph individuals, and used a range of analytical tools to search for genotype-to-phenotype associations across the genome. We identified a single locus on the Z chromosome in a small non-coding region upstream of *Follistatin* (FST) that was perfectly associated with head-colour morph. FST is a secreted glycoprotein that is widely expressed in the body, including in skin and gonads [21]. We hypothesize that the head-colour polymorphism and the suite of associated physiological and behavioural traits are the product of regulatory variation that modulates the timing and levels of *FST* expression in key tissues controlling coloration and behaviour.

2. Material and methods

Full methods and materials are detailed in the electronic supplementary material, methods.

(a) Light and electron microscopy

We used a combination of light and electron microscopy to characterize head feather morphology. We collected full colour macroscopic images through a Leica MZ7 dissecting microscope (Leica Microsystems, Inc., Buffalo Grove, IL, USA). We also collected images of the feathers with a Zeiss Merlin field emission scanning electron microscope and a JEOL JEM-1400 transmission electron microscope at the Washington University Center for Cellular Imaging.

(b) Carotenoid analysis

We collected 10 mature feathers from the heads of red and black males and extracted carotenoids with acidified pyridine and hexane following the methods of McGraw *et al*. [22]. The carotenoid content was analysed by high performance liquid chromatography.

(c) *De novo* genome assembly and annotation

A *de novo* genome assembly of the Gouldian finch (*E. gouldiae*) was generated using Chromium linked-read data (10× Genomics, San Francisco, USA). A single Chromium library was sequenced on an Illumina HiSeqX Ten using 2 × 150 bp paired-end reads and a reference genome was assembled with Supernova (v. 2.0) [23]. The reference genome was annotated using Maker2 [24] by combining *in silico* gene prediction tools, protein evidence from curated databases and newly generated RNA-sequencing (RNA-seq) data from regenerating mask skin. A quantitative assessment of assembly and annotation completeness was performed using BUSCO (v. 3) [25], as well as whole-genome alignments between Gouldian finch and zebra finch using Last [26].

(d) Whole-genome resequencing

Black ($n = 21$) and red-headed ($n = 21$) captive birds were obtained from 13 private aviaries in Portugal (electronic supplementary material, table S1). Individual paired-end libraries were prepared following a modified version [27] of Illumina's Nextera XT protocol (Illumina, San Diego, IL, USA). The libraries were sequenced using 2 × 125 bp paired-end reads on an Illumina Hiseq 1500 (electronic supplementary material, table S1).

(e) Population genetics and association analyses

Population structure was investigated with principal component analysis (PCA). To detect genomic regions associated with head-colour polymorphism, we performed a genome-wide association analysis using individual variants, and estimated genetic differentiation in sliding windows across the genome using the fixation index (F_{ST}) and the average number of pairwise differences per site (d_{xy}). Both F_{ST} and d_{xy} analyses were performed on a subset of the individuals (13 black and 12 red; see the electronic supplementary material, methods). To take into account uncertainty in genotype calls owing to low sequencing depth, these analyses were carried out using genotype probability methods as implemented in ANGSD [28].

(f) Single nucleotide polymorphism genotyping

For confirmation of the association, we genotyped a set of seven closely located single nucleotide polymorphisms (SNPs) found to be significantly associated with the red and black alleles using the whole-genome resequencing data (primers given in the electronic supplementary material, table S2). The read data for these SNPs also supported the expected inheritance: red individuals carried one variant allele or both, while the black individuals carried the alternative allele.

(g) Haplotype analyses

We estimated the relative node depth (RND) statistic [29] by dividing nucleotide divergence between the two haplotypes derived from the Gouldian finch diploid assembly by nucleotide divergence between Gouldian finch and zebra finch. This statistic was calculated using the Perl script *calculate-dxy.pl* [30] from alignments obtained using progressiveMauve [31].

(h) Structural rearrangements

To search for structural rearrangements between the red and black alleles, we used both the linked-read and whole-genome resequencing data. We started by inspecting the alignment between the red and black haplotypes from the diploid assembly by means of a dot plot. We also searched for structural rearrangements in the candidate region by visualizing the data in IGV (Integrative Genomics Viewer) [32] and by using structural rearrangement detection software that take advantage of multiple aspects of the read data: *long ranger* toolbox from 10× Genomics, Breakdancer [33], Delly [34] and Lumpy [35].

(i) RNA-sequencing

We generated RNA-seq data from regenerating skin samples of three black and five non-black males (three orange and two red;

electronic supplementary material, table S3). Orange individuals are identical to red individuals at the *Red* locus and also express carotenoids in their masks [19]. Feather regeneration was induced by plucking feather patches from the mask region, and new feathers were allowed to regrow for 10 days prior to skin excision. Libraries were sequenced on an Illumina HiSeq 1500 with 2 × 125 bp paired-ends reads. Expression analysis was conducted using edgeR 3.4.2 [36,37].

(j) Transcript isoform characterization

We searched for transcript isoforms of the candidate genes by mapping RNA-seq reads to the EryGou1.0 genome and quantitating exon–exon spanning reads with *sashimi_plot* [38]. We confirmed the presence of alternative transcripts by Sanger sequencing of cDNA derived from regenerating skin of red and black morphs.

(k) Quantitative polymerase chain reaction

To examine the expression of *FST* and *MOCS2* during feather regeneration, we sampled the regenerating skin from the masks of three red and three black males. We plucked small patches of feathers from the mask region of these birds 2 and 4 days prior to skin excision. Testis were harvested from four red and four black 1-year-old adult males purchased from different breeders and then housed together in a single cage for 8 days to induce social competition in the group [11]. We designed primers to target the coding sequences of each isoform of *FST*, both isoforms of *MOCS2*, and *GAPDH* (electronic supplementary material, table S2).

3. Results and discussion

(a) Feathers of black and red morphs have distinct pigmentation and morphology

Consistent with an earlier report [39], we found that the head feathers of black morph Gouldian finches are pigmented with melanin, devoid of carotenoids, and have relatively thin barbs with numerous barbules along the entire length of the feather (figure 1*c*,*e*,*g*; electronic supplementary material, figure S1). By contrast, the feathers of the red morph are pigmented with the ketocarotenoids papilioerythrinone, astaxanthin and adonirubin, and have enlarged barbs that entirely lack barbules along their distal half (figure 1*d*,*f*,*g*; electronic supplementary material, figure S1). Melanin is almost entirely absent from the distal barbule-free portion of the barb but is present within more proximal portions (figure 1*d*). These differences suggest that genetic variation underlying head-colour polymorphism influences multiple aspects of feather development and regeneration, including pigment deposition and barbule morphogenesis [40].

(b) Assembly of a reference genome for the Gouldian finch

To assemble a draft genome of the Gouldian finch (EryGou1.0), we generated and sequenced a microfluidically partitioned genomic library using 10× Chromium technology. This method produces linked-read data that allows us to integrate long-range information from barcoded reads belonging to the same DNA molecules and produce locally phased assemblies, where both haplotypes in a diploid individual are recovered [23]. The Chromium library was produced from a male, heterozygous for the red and black alleles. The library was sequenced using 150 bp paired-end Illumina reads and assembled using the Supernova Assembler (v. 2.0), a software package for *de novo* assembly from Chromium linked reads [23]. The final assembly comprised 1.07 Gb divided in 831 scaffolds larger than 10 kb (N50 of 18.97 Mb), slightly lower than the predicted genome size of 1.22 Gb (electronic supplementary material, table S4).

We evaluated the quality of our assembly in two ways. First, we assessed the representation of highly conserved protein-coding genes in our genome sequence using the BUSCO software. Among the tested set of genes ($n = 4915$), we detected a high percentage of complete protein-coding sequences (94.2%; electronic supplementary material, table S4). Second, we aligned the Gouldian finch genome assembly with that of the zebra finch—the most complete and contiguous assembly of a passerine bird—to assess genome fragmentation and completeness per chromosome. Of particular interest to this study, we found that the Z chromosome was more fragmented (i.e. consisted of a larger number of scaffolds) than autosomes of similar size (e.g. 1A, 4 and 5), which is expected given the higher repeat content that characterizes sex chromosomes [41], but the completeness of the sequence for this chromosome compares well with those of the autosomes (electronic supplementary material, table S5).

The annotation of the reference genome produced 18 989 protein-coding gene models (electronic supplementary material, table S4), which is similar to the number found for the zebra finch (*Taeniopygia guttata*; $n = 19\,334$). A quantitative assessment with BUSCO and the same set of conserved genes as above ($n = 4915$) revealed that 86% of the gene models in our annotation were complete (electronic supplementary material, table S4). Incomplete (fragmented) gene models were found for an additional 7.8% (383). Overall, these results indicate that both our reference genome sequence and genome annotation are highly complete and accurate.

(c) The head-colour polymorphism maps to a small genomic interval on the Z chromosome

To map the genomic region controlling the red and black head colour, we performed whole-genome Illumina sequencing at low coverage (average = 1.63-fold) for a total of 42 captive males (electronic supplementary material, table S1). We sampled 21 males with a black head, which are expected to be homozygous for the black allele, and 21 males with a red head, which are expected to be homozygous for the red allele or heterozygous for the black and red alleles. To take into account biases and the uncertainty in genotype calling associated with low-coverage datasets, we used probabilistic methods that rely on genotype likelihoods [28,42].

Given the known effects of population structure on genetic mapping studies, we started by assessing genetic differentiation within our sample of 42 individuals using whole-genome information. However, we did not detect any signature of population structure based on head colour using PCA because black and red individuals appear intermingled (electronic supplementary material, figure S2). This result is consistent with the fact that breeders rarely keep different morphs as separate lines, but instead cross the two extensively. The lack of significant differentiation between head morphs is favourable for genetic mapping.

To search for genomic regions of elevated genetic differentiation between black and red morphs, we began by performing a genome-wide screen based on the fixation index

(F_{ST}). Using a sliding-window approach, we uncovered a single clearly differentiated genomic region on scaffold 11 that is consistent with a simple Mendelian genetic basis for this trait (figure 2*a*). This genomic region was 75 kb long (scaffold 11: 19 805 000–19 880 000 bp), contained the top 16 values of the empirical distribution of F_{ST} (top value $F_{ST} = 0.24$) and is homologous to the zebra finch Z chromosome as determined by whole-genome alignments. We performed a second window-based analysis using an absolute measure of divergence (d_{XY}) to complement the F_{ST} analysis. This analysis revealed that the region previously identified using F_{ST} was also a clear genome-wide outlier (figure 2*b*). There was another region with a slightly higher value (scaffold 326); however, this region is unlikely to be implicated in the head-colour polymorphism, because it is homologous to zebra finch chromosome 4, and not the Z chromosome as expected from the causative region containing the red and black alleles. In addition, a closer inspection of scaffold 326 suggests that the elevated d_{XY} might be a consequence of a mis-assembly of the genome, where two or more similar sequences are being collapsed into a single scaffold. For example, we found that read coverage averaged over the scaffold is approximately twice the genome-wide average and many positions across the entire length of the scaffold are heterozygous for all individuals that have five or more reads overlapping a given site.

We next performed a genome-wide association analysis based on individual variants using a likelihood ratio test (LRT) for differences in allele frequencies between the two groups [43]. We evaluated 8 101 896 SNPs and found that the top 20 most significant variants were all located in the same region uncovered by the sliding-window analysis (figure 2*c*). In fact, only two out of the 52 SNPs reaching the standard threshold for genome-wide significance ($p < 5 \times 10^{-8}$; approximately LRT > 28) were not located in this region. When we applied an even more stringent Bonferroni correction ($p \leq 6.17 \times 10^{-9}$; approximately LRT > 34), the nine variants exceeding the significance threshold were located on the aforementioned interval on scaffold 11. The 50 SNPs mentioned above defined a chromosomal segment of approximately 73 kb (scaffold 11: 19 803 083–19 876 349 bp), nearly identical to the segment defined by the F_{ST} and d_{XY} analysis (figure 2*d*). We note that scaffold 11 is the largest Z-linked sequence in our genome assembly (22.4 Mb), and this segment is well removed from the extremities of the scaffold (greater than 2.5 Mb).

Next, we selected seven of these highly significant variants contained within a small amplicon and genotyped the 42 individuals using Sanger sequencing (scaffold 11: 19 840 503–19 840 778). Consistent with a recessive mode of inheritance of the black head colour [20], all black males were homozygous ($n = 21$), whereas all red males were either heterozygous ($n = 15$) or homozygous ($n = 6$) for the alternative variant (electronic supplementary material, table S6). Importantly, the Z-linked candidate region is also located within the approximately 7.2 cM interval previously shown using linkage mapping to contain the red/black locus [20]. Taken together, our results strongly suggest that our candidate region contains the gene underlying red and black head coloration in the Gouldian finch.

We examined the gene content of the candidate region using the annotation of the Gouldian finch genome described above. We failed to identify any protein-coding genes within the candidate interval (figure 2*d*), suggesting that the underlying causative variant, or variants, is likely to be regulatory. The homologous genomic regions both in the chicken and zebra finch genomes are also devoid of protein-coding genes. We found, however, two protein-coding genes flanking the candidate region (figure 2*d*): *FST* and *MOCS2*. *MOCS2* encodes a molybdopterin synthase associated with the biosynthesis of the molybdenum cofactor, which is essential to the function of several enzymes [44]. *FST* encodes follistatin, which functions as an antagonist of transforming growth factor beta 1 family members (including bone morphogenetic proteins and activin) and is widely expressed, including in the growing feather bud and gonad [21,45]. *FST* is thus a strong candidate to explain the behavioural, morphological and pigmentary differences between red and black head morphs.

(d) Deep divergence between red and black alleles but no evidence for structural rearrangements

Next, we investigated the existence of large structural rearrangements that might differ between haplotypes (i.e. inversions, deletions, duplications or translocations). First, we aligned the sequence surrounding our candidate region with zebra finch. This alignment showed that the Gouldian and zebra finch genomes are syntenic across this region and that it is unlikely that the Gouldian sequence contains any major gaps that could reduce our chances of finding structural variation (electronic supplementary material, figure S3). We then took advantage of our diploid assembly, and the fact that we sequenced an individual heterozygous for the red and black alleles, to produce an alignment between the red and black haplotypes across the candidate genomic interval. This analysis revealed that the two haplotypes had a similar overall structure with no evidence for large-scale rearrangements or copy number variation (figure 3*a*). Finally, we applied to both linked-read and whole-genome resequencing data several tools for detection of structural rearrangements (see Material and methods), but we also failed to detect large structural rearrangements overlapping the candidate region.

We noted, however, that the red and black haplotypes were highly divergent. To test how unusual this level of divergence was relative to the remainder of the Z chromosome, we estimated the RND statistic between the paternal and maternal haplotypes in our diploid assembly using a sliding window of 20 kb. RND values vary between 0 and 1 and provide a relative estimate of divergence between sequences by correcting for mutation rate using an outgroup species. We used the zebra finch as the outgroup, and the analysis was restricted to the Z chromosome because the Z chromosome and autosomes differ in multiple aspects that can cause systematic biases in estimates of divergence, including mutation, recombination and effective population size, making those estimates not directly comparable between different genomic compartments.

We evaluated 11 568 windows and found that our region of association was the most divergent region on the Z chromosome and contained the top 12 windows of the empirical distribution of RND (figure 3*b*). The maximum nucleotide divergence between haplotypes in the candidate region was much higher (2.5%) than the average for the Z chromosome (0.5%; figure 3*c*). Based on an RND value in the candidate region of 0.41 (approx. 40% of the divergence between Gouldian and zebra finch), and a split-time between zebra and Gouldian finch of approximately 10 Myr [46], we estimate that the red and black alleles may have started diverging approximately 4 Ma. These results suggest that the head

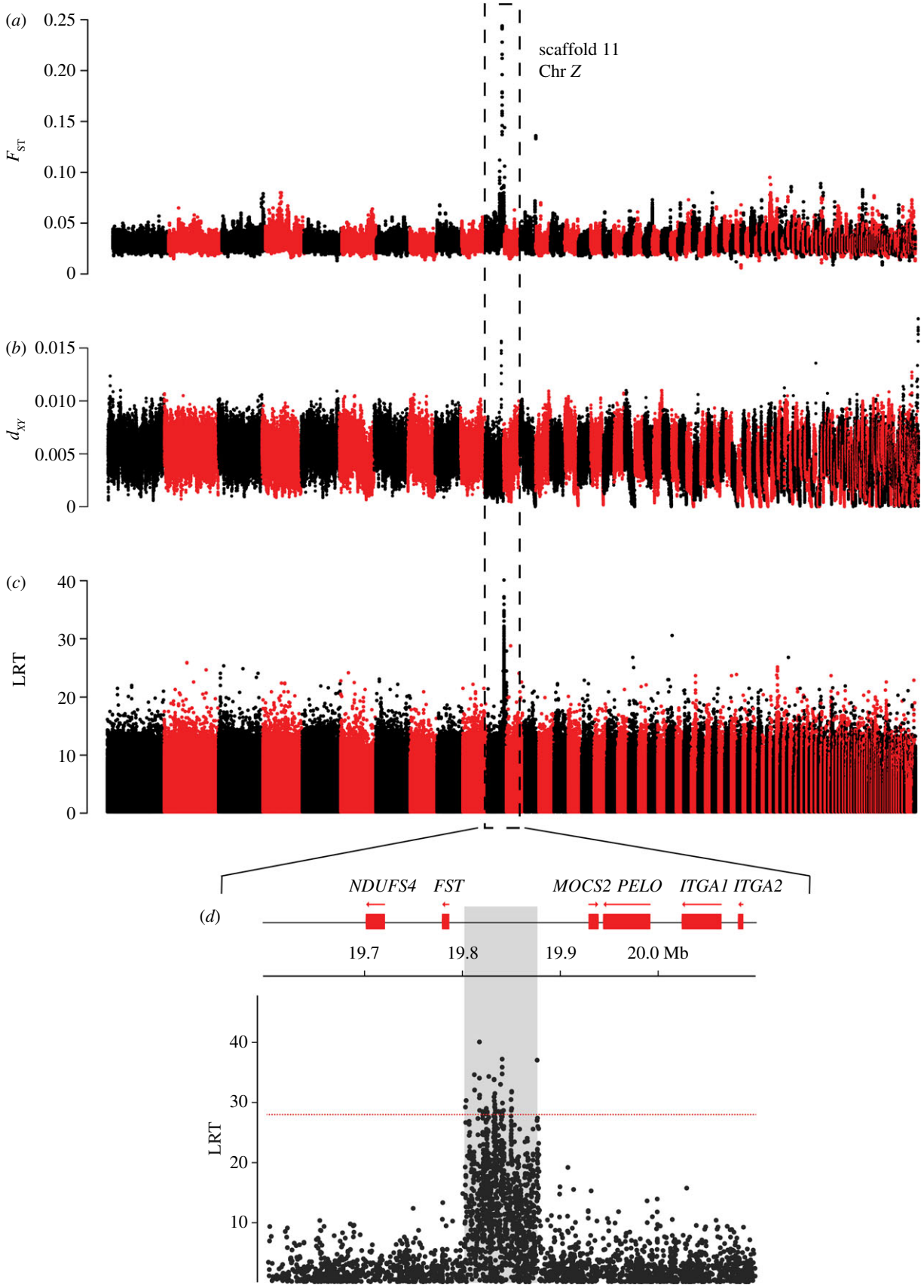

Figure 2. Genetic mapping of the head-colour polymorphism using whole-genome sequencing. (*a*) Manhattan plot summarizing F_{ST} values between red and black individuals across the genome. Each dot represents F_{ST} estimated in 20 kb windows with 5 kb steps. Windows with less than 80% of the positions passing filters were excluded. All scaffolds of the reference genome containing at least one valid window are presented along the *x*-axis. (*b*) Manhattan plot summarizing d_{XY} between red and black individuals across the genome. Each dot represents d_{XY} estimated in windows of 5000 positions passing filters (both variant and invariant) with 1000 positions steps. All scaffolds of the reference genome containing at least one valid window are presented along the *x*-axis. (*c*) Manhattan plot summarizing likelihood ratio tests (LRTs) for differences in allele frequencies between red and black individuals across the genome. Each dot represents an individual SNP. The scaffolds of the reference genome are presented along the *x*-axis. (*d*) Close-up of the LRTs across the candidate region. The horizontal line indicates the standard genome-wide significance threshold of $p = 5 \times 10^{-8}$ (approx. LRT > 28). The shaded area represents the region of association. The protein-coding genes identified by the genome annotation are represented by red boxes. (Online version in colour.)

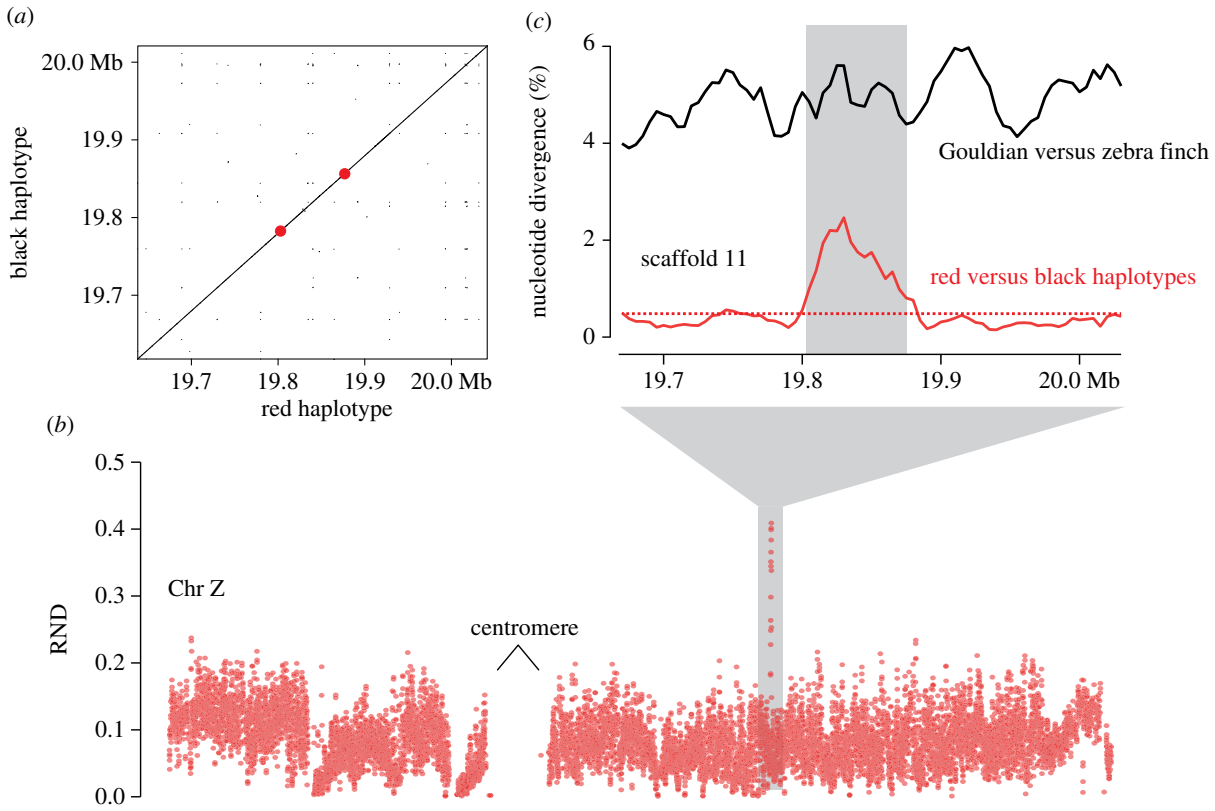

Figure 3. Haplotype analysis in the candidate region and across the genome. (*a*) Dot plot summarizing the alignment between the red and black alleles in the region of association (marked by red dots) and flanking regions. (*b*) RND across the Z chromosome between haplotypes inferred from our diploid genome assembly. RND was estimated in 20 kb windows with 5 kb steps. Coordinates along the *x*-axis are from the zebra finch. (*c*) Nucleotide divergence between the red and black alleles and between Gouldian finch and zebra finch within and around the region of association. Nucleotide divergence was estimated in 20 kb windows with 5 kb steps. The horizontal line indicates the average between haplotypes across the Z chromosome. The shaded area represents the region of association. (Online version in colour.)

polymorphism is evolutionarily old as expected for genetic variation preserved by balancing selection.

(e) Expression profiling of regenerating feathers and testis demonstrates modest gene expression differences between morphs

Initially, we analysed differential gene expression in the regenerating mask skin of black and red morphs by RNA-seq (electronic supplementary material, table S3). We harvested regenerating skin from eight males (three black and five non-black (three orange and two red)) that had their mask feathers plucked 10 days prior to induce feather regeneration. From the 20 058 genes present in our annotation, we evaluated expression differences for 16 535 genes (82.4%). The remaining genes were excluded from the analysis owing to an insufficient number of mapped reads.

We detected 14 genes that were differentially expressed between the two groups after controlling for multiple testing (figure 4*a*; electronic supplementary material, table S7). None of these genes were located near our candidate region and none seemed to have functions consistent with a role in carotenoid or melanin pigmentation. Both *FST* and *MOCS2*, the two genes immediately flanking the candidate region, showed no significant differences in expression between groups by RNA-seq. The lack of transcriptomic differences between the morphs in this experiment may be the result of sampling a late stage of feather regeneration, at a time after the first half of barb development is already complete.

Upon closer analysis of the RNA-seq data, we identified multiple transcript isoforms of *FST* and *MOCS2* expressed in the regenerating mask skin. *FST* occurs as two isoforms we have termed *FST X1* and *FST X2* (electronic supplementary material, figures S4 and S5A). These isoforms differ in whether or not splicing has occurred between exons 5 and 6 and results in different protein products that probably have distinct functions (electronic supplementary material, figure S5B). *FST X1* is homologous to the *FST315* transcript in mice that includes an acidic C-terminus and is the primary form found in the circulation (electronic supplementary material, figure S5B) [46,47]. *FST X2*, by contrast, is homologous to the *FST288* transcript in mice that lacks the acidic C-terminus and is typically membrane-associated and acts locally within tissues (electronic supplementary material, figure S5B) [47,48]. We quantified expression of these transcript isoforms by quantitative reverse transcription polymerase chain reaction (RT-qPCR) and found that the relative proportion of isoform *X1* to *X2* expressed in regenerating skin did not differ significantly between the morphs (electronic supplementary material, figure S5C).

The *MOCS2* transcript contains two separate open reading frames that encode subunits A and B of the enzyme [49,50]. Studies of the human *MOCS2* indicate that splicing of an alternative first exon determines which open read frame will be translated [49,50]. Consistent with a similar mechanism in birds, we observed evidence of alternative splicing of the first

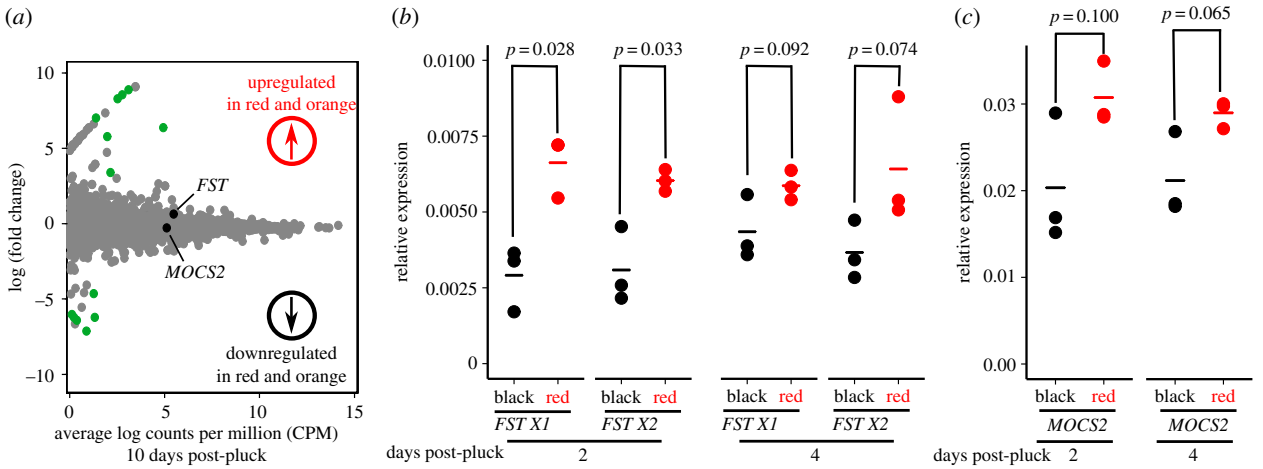

Figure 4. Gene expression in the skin of black and red morphs. (*a*) Volcano plot of RNA-seq analysis performed on 10 days post-pluck regenerating mask skin from black and non-black (red and orange) Gouldian finches. Genes that are differentially expressed (false discovery rate < 0.05) are shown in green. Positive log (fold change) values indicate upregulation in non-black morphs. (*b*) qPCR measurements of *FST* isoform *X1* and *X2* relative to *GAPDH* in the regenerating mask skin at 2 and 4 days post-pluck. The points represent samples from individual birds, and bar indicates mean for each morph at each time point. (*c*) qPCR measurements of *MOCS2* expression relative to *GAPDH*.

exon *MOCS2* in the Gouldian finch with RNA-seq reads spanning either exons 1 and 2 or exons 1 and 3 (electronic supplementary material, figure S6). However, the frequency of these splicing events did not differ significantly between the colour morphs, indicating that the abundance of these splice isoforms is similar between morphs (electronic supplementary material, figure S7)

To examine gene expression at early stages of feather regeneration, we harvested regenerating mask skin from six males (three black and three red) that had their feathers plucked 2 and 4 days prior and measured the expression of each isoform of *FST* and *MOCS2* by RT-qPCR (figure 4*b*,*c*). After 2 days of feather regeneration red morph males had modestly higher levels of expression of both isoforms of *FST* (*X1*: $t_4 = -3.36$, $p = 0.028$, *X2*: $t_4 = -3.20$, $p = 0.033$), but there were no significant differences between the morphs after 4 days of feather development (*X1*: $t_4 = -2.20$, $p = 0.092$, *X2*: $t_4 = -2.40$, $p = 0.074$). There were no significant differences in *MOCS2* expression between the morphs at either time point (day 2: $t_4 = -2.14$, $p = 0.10$, day 4: $t_4 = -2.52$, $p = 0.065$).

Red and black head-colour morphs have distinct aggressive behavioural phenotypes and dynamics of testosterone expression that might be mediated by the action of the candidate genes within the gonads. To determine if and how *FST* and *MOCS2* expression varies between the head-colour morphs, we analysed expression in the testis of four red and four black morph males by qPCR (electronic supplementary material, figure S8). We found that red males expressed higher levels of *FST* in the testis compared to black males. However, *FST* expression was highly variable among red individuals, and only isoform *X1* differed significantly between red and black morphs (*X1*: $t_6 = -3.65$, $p = 0.011$, *X2*: $t_6 = -2.18$, $p = 0.072$). *MOCS2* expression in the testis did not differ significantly between the morphs ($t_{3.04} = -1.76$, $p = 0.18$).

4. Conclusion

Intraspecific colour morphs coexisting in sympatry are classic examples of evolutionary diversification driven by both natural and sexual selection [1]. In this study, we generated a reference genome and performed whole-genome sequencing to dissect the genetic underpinnings of the pigmentation differences and correlated traits observed between head-colour morphs of the Gouldian finch. Regardless of the existence of assortative mating with respect to colour [15] and the existence of distinct physiological and behavioural phenotypes, we show that the red and black morphs have nearly identical genome sequences, and that genetic differences between morphs are restricted to a relatively small genomic interval that was found to be perfectly associated with the head-colour polymorphism. This interval is in a non-coding region upstream of *FST*. *FST* has been indirectly linked to plumage and coloration differences between species of flycatchers and warblers [51,52], suggesting that this gene might be an important player in the evolution of pigmentation patterns in birds.

Our results further suggest that the distinct feather colour and morphology of the two morphs may be the product of *cis*-regulatory differences that drive differential expression of *FST* during a discrete period early in feather development/regeneration. It is also possible that this same regulatory variant(s) mediate(s) differences in gonadal *FST* expression that might explain the endocrinological and behavioural differences between the two morphs. Future investigations will address these intriguing possibilities. Overall, the identification of the genomic region governing the red and black head morphs is a crucial first step towards understanding the mechanisms promoting the maintenance of sympatric colour polymorphisms and their correlated traits, and answering why colour polymorphic species evolve so frequently in nature.

Ethics. Animal care followed the recommendations of the Guide for the Care and Use of Laboratory Animals (National Research Council—USA) and National and European regulations for the maintenance of birds in captivity (Federation of European Laboratory Animal Science Associations). Birds were euthanized following practices outlined by the American Veterinary Medical Association and approved by the ethical commission CIBIO-ORBEA (02/2018).

Data accessibility. The reference genome sequence is publicly available at NCBI (accession no. SAMN09689584). The linked-read, whole-genome resequencing and RNA-sequencing data are available at the NCBI Sequence Read Archive (accession no. SRP1529467). Isoform sequences

and qPCR raw results are available from Dryad Digital Repository at: http://dx.doi.org/10.5061/dryad.8p97m88 [53].

Authors' contributions. M.B.T., R.J.L., J.C.C. and M.C. designed research. M.B.T., C.I.M., P.M.A., S.A. and R.J.L. performed research. M.B.T., C.I.M., P.A., S.S., M.A.G. and M.C. analysed data and M.B.T., C.I.M., J.C.C. and M.C. wrote the paper with input from all authors.

Competing interests. We declare we have no competing interests.

Funding. This 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 post-doc fellowships to R.J.L. (SFRH/BPD/84141/2012) and S.S. (SFRH/BPD/99138/2013)); by research fellowships attributed to M.A.G. (PD/BD/114042/2015) and P.A. (PD/BD/114028/2015) in the scope of the Biodiversity, Genetics, and Evolution (BIODIV) PhD programme at CIBIO/InBIO and University of Porto, and by grants to J.C.C. from the National Institutes of Health (EY024958, EY025196 and EY026672).

Acknowledgements. We gratefully acknowledge the electron microscopy expertise of Greg Strout, Matt Joens and Dr James Fitzpatrick at the Washington University Center for Cellular Imaging (WUCCI), which is supported by the Washington University School of Medicine, The Children's Discovery Institute of Washington University and St Louis Children's Hospital (CDI-CORE-2015-505) and the Foundation for Barnes-Jewish Hospital (3770).

References

1. McKinnon JS, Pierotti MER. 2010 Colour polymorphism and correlated characters: genetic mechanisms and evolution. *Mol. Ecol.* **19**, 5101–5125. (doi:10.1111/j.1365-294X.2010.04846.x)
2. Gray SM, McKinnon JS. 2007 Linking color polymorphism maintenance and speciation. *Trends Ecol. Evol.* **22**, 71–79. (doi:10.1016/j.tree.2006.10.005)
3. Hugall AF, Stuart-Fox D. 2012 Accelerated speciation in colour-polymorphic birds. *Nature* **485**, 631–634. (doi:10.1038/nature11050)
4. Poelstra JW *et al.* 2014 The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science* **344**, 1410–1414. (doi:10.1126/science.1253226)
5. Tuttle EM. 2003 Alternative reproductive strategies in the white-throated sparrow: behavioral and genetic evidence. *Behav. Ecol.* **14**, 425–432. (doi:10.1093/beheco/14.3.425)
6. Tuttle EM *et al*. 2016 Divergence and functional degradation of a sex chromosome-like supergene. *Curr. Biol.* **26**, 344–350. (doi:10.1016/j.cub.2015.11.069)
7. Küpper C *et al*. 2016 A supergene determines highly divergent male reproductive morphs in the ruff. *Nat. Genet.* **48**, 79–83. (doi:10.1038/ng.3443)
8. Lamichhaney S *et al*. 2016 Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax*). *Nat. Genet.* **48**, 84–88. (doi:10.1038/ng.3430)
9. Franklin DC, Dostine PL. 2000 A note on the frequency and genetics of head colour morphs in the Gouldian finch. *Emu-Austral Ornithol.* **100**, 236–239. (doi:10.1071/MU00911)
10. Pryke SR. 2007 Fiery red heads: female dominance among head color morphs in the Gouldian finch. *Behav. Ecol.* **18**, 621–627. (doi:10.1093/beheco/arm020)
11. Pryke SR, Astheimer LB, Buttemer WA, Griffith SC. 2007 Frequency-dependent physiological trade-offs between competing colour morphs. *Biol. Lett.* **3**, 494–497. (doi:10.1098/rsbl.2007.0213)
12. Pryke SR, Griffith SC. 2006 Red dominates black: agonistic signalling among head morphs in the colour polymorphic Gouldian finch. *Proc. R. Soc. B* **273**, 949–957. (doi:10.1098/rspb.2005.3362)
13. Pryke SR, Griffith SC. 2009 Postzygotic genetic incompatibility between sympatric color morphs. *Evolution* **63**, 793–798. (doi:10.1111/j.1558-5646.2008.00584.x)
14. Pryke SR, Griffith SC. 2009 Genetic incompatibility drives sex allocation and maternal investment in a polymorphic finch. *Science* **323**, 1605–1607. (doi:10.1126/science.1168928)
15. Pryke SR, Griffith SC. 2007 The relative role of male vs. female mate choice in maintaining assortative pairing among discrete colour morphs. *J. Evol. Biol.* **20**, 1512–1521. (doi:10.1111/j.1420-9101.2007.01332.x)
16. Templeton JJ, Mountjoy DJ, Pryke SR, Griffith SC. 2012 In the eye of the beholder: visual mate choice lateralization in a polymorphic songbird. *Biol. Lett.* **8**, 924–927. (doi:10.1098/rsbl.2012.0830)
17. Bolton PE, Rollins LA, Brazill-Boast J, Kim KW, Burke T, Griffith SC. 2017 The colour of paternity: extra-pair paternity in the wild Gouldian finch does not appear to be driven by genetic incompatibility between morphs. *J. Evol. Biol.* **30**, 174–190. (doi:10.1111/jeb.12997)
18. Southern HN. 1945 Polymorphism in *Poephila gouldiae* gould. *J. Genet.* **47**, 51–57. (doi:10.1007/BF02989037)
19. Stradi R. 1999 Pigmenti e sistematica degli uccelli. In *Colori in volo: il piumaggio degli uccelli* (eds LC Brambilla, E Giovanni, C Mannucci, R Massa, N Saino, R Stradi, G Zerbi), pp. 117–146. Milan, Italy: Università degli Studi di Milano.
20. Kim KW, Griffith SC, Burke T. 2016 Linkage mapping of a polymorphic plumage locus associated with intermorph incompatibility in the Gouldian finch (*Erythrura gouldiae*). *Heredity* **116**, 409–416. (doi:10.1038/hdy.2015.114)
21. Uhlén M *et al*. 2015 Tissue-based map of the human proteome. *Science* **347**, 1260419. (doi:10.1126/science.1260419)
22. McGraw KJ, Hudon J, Hill GE, Parker RS. 2005 A simple and inexpensive chemical test for behavioral ecologists to determine the presence of carotenoid pigments in animal tissues. *Behav. Ecol. Sociobiol.* **57**, 391–397. (doi:10.1007/s00265-004-0853-y)
23. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017 Direct determination of diploid genome sequences. *Genome Res.* **27**, 1–11. (doi:10.1101/gr.214874.116)
24. Holt C, Yandell M. 2011 MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491. (doi:10.1186/1471-2105-12-491)
25. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015 BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212. (doi:10.1093/bioinformatics/btv351)
26. Frith MC, Hamada M, Horton P. 2010 Parameters for accurate genome alignment. *BMC Bioinformatics* **11**, 80. (doi:10.1186/1471-2105-11-80)
27. Tan JA, Mikheyev AS. 2016 A scaled-down workflow for Illumina shotgun sequencing library preparation: lower input and improved performance at small fraction of the cost. *PeerJ Prepr.* **4**, e2475v1. (doi:10.7287/PEERJ.PREPRINTS.2475V1)
28. Korneliussen TS, Albrechtsen A, Nielsen R. 2014 ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356. (doi:10.1186/s12859-014-0356-4)
29. Feder JL, Xie X, Rull J, Velez S, Forbes A, Leung B, Dambroski H, Filchak KE, Aluja M. 2005 Mayr, Dobzhansky, and Bush and the complexities of sympatric speciation in *Rhagoletis*. *Proc. Natl Acad. Sci. USA* **102**, 6573–6580. (doi:10.1073/pnas.0502099102)
30. Kofler R, Orozco-terWengel P, de Maio N, Pandey RV, Nolte V, Futschik A, Kosiol C, Schlötterer C. 2011 Popoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS ONE* **6**, e15925. (doi:10.1371/journal.pone.0015925)
31. Darling AE, Mau B, Perna NT. 2010 Progressivemauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147. (doi:10.1371/journal.pone.0011147)
32. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011 Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26. (doi:10.1038/nbt.1754)

33. Chen K *et al*. 2009 BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6**, 677–681. (doi:10.1038/nmeth.1363)
34. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012 DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339. (doi:10.1093/bioinformatics/bts378)
35. Layer RM, Chiang C, Quinlan AR, Hall IM. 2014 LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84. (doi:10.1186/gb-2014-15-6-r84)
36. Robinson MD, McCarthy DJ, Smyth GK. 2010 edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140. (doi:10.1093/bioinformatics/btp616)
37. Trapnell C *et al*. 2012 Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578. (doi:10.1038/nprot.2012.016)
38. Katz Y *et al*. 2015 Quantitative visualization of alternative exon expression from RNA-seq data. *Bioinformatics* **31**, 2400–2402. (doi:10.1093/bioinformatics/btv034)
39. Brush AH, Seifried H. 1968 Pigmentation and feather structure in genetic variants of the Gouldian finch, *Poephila gouldiae*. *Auk* **85**, 416–430. (doi:10.2307/4083290)
40. Chen C-F, Foley J, Tang P-C, Li A, Jiang TX, Wu P, Widelitz RB, Chuong CM. 2015 Development, regeneration, and evolution of feathers. *Annu. Rev. Anim. Biosci.* **3**, 169–195. (doi:10.1146/annurev-animal-022513-114127)
41. Bellott DW *et al*. 2010 Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. *Nature* **466**, 612–616. (doi:10.1038/nature09172)
42. Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011 Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12**, 443–451. (doi:10.1038/nrg2986)
43. Kim S *et al*. 2011 Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics* **12**, 231. (doi:10.1186/1471-2105-12-231)
44. Reiss J, Johnson JL. 2003 Mutations in the molybdenum cofactor biosynthetic genes *MOCS1*, *MOCS2*, and *GEPH*. *Hum. Mutat.* **21**, 569–576. (doi:10.1002/humu.10223)
45. Patel K, Makarenkova H, Jung HS. 1999 The role of long range, local and direct signalling molecules during chick feather bud development involving the BMPs, Follistatin and the Eph receptor tyrosine kinase Eph-A4. *Mech. Dev.* **86**, 51–62. (doi:10.1016/S0925-4773(99)00107-0)
46. Hooper DM, Price TD. 2015 Rates of karyotypic evolution in Estrildid finches differ between island and continental clades. *Evolution* **69**, 890–903. (doi:10.1111/evo.12633)
47. Kimura F, Sidis Y, Bonomi L, Xia Y, Schneyer A. 2010 The follistatin-288 isoform alone is sufficient for survival but not for normal fertility in mice. *Endocrinology* **151**, 1310–1319. (doi:10.1210/en.2009-1176)
48. Sidis Y, Mukherjee A, Keutmann H, Delbaere A, Sadatsuki M, Schneyer A. 2006 Biological activity of follistatin isoforms and follistatin-like-3 is dependent on differential cell surface binding and specificity for activin, myostatin, and bone morphogenetic proteins. *Endocrinology* **147**, 3586–3597. (doi:10.1210/en.2006-0089)
49. Hahnewald R, Leimkühler S, Vilaseca A, Acquaviva-Bourdain C, Lenz U, Reiss J. 2006 A novel *MOCS2* mutation reveals coordinated expression of the small and large subunit of molybdopterin synthase. *Mol. Genet. Metab.* **89**, 210–213. (doi:10.1016/j.ymgme.2006.04.008)
50. Reiss J, Dorche C, Stallmeyer B, Mendel RR, Cohen N, Zabot MT. 1999 Human molybdopterin synthase gene: genomic structure and mutations in molybdenum cofactor deficiency type B. *Am. J. Hum. Genet.* **64**, 706–711. (doi:10.1086/302296)
51. Lehtonen PK *et al*. 2012 Candidate genes for colour and vision exhibit signals of selection across the pied flycatcher (*Ficedula hypoleuca*) breeding range. *Heredity* **108**, 431–440. (doi:10.1038/hdy.2011.93)
52. Toews DPL, Taylor SA, Vallender R, Brelsford A, Butcher BG, Messer PW, Lovette IJ. 2016 Plumage genes and little else distinguish the genomes of hybridizing warblers. *Curr. Biol.* **26**, 2313–2318. (doi:10.1016/j.cub.2016.06.034)
53. Toomey MB *et al*. 2018 Data from: A non-coding region near *Follistatin* controls head colour polymorphism in the Gouldian finch. Dryad Digital Repository. (doi:10.5061/dryad.8p97m88)