Light and electron microscopy

To characterize head feather morphology, we used a combination of light and electron microscopy. We collected full color macroscopic images by mounting whole feathers on white card stock and imaging them through a Leica MZ7 dissecting microscope (Leica Microsystems Inc., Buffalo Grove, IL). We then embedded unfixed feather samples in Tissue-Tek OCT compound (Sakura, Torrance, CA), cut 12 µm cross-sections of the feather barb tips, and mounted these on Superfrost® microscopes slides (VWR inc. Radnor, PA) with gelvatol mounting medium. We imaged the barb sections at 400x magnification on an Olympus BX-51 microscope.

To acquired more detailed images of the feather structure, we imaged whole feathers with a Zeiss Merlin field emission scanning electron microscope (FE-SEM) at the Washington University Center for Cellular Imaging. Whole red and black feathers were mounted on iridium-coated 12 mm circular glass coverslips using carbon conductive adhesive to minimize charging artifacts. The coverslips were then mounted on aluminum stubs using carbon sticky tabs and silver paint, coated with 12 nm of iridium using a Leica EM ACE 600 sputter coater. The feathers were imaged at a 7.8 mm operating distance and 3 kV accelerating voltage.

To characterize the ultrastructure of the head feathers we carried out transmission electron microscopy (TEM) at the Washington University Center for Cellular Imaging. Feather barb samples were cut from the distal ~2 mm of multiple feathers of each color variant, washed in 0.25 M sodium hydroxide and 0.1% Tween-20 for 30 minutes, then transferred to a solution of 2 parts formic acid to 3 parts ethanol for 3 hours. The samples were dehydrated in 100% ethanol (3x10 minutes each), infiltrated in 15%, 50%, 70%, and 100% (3x) Spurr's Resin (18-24 hrs. each step), and cured at 60ºC for 48 hours. Cross

sections of feather barbs were cut using a diamond knife on a Leica EM UCT7 ultramicrotome (Leica Microsystems GmbH, Wetzlar, Germany), then post stained in 2% Uranyl Acetate followed by Reynolds lead citrate [1], and viewed on a JEOL 1400 Plus transmission electron microscope operated at 120kV.

Carotenoid analysis

We collected ten mature feathers from the heads of red and black morph male Gouldian finches and extracted carotenoids with acidified pyridine and hexane following the methods of McGraw et al. (2005) [2]. We then evaporated the extracts to dryness with a stream of nitrogen, resuspended them in 120 µl of methanol:acetonitrile 1:1 (vol:vol), and injected them into an Agilent 1100 series high performance liquid chromatography (HPLC) instrument fitted with a 5.0 µm carotenoid column (4.6 mm × 250 mm, YMC). The mobile phase consisted of acetonitrile:methanol:dichloromethane (44:44:12) (vol:vol:vol) for 11 minutes followed by a ramp up to acetonitrile:methanol:dichloromethane (35:35:30) in minutes 11-21 then with continued isocratic conditions through 35 minutes. We used a flow rate of 1.2 ml min^{-1}, and a column temperature of 30°C. We monitored the samples with a photodiode array detector at 400, 445, and 480 nm, and identified carotenoids by comparison to authentic standards (astaxanthin, a gift of DSM nutritional products, Heerlen, the Netherlands) or by comparison to published accounts [3–5].

***De novo* genome assembly and annotation**

Sampling and DNA extraction. Genomic DNA of a single male bird was subjected to sequencing to generate a *de novo* genome assembly of the Gouldian finch (*Erythrura*

gouldiae). This individual was heterozygous for the red and black alleles and wild-type for all other color phenotypes found in captive populations of this species. The individual was euthanized by manual cervical dislocation after being rendered unconscious with an anaesthetic gas (isoflurane). Immediately after sacrifice and dissection, tissues were snap-frozen in dry ice and stored at -80°C. High molecular weight genomic DNA was extracted from muscle tissue using the MagAttract HMW DNA Kit (QIAGEN, Germantown, USA), followed by pulsed-field gel electrophoresis to evaluate DNA integrity. The size of the DNA was estimated from the gel to be around 150 kb with minimal shearing.

Chromium library preparation and sequencing. Linked-read data for genome assembly was obtained using 10x Chromium technology (10x Genomics, San Francisco, USA). A Chromium library was prepared at the Genomics Services Laboratory of the HudsonAlpha Institute for Biotechnology following manufacturer's instructions. Prior to sequencing, the library was quantified by qPCR using the KAPA Library Quantification Kit (Kapa Biosystems, Wilmington, USA), and DNA fragment size and distribution was assessed on a Bioanalyzer (Agilent Technologies, Waldbronn, Germany). The library was then sequenced on a single lane of an Illumina HiSeq X Ten sequencer using 2 × 150 bp paired-end reads. We obtained ~477 million reads representing a raw coverage of ~115-fold.

Genome assembly. Linked-read data was assembled using the *Supernova* Assembler (version 2.0) [6]. The recommended sequencing coverage for genome assembly using *Supernova* is between 38-fold and 56-fold. However, it has been reported that coverage >56-fold can improve the results [6]. We therefore attempted several assemblies using variable read depth (45-, 56-, 60-, and 70-fold). Other than sequencing coverage, all options were set to default.

The assembly with the highest N50 values was obtained using 60X read coverage and was selected as the final assembly.

Genome annotation. The reference genome was annotated by analysis of sequence composition and by generating *ab initio* gene models using transcriptome and protein data. Repeat regions were identified and masked using *RepeatMasker* [7] with a finch-specific repeat library generated for our *de novo* finch genome using *RepeatModeler* (http://www.repeatmasker.org/RepeatModeler/) and RepBase23.02. Intron/exon boundaries were inferred by aligning RNA-seq data generated from skin tissue (see RNA-seq methods below), that was trimmed using *Trimmomatic* (v0.3220) [8], to our finch genome using *HISAT2* [9]. These aligned reads were used to obtain a genome-guided transcriptome assembly using *Cufflinks* (v2.2.1)[10]. The trimmed, skin RNA-seq data was also assembled into a transcriptome using *TRINITY* (v2.2.0)[11].

An initial run of *Maker2* [12] was conducted on the repeat-masked genome using the output from *Cufflinks*, the *de novo* transcriptome assembly from *Trinity*, and high-confidence protein sequence evidence from the Uniprot Swiss-Prot and Uniprot zebra finch (*Taeniopygia guttata*) databases. The genome was then re-annotated with *Maker2* using gene models generated with *GeneMark-ES* [13], *SNAP* [14], and *AUGUSTUS* [15] runs using default parameters. The *AUGUSTUS* and *SNAP* models were trained according to author recommendations. To characterize the functions of the resulting transcripts, we ran *InterProscan* (v5.7–48)[16] to retrieve *Interpro* [17], *PFAM* [18] and *GO* [19] terms. Transcripts were also compared to the Uniprot protein databases mentioned earlier using *BLASTp* (v2.2.28+). In addition, tRNAs were identified using *tRNAscan-SE-2.0* [20].

Quantitative assessment of assembly and annotation completeness. We quantified the number of highly conserved single-copy orthologs present both in the reference sequence and in the annotation using the Benchmarking Universal Single-Copy Orthologs (BUSCO) software (v. 3) [21,22]. We used the aves_0db9 standard data set, which includes 4915 genes conserved among 40 representative bird species. To assess the quality of the assembly per chromosome, we aligned the Gouldian finch and the zebra finch reference sequences using *LAST* [23]. We kept a unique best alignment for each region using *last-split* and alignments with error probability $>10^{-5}$ were discarded. We only considered a scaffold homologous to a given zebra finch chromosome when we obtained at least one alignment block between the two sequences larger than 5,000 bp.

Whole genome resequencing

Sampling and initial laboratory procedures. To obtain genome-wide polymorphism data we performed whole-genome resequencing. Black (n=21) and red-headed (n=21) captive birds were obtained from 13 private aviaries in Portugal (Table S1). Blood was collected in a heparin-free capillary tube by brachial venipuncture with a sterile needle and transferred into a vial containing 96% ethanol. Genomic DNA was extracted using an EasySpin Genomic DNA Tissue Kit SP-DT-250 (Citomed, Lisbon, Portugal), and RNA was removed with a RNAse A digestion step.

Library preparation and sequencing. Individual paired-end libraries were prepared following a modified version [24] of Illumina's Nextera XT protocol (Illumina, San Diego, USA). The libraries were quantified by qPCR using the KAPA Library Quantification Kit, pooled, and

sequenced at low coverage (average = 1.63X) using 2 x 125 bp paired-end reads on an Illumina Hiseq 1500 machine (Table S1).

Read quality control and mapping. After sequencing, read quality was inspected with *FastQC v0.10.1* [25]. Since the protocol to generate the libraries included PCR steps, we identified and removed duplicates using *Picard MarkDuplicates* (http://broadinstitute.github.io/picard). Sequencing reads were then mapped to the Gouldian finch reference genome assembly using *BWA-MEM* with default settings [26]. Sequencing and mapping summary statistics were computed using *SAMtools* [27] (Table S1).

Assessment of population structure. We investigated patterns of population structure among the sequenced individuals using Principal Component Analysis (PCA). In order to take into account uncertainty in genotype calls due to low sequencing depth, this analysis was carried out using genotype probability methods as implemented in *ANGSD* [28,29], instead of relying on standard hard filter approaches for inferring genotypes. To further avoid errors and biases derived from uneven depths of sequencing coverage among individuals, we followed a single-read per position sampling approach. Genotype likelihoods from variable sites were used to estimate a covariance matrix between individuals, as implemented in the *ngsPopGen* package from *ngsTools* [30–32]. The resulting matrix was decomposed to principal components and plotted.

Genetic mapping. To identify the genomic region associated with the two alternative color morphs, we took a two-folded approach. As before, the analyses described below were based on probabilistic methods tailored for low-coverage sequencing datasets as implemented in

143 *ANGSD*. First, we performed a genome-wide association analysis using individual variants
144 [29]. Genome-wide allele frequency differentiation between groups was computed following
145 a case-control approach, in which black-headed finches were considered cases and red-
146 headed finches controls. Differences between cases and controls were summarized using a
147 likelihood-ratio test (LRT) and we interpreted the results using two significance thresholds:
148 (1) the standard threshold for significance commonly applied in genome-wide association
149 studies ($P <= 5.00 \times 10^{-8}$; approximately LRT > 28), and (2) an even more stringent
150 Bonferroni correction ($P <= 6.17 \times 10^{-9}$; approximately LRT > 34). Variants with a minor
151 allele frequency lower than 20% were excluded from the analysis. A quantile–quantile plot
152 summarizing the distribution of observed and expected LRT values was calculated using
153 *SNPStats* [33], and the resulting plot indicated a reasonably good fit to the null expectation
154 throughout most of the distribution (Figure S9). We observed a departure from the expected
155 distribution at higher LRT values. However, this seems to be driven to a large extent by SNPs
156 mapping to our candidate region that are associated with the phenotype.
157 Second, we estimated genetic differentiation across the genome using a sliding-
158 window approach by means of the fixation index (F_{ST}) and the average number of pairwise
159 differences per site (d_{XY}). d_{XY} values per SNP were obtained using a script provided with
160 *ANGSD* (https://github.com/mfumagalli/ngsPopGen/tree/master/scripts). For both statistics,
161 we required reads with a mapping quality ≥ 30 and an individual base quality ≥ 20. To
162 consider a position, we required a minimum of five individuals having reads overlapping a
163 given position and a maximum coverage twice the average coverage in each group (red and
164 black). Prior to the calculation of the statistics, we also performed two sliding-window filters
165 of the called variants: (1) to avoid considering false SNPs that can arise from local
166 misalignments, we excluded clusters of three or more SNPs within a 5 bp window; and (2)

to avoid false SNPs that can arise from incorrectly aligned reads, we excluded windows of 20 bp characterized by a high density of SNPs (five or more). F_{ST} values were averaged over 20 kb windows with a 5 kb step across each scaffold and windows with less than 80% of the positions passing filters were excluded. Other window sizes were attempted (10 and 50 kb), and the results were qualitatively the same. d_{XY} values were averaged over 5,000 positions passing filters (both polymorphic and monomorphic positions) with steps of 1,000 positions. Windows in small scaffolds or at the end of scaffolds that did not reach 5,000 positions were not considered. Since our full dataset comprised captive-bred Gouldian finches bearing other color mutations that emerged during the domestication process, the F_{ST} and d_{XY} analysis was restricted to the subset of individuals characterized by wild-type coloration (13 black and 12 red; Table S1) in order to avoid signatures of elevated differentiation in genomic regions harboring artificially selected mutations.

SNP genotyping

To confirm the association, we selected a set of seven closely located SNPs contained within a 275 bp fragment that were found to be strongly associated with the red and black alleles using the whole-genome resequencing data. These SNPs were genotyped by Sanger sequencing on a BiosystemsTM 3130XL Sequencer following PCR amplification. The amplified fragment was located between positions 19,840,503 and 19,840,778 on scaffold 11. Primers sequences are given in Table S2.

Haplotype analyses

Relative node depth. The relative node depth (RND) is a measure of divergence that aims at correcting for mutation rate variation along the genome [34]. This statistic was obtained by

dividing nucleotide divergence between the two haplotypes derived from the Gouldian finch genome by nucleotide divergence between Gouldian finch and zebra finch. Prior to the calculation, the Gouldian sequences were aligned against the zebra finch Z-chromosome using *progressiveMAUVE* [35]. Individual nucleotide divergence values were calculated using the Perl script *calculate-dxy.pl* from the *PoPoolation* toolbox [36]. Estimates were computed in windows of 20 kb with 5 kb steps assuming the Z-chromosome coordinates of the zebra finch genome. Windows with more than 20% missing data were excluded from the analysis.

Structural rearrangements. To search for structural rearrangements between the red and black alleles, we used several methodologies. First, we produced an alignment between the haplotypes associated with each color morph across the candidate region and summarized this alignment using a dot plot. Second, we visualized both the linked-read and the whole-genome resequencing data in *IGV* (Integrative Genomics Viewer)[37]. Finally, we used several structural rearrangement detection methods that we applied both to the linked-read and whole-genome resequencing data. For the linked-read data, we used the *long ranger* toolbox from 10X Genomics. For the whole-genome resequencing data, we merged the reads of all black and red individuals into two larger files (i.e. one per morph) and applied several algorithms that take advantage of multiple aspects of the read data: *BREAKDANCER* [38], *DELLY* [39], and *LUMPY* [40]. *BREAKDANCER* uses discordant read pairs (i.e. read pair orientation and insert size information). *DELLY* uses both discordant read pairs and split-read information (i.e. single continuous reads that map on two genomic locations). *LUMPY* utilizes a combination of discordant read pairs, split-read, and read-depth information.

Gene expression analysis

RNA-sequencing. We generated RNA-seq data from regenerating skin samples of eight individuals, including three black morphs and five non-black morphs (three orange and two red head morphs; Table S3). Orange individuals are identical to red morphs at the *Red locus* and also express carotenoids in their masks (unpublished data). These individuals were kept in experimental facilities under the same conditions until harvesting. All birds were adult males and originated from different breeders. Feather regeneration was induced by plucking small feather patches from the mask region, and new feathers were allowed to regrow for ten days prior to skin excision. Subjects were euthanized by manual cervical dislocation after being rendered unconscious with an anaesthetic gas (isoflurane). Tissues were snap-frozen in dry ice and stored at -80°C.

Total RNA was isolated and purified using the RNeasy® Mini Kit (QIAGEN). An additional RNase-Free DNase® digestion step was performed for more complete removal of contaminating DNA. Following isolation, initial estimates of RNA concentration and purity were made using Qubit® RNA BR assay kit. Illumina libraries were prepared using 1µg of total RNA per sample, following the TruSeq® RNA Sample Preparation Kit v2 protocol and sequenced on Illumina HiSeq 1500 with 2 x 125 bp paired-ends reads.

Sequencing data was submitted to pre-processing steps, including removal of TruSeq Illumina adapters using *Cutadapt* (v1.7.119) [41] and quality filtering with *Trimmomatic*. In *Trimmomatic*, reads were scanned through a 4 bp sliding window and trimmed whenever the average quality dropped below a Phred quality score of 15 (SLIDINGWINDOW=4:15). All reads shorter than 30 bp were discarded. Data quality between each filtering step was assessed with *FastQC*.

Gene expression analysis was conducted using *edgeR 3.4.2* [10,42], as implemented in *TRINITY* (v2.2.0). First, we mapped using *bowtie* (v1.1.0) [43] RNA-seq reads for each individual to a reference transcriptome containing all transcripts identified by the genome annotation. This produced ungapped alignments which were then used to estimate transcript abundances for each individual (TPM, transcripts per million) using *RSEM* [44]. TPM values for each of the eight individuals were used to calculate differential expression using *edgeR*, grouping individuals into two different groups based on head color morph: black (three replicates) vs. non-black (five replicates). We used a False Discovery Rate (FDR) of 0.05 as our significance threshold.

To search for transcript isoforms of the candidate genes we re-mapped RNA-seq reads to the EryGou1.0 genome with *HISAT2*. We then visualized and quantitated exon-exon spanning reads with *sashimi_plot* as implemented in *IGV* (Integrative Genomics Viewer) [45]. We examined the frequency of alternative splicing events in *MOCS2*, between the color morphs, by comparing the proportions of RNA-seq reads spanning exon 1 to 3 relative to reads spanning exon 2 to 3, and exon 3 to 5 relative to reads spanning exon 4 to 3 with a student's *t*-test. We confirmed the alternative splice products of *FST* by PCR amplification with primers located in the 5' and 3'-UTR regions common to both isoforms (Figure S4, Table S2) followed by Sanger sequencing (Eurofins Genomics, Louisville, KY). We examined the abundance of *FST* isoforms between the color morphs by qPCR with primers targeting the exon 5 to 6 junction to amplify isoform X1 or the intron between exons 5 and 6 to amplify isoform X2 following the methods described below and using primers presented in Table S2.

Quantitative PCR. To examine the expression of *FST* and *MOCS2* during feather regeneration we sampled the regenerating skin from the heads of three red and three black males. We plucked small patches of feathers from the mask region of these birds two and four days prior to skin excision. Testis were harvested from four red and four black one year-old adult males purchased from different breeders and then housed together in a single cage for eight days to induce social competition in the group [46]. Harvested tissue was frozen at -80°C, and RNA was later extracted as described above. We generated cDNA from ~1 μg of RNA with the GRS cDNA Synthesis Kit (GRiSP, Porto, Portugal) according to the manufacturer's instructions. We designed primers to target the coding sequences or 3' UTR each isoform of *FST*, both isoforms of *MOCS2*, and *GAPDH* (Table S2).

Primer efficiency was determined by assaying a dilution series of cDNA pooled from all experimental samples. Each of the primers produced a single amplicon as indicated by melt curve analyses and were 99.1-109.1% efficient at the analysis threshold. We measured three technical replicates of each sample with Power Sybr® Green PCR master mix (Life Technologies, 4367659) using an Applied Biosystems StepOne real-time PCR system. We then calculated the mean C_t values among the technical replicates and compared expression of each gene (ΔC_t) relative to *GAPDH* between the morphs with student's *t*-tests.

REFERENCES

1. Reynolds ES. 1963 The use of lead citrate at high pH as an electron-opaque stain in electron microscopy. *J. Cell Biol.* **17**, 208–212. (doi:10.1083/jcb.17.1.208)
2. 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)
3. Brush AH, Seifried H. 1968 Pigmentation and feather structure in genetic variants of the Gouldian finch, *Poephila gouldiae*. *Auk* **85**, 416–430.
4. Stradi R. 1999 Pigmenti e sistematica degli uccelli. In *Colori in volo: il piumaggio degli uccelli ricerca scientifica e cultura umanistica* (eds LC Brambilla, E Giovanni, C.Mannucci, R Massa, N Saino, R Stradi, G Zerbi), p. 182. Milan.
5. Harashlma K, Nakahara J ichiro, Kato G. 1976 Papilioerythrinone: A new ketocarotenoid in integuments of orange pupae of a swallowtail, *papilio xuthus*, and carapaces of a crab, *paralithodes brevipes* (hanasakigani in japanese). *Agric. Biol. Chem.* **40**, 711–717. (doi:10.1080/00021369.1976.10862113)
6. 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)
7. Tarailo-Graovac M, Chen N. 2009 Using RepeatMasker to identify repetitive elements in genomic sequences. In *Current Protocols in Bioinformatics*, p. 4.10.1-14. (doi:10.1002/0471250953.bi0410s25)
8. Bolger AM, Lohse M, Usadel B. 2014 Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. (doi:10.1093/bioinformatics/btu170)

9. Kim D, Langmead B, Salzberg SL. 2015 HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360. (doi:10.1038/nmeth.3317)
10. 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)
11. Grabherr MG *et al.* 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652. (doi:10.1038/nbt.1883)
12. 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)
13. Borodovsky M, Lomsadze A. 2011 Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. *Curr. Protoc. Bioinformatics* **Chapter 4**, Unit 4.6.1-10. (doi:10.1002/0471250953.bi0406s35)
14. Korf I. 2004 Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59. (doi:10.1186/1471-2105-5-59)
15. Stanke M, Waack S. 2003 Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, ii215-ii225. (doi:10.1093/bioinformatics/btg1080)
16. Jones P *et al.* 2014 InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240. (doi:10.1093/bioinformatics/btu031)
17. Hunter S *et al.* 2012 InterPro in 2011: New developments in the family and domain prediction database. *Nucleic Acids Res.* **40**, D306–D312. (doi:10.1093/nar/gkr948)
18. Finn RD *et al.* 2014 Pfam: The protein families database. *Nucleic Acids Res.* **42**, D222–D230. (doi:10.1093/nar/gkt1223)

19. Ashburner M *et al.* 2000 Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29. (doi:10.1038/75556)

20. Lowe TM, Eddy SR. 1996 TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964. (doi:10.1093/nar/25.5.0955)

21. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. 2015 BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212. (doi:10.1093/bioinformatics/btv351)

22. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva E V, Zdobnov EM. 2017 BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* **35**, 543–548. (doi:10.1093/molbev/msx319)

23. Frith MC, Hamada M, Horton P. 2010 Parameters for accurate genome alignment. *BMC Bioinformatics* **11**, 80. (doi:10.1186/1471-2105-11-80)

24. 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.* (doi:10.7287/PEERJ.PREPRINTS.2475V1)

25. Andrews S. 2010 FastQC: A quality control tool for high throughput sequence data. *Http://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/*. , http://www.bioinformatics.babraham.ac.uk/projects/. (doi:citeulike-article-id:11583827)

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

27. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,

Durbin R. 2009 The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079. (doi:10.1093/bioinformatics/btp352)

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. 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)

30. Fumagalli M. 2013 Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS One* **8**, e79667. (doi:10.1371/journal.pone.0079667)

31. Fumagalli M, Vieira FG, Linderoth T, Nielsen R. 2014 NgsTools: Methods for population genetics analyses from next-generation sequencing data. *Bioinformatics* **30**, 1486–1487. (doi:10.1093/bioinformatics/btu041)

32. Fumagalli M, Vieira FG, Korneliussen TS, Linderoth T, Huerta-Sánchez E, Albrechtsen A, Nielsen R. 2013 Quantifying population genetic differentiation from next-generation sequencing data. *Genetics* **195**, 979–992. (doi:10.1534/genetics.113.154740)

33. Solé X, Guinó E, Valls J, Iniesta R, Moreno V. 2006 SNPStats: a web tool for the analysis of association studies. *Bioinformatics* **22**, 1928–1929. (doi:10.1093/bioinformatics/btl268)

34. 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.* **102**, 6573–6580. (doi:10.1073/pnas.0502099102)

35. 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)

36. 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)

37. 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)

38. 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)

39. 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)

40. 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)

41. Martin M. 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12. (doi:10.14806/ej.17.1.200)

42. 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)

43. Langmead B, Trapnell C, Pop M, Salzberg S. 2009 Ultrafast and memory-efficient

403 alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25.
404 (doi:10.1186/gb-2009-10-3-r25)
405 44. Li B, Dewey CN. 2011 RSEM: Accurate transcript quantification from RNA-Seq data
406 with or without a reference genome. *BMC Bioinformatics* **12**, 323. (doi:10.1186/1471-
407 2105-12-323)
408 45. Katz Y *et al.* 2015 Quantitative visualization of alternative exon expression from
409 RNA-seq data. *Bioinformatics* **31**, 2400–2402. (doi:10.1093/bioinformatics/btv034)
410 46. Pryke SR, Astheimer LB, Buttemer WA, Griffith SC. 2007 Frequency-dependent
411 physiological trade-offs between competing colour morphs. *Biol. Lett.* **3**, 494–497.
412 (doi:10.1098/rsbl.2007.0213)
413