Appendix S1 for:

Long-term balancing selection on chromosomal variants associated with crypsis in a stick insect

Dorothea Lindtke, Kay Lucek, Víctor Soria-Carrasco, Romain Villoutreix, Timothy E. Farkas, Rüdiger Riesch, Stuart R. Dennis, Zach Gompert & Patrik Nosil

DOI: 10.1111/mec.14280

Contents

A1 Supplemental Methods and Results

A1.1 Genotyping-by-sequencing (GBS)

For population N1, we filtered raw sequences (minimum read and base quality score 20, minimum read length 50 bp after trimming) and obtained 376 294 982 DNA sequences (mean 865 046 reads per individual with mean read length 83.9 bp). We mapped 94.4% (355 369 015) of reads to the *T. cristinae* reference genome v0.2 (`https://www.ncbi.nlm.nih.gov/nuccore/MSSY00000000.2`; Soria-Carrasco *et al.*, 2014) using Bowtie2 version 2.1.0 (Langmead & Salzberg, 2012) with the '--very-sensitive-local' preset. We called variants using Samtools mpileup and Bcftools versions 0.1.19 (Li, 2011) using the full prior and requiring the probability of the data being homozygous for the reference allele to be less than 0.01. We further discarded variants with low confidence calls of the alternative allele (phred-scaled quality score below 20) and where less than 90% of samples were covered. We retained 304 168 bi-allelic SNPs with mean coverage depth per SNP per individual of ∼5x (per SNP average 2.2–22.7x; per individual average 1.2–7.5x). We then re-assigned scaffolds to the updated *T. cristinae* draft genome v0.3 (Riesch *et al.*, 2017) prior to further analyses.

For population FHA, we called SNPs from the original data set (Comeault *et al.*, 2015) as above, and retained 384 611 bi-allelic SNPs with mean coverage depth per SNP per individual of ∼7x (per SNP average 2.2–16.6x; per individual average 0.8–13.4x). We again re-assigned scaffolds to draft genome v0.3 prior to further analyses.

For the combined data set of 21 populations, we filtered sequences, mapped reads, and called SNPs as above. We obtained 310 066 248 DNA sequences after filtering (mean 849 497 reads per individual with mean read length 83.9 bp) and mapped 94.2% (292 054 983) of reads to draft genome v0.3 (Riesch *et al.*, 2017) using Bowtie2 version 2.2.3 (Langmead & Salzberg, 2012). We retained 626 854 bi-allelic SNPs with mean coverage depth per SNP per individual of ∼5x (per SNP average 2.2–27.2x; per individual average 1.0–10.3x).

A1.2 Identification of genomic clusters

A1.2.1 PCA outlier removal

For populations N1 and FHA, we sequentially removed genomewide outliers similar to Price *et al.* (2006) to eliminate confounding random population structure (e.g., potentially generated by immigrants or their descendants from outside our focal populations, or genetic structure caused by different age classes). In particular, for N1, we removed 48 samples that were more than five standard deviations from their mean on any of the first ten axes, for five iterations; for FHA, we removed 37 samples that were more than six standard deviations from their mean on any of the first five axes, for five iterations. We plotted and visually inspected the rotated data during this process (phenotypes were blinded; Figs. S1 and S2). For N1, non-outlier versus outlier samples were homogeneous with respect to main karyotype ($\chi^2_{(2,N=435)} = 0.69$, p-value $= 0.728$, Cramér's V $= 0.04$) and sex ($\chi^2_{(1,N=357)} = 0.11$, p-value $= 0.859$, Cramér's V $= 0.02$), but not with respect to morph ($\chi^2_{(3,N=409)} = 12.52$, p-value $= 0.006$, Cramér's V $= 0.17$), where fewer than expected green-unstriped but more green-striped and green-incomplete samples were classified as outliers (perhaps because N1 is dominated by *Ceanothus* plants where the green-unstriped morph is more frequent). For FHA, non-outlier versus outlier samples were homogeneous with respect to main karyotype ($\chi^2_{(2,N=600)} = 4.45$, p-value $= 0.114$, Cramér's V $= 0.09$) and morph ($\chi^2_{(3,N=588)} = 2.92$, p-value $= 0.391$, Cramér's V $= 0.07$), but not with respect to sex ($\chi^2_{(1,N=600)} = 5.02$, p-value $= 0.029$, Cramér's V $= 0.09$), where fewer than expected females but more males were classified as outliers. We used the reduced data sets of 387 and 563 non-outlier individuals for all analyses except for determining heterokaryotype excess (although test statistics were unaffected by outlier removal; see section A1.6 for further discussion).

A1.2.2 Structure input file preparation and settings

As the Structure software (version 2.3.4; Pritchard *et al.*, 2000; Falush *et al.*, 2003) cannot directly take genotype likelihoods from Bcftools as input, we first assigned discrete allele states for sites with genotype likelihoods $\geq 95\%$, setting the allele as missing otherwise. We excluded loci with MAF $< 1\%$ and phased our data using fastPhase version 1.4.8 (Scheet & Stephens, 2006, 2008), thereby also imputing missing genotypes. We generated Structure input files from fastPhase output, coding alleles as missing when the posterior probability of the most-likely genotype was $< 90\%$. We then ran Structure using physical distance between loci (in bp) divided by 400 000 as a proxy for cM (Wilfert *et al.*, 2007), conservatively considering different scaffolds as fully unlinked.

To test if individuals from different PCA clusters represent homozygous and heterozygous combinations of main genetic variants, we obtained locus-specific estimates of ancestry for SNPs on LG8. We ran three independent chains for each $k = 2$ or $k = 3$, using the admixture model and correlated allele frequencies with 200 000 iterations as burn-in, followed by 200 000 estimation steps, and by setting priors according to shorter initial test runs that showed the best mixing. As the site-by-site output cannot be pooled among chains, we used only the best chain of each k for further analysis.

To obtain karyotype assignments for tests of HWE and phenotypic differences among karyotypes, and to test if variants are geographically spread, we used SNPs on scaffolds 931, 318, and 1440 and set $k = 2$. We ran three independent chains for $k = 2$ with 50 000 iterations as burn-in and 50 000 estimation steps for populations N1 and FHA, and with 200 000 burn-in and 200 000 estimation steps for the combined data set of 21 populations.

A1.3 Multilocus genomewide association mapping

We used the software Gemma version 0.94 (Zhou *et al.*, 2013) to map colour and pattern traits, as in previous work (Comeault *et al.*, 2015, 2016; Riesch *et al.*, 2017). We ran five independent chains with 2 million burn-in and 6 million sampling steps, applying BSLMMs with the probit model. Parameter states were recorded every 100 and written every 10 000 steps, and all other options were set to default values. BSLMMs in Gemma control for population structure and model all SNPs jointly, thereby accounting for LD. The model provides estimates of total phenotypic variance explained by all SNPs (PVE), the proportion of genetic variance that can be explained by SNPs with measurable effects on phenotype (PGE), and the number of SNPs with a measurable effect (n-SNPs). Further, for individual SNPs Gemma reports the effect size of a SNP (β) and the fraction of MCMC steps it was retained as a SNP with measurable effect (posterior inclusion probability; i.e., the weight of evidence that a SNP has a measurable effect on phenotype).

A1.4 Population genomic statistics

A1.4.1 Relative measure of divergence (F_{ST})

We obtained allele frequency estimates for each of the genomic clusters identified by PCA for N1 and FHA (Fig. 1c; Fig. S4b), using the full sets of 304 168 and 384 611 bi-allelic SNPs, respectively. For each cluster and variant, we first inferred maximum-likelihood allele frequencies from genotype likelihoods using an iterative soft expectation-maximization algorithm (EM) that accounts for uncertainty in individual genotypes (Li, 2011), as in previous work (Riesch *et al.*, 2017). We set the maximum number of EM iterations to 50 and the tolerance for EM convergence to 0.0001. We then used these maximum-likelihood estimates of allele frequencies to calculate Hudson's F_{ST} (Hudson *et al.*, 1992) for non-overlapping 20-kb windows, excluding SNPs with MAF $< 5\%$, as $F_{\mathrm{ST}} = 1 - H_w/H_b$, where $H_w = p_1(1-p_1) + p_2(1-p_2)$ and $H_b = p_1(1-p_2) + p_2(1-p_1)$, where p_j is the allele frequency of a SNP in cluster j. We calculated mean F_{ST} within scaffolds, or F_{ST} per 20-kb window, as ratio of averages (i.e., H_w and H_b are averaged separately across multiple SNPs), as suggested by Weir & Cockerham (1984) and Bhatia *et al.* (2013).

A1.4.2 Nucleotide diversity (π) and absolute measure of divergence (D_{xy})

As calculation of π and D_{xy} requires knowledge of the total number of sites per sequence, we modified variant-calling scripts to output all sites per sequence read that passed filtering. In particular, we used identical filtering thresholds and steps as for SNP calling but retained both variant and invariant sites, discarding sites where less than 90% of samples were covered. We obtained a total of 5 204 739 and 6 426 888 sites for N1 and FHA, respectively, of which in both cases 94% were invariant. Using the maximum-likelihood estimates of allele frequencies for SNPs as above, we calculated nucleotide diversity for non-overlapping 20-kb windows for genomic cluster j and window l as $\pi_{jl} = 2\mathrm{n}\Sigma_i(p_{ij}(1-p_{ij}))/(\mathrm{k}_l(\mathrm{n}-1))$, where p_{ij} is the allele frequency for SNP i in window l and in genomic cluster j, n is the number of sampled chromosomes, and k_l is the total number of nucleotides that passed filtering in a 20-kb window (Nei & Li, 1979). Similarly, we determined absolute divergence between two genomic clusters X and Y with allele frequencies p_{xi} and p_{yi} as $D_{xyl} = \Sigma_i(p_{xi}(1 - p_{yi}) + (1 - p_{xi})p_{yi})/\mathrm{k}_l$ for 20-kb windows. To generate smoothed lines for π along LG8 in Fig. 6, we applied a cubic Savitzky-Golay filter (Savitzky & Golay, 1964) with a filter length of 15 windows using the *savgol* function from the pracma package in R.

A1.4.3 Linkage Disequilibrium (LD) between genomic clusters

We used $\mathrm{Z_g}$ as a measure of between-cluster, intra-locus LD (Storz & Kelly, 2008). We estimated this statistic for 20-kb windows between pairs of homokaryotypic clusters using maximum-likelihood estimates of allele frequencies as above (excluding SNPs with MAF < 1%), and by considering each 20-kb window as a locus. $\mathrm{Z_g}$ measures the between-cluster component of LD and only depends on allele frequencies and their covariance within clusters. The statistic was originally developed to detect loci under spatially varying selection, where peaks in $\mathrm{Z_g}$ are expected to arise when alleles from different subpopulations are under alternative selection pressure (Storz & Kelly, 2008). Here, we calculated $\mathrm{Z_g}$ between pairs of homokaryotypic clusters to investigate the genome for indications of balancing selection that results in divergence between clusters, but not divergence within clusters, akin to subpopulations subject to spatially varying selection.

A1.4.4 Haplotype homozygosity

Using phased genotypes on LG8 for all non-outlier individuals (above), we obtained two summary statistics of extended haplotype homozygosity (EHH), each of which is assumed indicative of recent positive selection (Sabeti *et al.*, 2002). The first is the integrated site-specific EHH (iES) within each homokaryotypic cluster, and the second is the standardized log-ratio of iES between two homokaryotypic clusters (Rsb; Tang *et al.*, 2007). We calculated iES and Rsb using the rehh package in R (Gautier & Vitalis, 2012), and averaged values across SNPs within 20-kb windows. Extreme values for Rsb (positive or negative) are indicative of selective sweeps unique to one of the two tested clusters. Heterogeneous recombination rates along the genome that can affect iES are thus canceled out, as are the effects of recent positive selection shared between clusters (Tang *et al.*, 2007).

A1.4.5 LD within populations

We obtained estimates of Burrow's composite measure of LD (Δ; Weir, 1979) for pairs of SNPs within populations. Δ does not require phased data, but instead provides a joint metric of intra- and inter-gametic components of LD without assuming random union of gametes (Weir, 1979). We determined Δ within populations N1 and FHA. We excluded genetic variants with MAF < 10%, and randomly selected SNPs to achieve at least 100 bp distance among variants, retaining 1 950 or 2 269 SNPs on LG8, and 39 396 or 45 347 SNPs on the remaining LGs for N1 or FHA, respectively. We obtained an estimate of Δ for each pair of SNPs by randomly sampling discrete genotypes from posterior genotype probabilities that were determined by MCMC, as described for PCA (main article), and then computed Δ as the mean over 1 000 iterations. Calculations were implemented in C++, using the GNU Scientific Library (Galassi *et al.*, 2009), as in previous work (Nosil *et al.*, 2012; Gompert *et al.*, 2014). We subsequently fitted models for the decay of Δ with distance r between pairs of SNPs (in bp) as $\Delta = br^z$, using non-linear regression (*nls* function in R; setting regression parameters $b = 0.1$ and $z =$ -0.2 as starting values). We only used Δ and r computed for pairs of SNPs within scaffolds, and then pooled these values for model-fitting for three sets of scaffolds on LG8: (i) scaffolds 931, 318, and 1440; (ii) high-differentiation scaffolds common to all three pairwise combinations of homokaryotypic clusters; and (iii) scaffolds that did not show high-differentiation in any of the three pairwise combinations (i.e., that might not be under balancing selection). For comparison, we additionally fitted models for scaffolds separately pooled for each of the remaining LGs.

A1.4.6 Tajima's D

We computed Tajima's D (Tajima, 1989) within populations N1 and FHA by randomly sampling discrete genotypes from the posterior genotype probabilities determined by MCMC, as described above. We computed the mean value of Tajima's D per 20-kb window over 100 iterations, with the number of invariant sites per window obtained as above. Calculations were implemented in C++, using the GNU Scientific Library. We subsequently summarized estimates for three sets of scaffolds on LG8 (as above), and all scaffolds from the remaining LGs, using the *boxplot* function and the beanplot package in R.

As selective sweeps that differ from the standard full-sweep model can also affect allele frequency spectra in different ways (Przeworski *et al.*, 2005; Hermisson & Pennings, 2005; Pennings & Hermisson, 2006; Coop & Ralph, 2012), we tested whether recent selective sweeps on any of the chromosomal variants might have resulted in an incomplete sweep in the whole population, potentially causing an excess of intermediate-frequency alleles that could increase Tajima's D. Selective sweeps are expected to reduce genetic variation and increase haplotype homozygosity (Vitti *et al.*, 2013). We thus utilized π and iES computed in any homokaryotypic cluster to test whether these statistics are associated with Tajima's D for 20-kb windows within high-differentiation scaffolds. A positive correlation between π and Tajima's D, or a negative correlation between iES and Tajima's D, are consistent with the expected outcomes of a standard sweep, while the opposite would suggest that one or more recent (incomplete) sweeps led to a conflict between these statistics.

A1.5 Divergence dating

A1.5.1 Beast 2

To estimate the approximate divergence times between all three chromosomal variants associated with colour or pattern morphs, we combined sequence reads from 60 *T. cristinae* samples (20 per homokaryotypic cluster from population N1) with data from 80 individuals across four species related to *T. cristinae* (20 individuals from a single population per species: *T. californicum*, population LICK; *T. knulli*, HB; *T. landelsensis*, BCBOG; *T. poppensis*, TBARN; data re-analyzed from Riesch *et al.*, 2017). Filtering, mapping, and variant calling followed the same steps and settings as described above, except requiring coverage for at least 85% of samples for variant calling. From 112 930 670 reads that passed filtering (mean 806 648 reads per individual with mean length of 81.9 bp), 84.7% (95 622 574 reads) were mapped to the *T. cristinae* draft genome v0.3 (Riesch *et al.*, 2017), and 93 519 bi-allelic variants were called with a mean coverage of ∼5x (per variant average 1.8–69.7x; per individual average 2.0–10.9x).

We subsequently used the program BEAST 2 version 2.4.5 (Bouckaert *et al.*, 2014) to estimate divergence times with two subsets of genetic data from LG8, both comprising 140 individuals: (i) 202 SNPs located on scaffolds 931, 318, and 1440; and (ii) 834 SNPs located on those high-differentiation scaffolds that were common to all three pairwise combinations of homokaryotypic clusters. We used custom PERL scripts to generate multiple alignments in NEXUS format from the genotype likelihoods, encoding heterozygotes as IUPAC ambiguities. We included information about invariant positions using 'constantSiteWeights' tags in the XML input files for BEAST 2, so that the effective total number of sites used for analyses was 1 528 and 7 151, respectively. We used a reversible-jump based substitution model approach (Bouckaert *et al.*, 2013), which allows sampling and averaging over a mixture of models (F81, HKY85, TAN93, TIM, EVS, and GTR). We accounted for rate heterogeneity among sites by estimating a gamma distribution of rates and a proportion of invariants. We used an uncorrelated lognormal relaxed clock model (UCLN; Drummond *et al.*, 2006), along with a coalescent Extended Bayesian Skyline tree prior (Heled & Drummond, 2008), which allow accounting for substitution rate variation among lineages and changes in effective population size through time. We set an uninformative proper gamma prior (shape = 0.001, scale = 0.01) on UCLN mean. We used default priors for all other parameters.

We fitted gamma distributions to the posterior distributions of between-species divergence times that were estimated previously (Riesch *et al.*, 2017) with the function *fitdistr* in R (MASS package). We evaluated the joint prior calibration distributions (i.e., effective priors) to assess interactions between the calibrations and the coalescent Extended Bayesian Skyline tree prior (Drummond *et al.*, 2006; Heled & Drummond, 2008; Warnock *et al.*, 2012). We ran one chain for 100 000 000 generations adding the tag 'sampleFromPrior="true"' and sampling parameters every 5 000 steps. We removed the first 85% of samples as burn-in, and compared the 95% highest posterior density intervals obtained to those of the initial gamma prior distributions (Table S10). We then ran, for each of the two datasets, six chains for 200 000 000 generations sampling parameters and trees every 5 000 steps. After confirming stationarity and convergence by visual inspection of trace plots and discarding the first 85% of samples as burn-in, we combined all six chains with LOGCOMBINER, so that parameter estimates were based on 36 000 samples. Effective sample sizes were always above 500 for posterior, likelihood, and divergence times between chromosomal variants. We obtained maximum credibility trees with TREEANNOTATOR and summarized divergence times using the common ancestor tree approach (Heled & Bouckaert, 2013).

A1.5.2 Approximate Bayesian Computation (ABC)

We obtained estimates of divergence time (T) between the two main chromosomal variants (‘melanistic’ and ‘green-unstriped’), using forward-time simulations and all 6 608 sites (including 558 SNPs) from scaffolds 931, 318, and 1440 in N1. We assumed a Wright-Fisher population of constant size N and that all bi-allelic sites were located within a chromosomal inversion. The standard (*St*) and the inverted type (*In*) were assumed to represent the ‘melanistic’ and ‘green-unstriped’ variants, respectively, with dominance of the *In* allele. Recombination among sites was unrestricted within inversion homozygotes but gene exchange between *St* and *In* types was restricted to double-crossover events in inversion heterozygotes.

Simulations were initiated by mutating one *St* allele to an *In* allele by randomly sampling (2N - 1) *St* alleles and 1 *In* allele per site from the observed allele frequencies of the melanistic type, and by setting the carrying capacity for *St* homozygotes (St/St; i.e., melanistic individuals) to N - 1 and that for the inversion carriers (St/In or In/In; i.e., green individuals) to 1. During an initial growth phase, carrying capacity for the inversion carriers increased logistically while that of the *St* homozygotes decreased, until reaching their final values where the frequency of *In* carriers remained constant at value Q.

We used physical distances of the observed data in bp divided by 400 000 as a proxy for cM (Wilfert *et al.*, 2007) and by setting 0.2 cM distance between sites on different scaffolds and between the simulated inversion breakpoints and the first and last site. The total size of the simulated inversion (L) was 3.205 cM. We calculated site-specific flux (ϕ, i.e., gene flow between *St* and *In* types; Navarro *et al.*, 1997) from the probability of observing two crossover events within the inversion, adjusted by the distance of the site to the inversion breakpoints (allowing increased flux in the center of the inversion). We ignored genetic exchange through gene conversion or from more than one double-crossover event per gamete, and also assumed the absence of crossover interference. As we do not know the actual size of the inversion or the relative position of the studied scaffolds to the breakpoints, we scaled site-specific flux by a factor c (reducing or increasing site-specific flux similar to the effect of changing the total size of the inversion). In particular, site-specific flux was calculated as $\phi = \rho x(1-x)/S$, where $x = (0, 1)$ describes the relative position of the site within the inversion, $S = 1/6$ is the integral of $x(1-x)$ for scaling purpose, and ρ is the average flux rate per site, $\rho = \lambda^k e^{-k}/(3k!)$, i.e., ρ is Poisson distributed with number of observed events $k = 2$ and number of expected events $\lambda = c\mathrm{L}/100$, where the 3 in the denominator reflects the expected length of L/3 of the exchanged segment by double-crossovers. The number of alleles exchanged per generation and site from *St* to *In* were then drawn from a binomial distribution with the number of trials given by the total number of *In* alleles, and with probability ϕ times the proportion of *In* alleles found in heterokaryotypes relative to the total number of *In* alleles (or vice versa for *St* alleles).

Sites were allowed to mutate from one to another alternative state with probability μ per generation and site, regardless whether the site was already polymorphic or not.

We placed a log uniform prior on T with lower and upper bounds of 1 x 10^4 and 3 x 10^7, a uniform prior bounded by 0.0 and 1.0 on the site-specific flux scaling parameter c, and a normal prior on Q, bounded to be < 0.95, and with mean 0.80 and s.d. 0.15 (similar to observed frequencies of green *T. cristinae*; Table S1b in Comeault *et al.*, 2015). We restricted our simulations to relatively small N to reduce computation time, but scaled mutation rates accordingly to obtain equilibrium values of nucleotide diversity similar to the observed values. In particular, we placed a log uniform prior on N with lower and upper bounds of 1 x 10^3 and 1 x 10^4, and a gamma prior on μ with scale $\theta = (1 \text{ x } 10^{-10})/(6 \text{ x } 10^{-3})$ and shape $(6 \text{ x } 10^{-3})/(4\text{N}\theta)$.

We calculated summary statistics from allele frequencies estimated for the same number of sampled alleles as in the observed data (i.e., 56 *St* and 72 *In* alleles). We used four summary statistics considered to be informative about population history and that can account for recombination (Wakeley & Hey, 1997; Leman *et al.*, 2005; Becquet & Przeworski, 2007): the number of unique polymorphisms within *St* chromosomal variants ('unique *St*'); the number of unique polymorphisms within *In* chromosomal variants ('unique *In*'); the number of shared polymorphisms across variants ('shared'); and the number of sites that showed fixed differences between variants ('fixed'). We considered a site to be polymorphic if MAF $\geq$ 2%. Summary statistics were calculated over the full genomic region that was included in the simulations.

We performed 332 552 simulations with parameters sampled from their priors. Code for running simulations and calculating summary statistics was written in C++ using the GNU Scientific Library. To estimate the posterior distribution for parameters, we used local-linear ridge regression implemented in the R package abc (version 2.1; Csillery *et al.*, 2012), retaining the 0.5% of samples with summary statistics closest to the observed values. We used 4Nμ as composite parameter, log-transformed T, and logit-transformed c during inferences.

A1.6 Tests for HWE and heterokaryotype excess

We conducted tests for HWE and heterokaryotype excess using all individuals per population, including PCA outliers, as outliers are more likely to decrease but not to increase heterokaryotype excess. Nonetheless, outlier removal had no effect on our results: test statistics for population N1 were $F = -0.1554$ and p-value $= 0.00119$ including outliers, or $F = -0.1674$ and p-value $= 0.00123$ excluding outliers; for population FHA those were $F = -0.1671$ and p-value $= 0.00004$ including outliers, or $F = -0.1689$ and p-value $= 0.00006$ excluding outliers. Excluding outliers slightly decreased F, supporting the conservativism of our approach, while marginally increased p-values likely reflect reduced power of HWE tests due to decreased sample sizes.

A1.7 Mating preference models

An excess of heterozygotes relative to Hardy-Weinberg expectations can result from heterozygote advantage selection, negative assortative mating, or a combination of the two (Hedrick *et al.*, 2016). In addition, selective trade-offs, for example caused by universal mating advantage of one homozygote together with natural selection against this genotype could lead to heterozygote excess for certain parameter combinations. We built two models to parse the conditions under which heterozygote advantage selection versus mating preference is likely to contribute generating the genotype frequencies that we observed. The first model considered heterozygote advantage in combination with negative assortative mating between morphs, and the second model heterozygote advantage in combination with universal mating advantage of one morph (the latter motivated by results of previous mating trials; Comeault *et al.*, 2015).

In both models, we assumed that mating preferences are based on colour phenotype controlled by a single locus with two alleles, melanistic (m) and green (G), with full dominance of the G allele. We further assumed that the colour locus was additionally subject to heterozygote advantage selection with fitnesses $1 - s_1$, 1, and $1 - s_2$ for genotypes mm, mG, and GG, respectively ($0 \leq s_1 \leq 1$ and $0 \leq s_2 \leq 1$). Heterozygote advantage selection is realized when both s_1 and s_2 are nonzero. We note that selection does not necessarily need to act on colour itself for our models to be valid, but could instead result from pleiotropy or tight linkage to a second selected locus.

For the first model, we follow the formulas of Hedrick *et al.* (2016). We define A and $1 - A$ as the proportions of negative assortative mating and random mating, respectively. The proportions of the three different genotypes mm, mG, and GG in the current generation are denoted P, H, and Q, respectively. Assuming Mendelian segregation, the frequencies of zygotes available after mating can be computed from the proportions of genotypes in the current generation as:

$$zP = P^2(1 - A) + PH + \tfrac{1}{4}H^2(1 - A)$$
$$zH = 2PQ + PH + \tfrac{1}{2}H^2(1 - A) + HQ(1 - A)$$
$$zQ = Q^2(1 - A) + HQ(1 - A) + \tfrac{1}{4}H^2(1 - A)$$

In our second model, $1 - A$ specifies the probability of successful mating per green partner within a mating pair formed according to genotype proportions in the current generation, so that the mating probability is 1 for mm x mm pairs, $1 - A$ for mm x mG or mm x GG pairs, and $(1 - A)^2$ for mG x GG or GG x GG pairs. This model should better fit to the conditions in *T. cristinae* where universal mating advantage of the melanistic morph has been observed in experiments (Comeault *et al.*, 2015). Assuming Mendelian segregation as before, the frequencies of zygotes available after mating can be computed as:

$$zP = P^2 + PH(1 - A) + \tfrac{1}{4}H^2(1 - A)^2$$
$$zH = 2PQ(1 - A) + PH(1 - A) + \tfrac{1}{2}H^2(1 - A)^2 + HQ(1 - A)^2$$
$$zQ = Q^2(1 - A)^2 + HQ(1 - A)^2 + \tfrac{1}{4}H^2(1 - A)^2$$

For both models, zygote frequencies need to be adjusted by all progeny produced, $w = zP + zH + zQ$, to obtain genotype proportions in the progeny generation:

$P' = zP/w$

$H' = zH/w$

$Q' = zQ/w$

After incorporating selection on the three genotypes, the frequencies of surviving progeny are:

$sP = P'(1 - s1)$

$sH = H'$

$sQ = Q'(1 - s2)$

The proportion of genotypes available in the next generation is then given by:

$P'' = sP/w'$

$H'' = sH/w'$

$Q'' = sQ/w'$

Where $w' = sP + sH + sQ$.

Using these formulas and numerical simulation, we first computed equilibrium genotype proportions, which were quickly achieved after few generations, for different strengths of A, s_1, and s_2. We then obtained the probability of drawing the observed genotype counts in *T. cristinae* populations from these proportions, given the multinomial distribution. Probabilities were then standardized by the maximum probability that can be theoretically achieved with the observed data.

We performed a grid search for combinations of A, s_1, and s_2 ranging from 0 to 1 with a step size of 0.01. For each setting, we obtained equilibrium genotype proportions by a single simulation. We first draw a starting value for the frequency of the m allele, p, from a uniform distribution between 0.3 and 0.8, and initialized P, H, and Q according to Hardy-Weinberg expectations. Mating and selection occurred as specified above until convergence was reached (i.e., when the change in p from one generation to the next was less than 0.0001). We implemented our calculations in C++ using the GNU Scientific Library. Observed genotype counts were set to the observed karyotype counts obtained by Structure with $k = 2$, as in tests for HWE.

A1.8 Test for phenotypic differences among karyotypes

We excluded individuals that were PCA outliers, had ambiguous phenotype data, or where colour morph did not match expected karyotypic state, such that 539 samples from population FHA with complete observations were retained for analysis. We built linear models for body length and six continuous colour traits as response variables, including either binary colour state (M0) or karyotypic state (M1) as explanatory variable. We further included sex and % striped as covariates in the models. We compared M0 and M1 by analysis-of-deviance and by difference in Akaike's Information Criterion (ΔAIC). Further, to directly determine differences between the two phenotypically green karyotypes ('melanistic-green' and 'green-green'), we repeated the analyses excluding 'melanistic-melanistic' homokaryotypes (retaining 491 complete observations), and compared models excluding (M0) or including karyotypic state (M1), with sex and % striped as covariates. We adjusted *p*-values using the method of Benjamini & Hochberg (1995) and by Bonferroni correction. We used functions *lm*, *anova*, *AIC*, and *p.adjust* from the stats package in R for these calculations.

A2 Supplemental Tables

Table S1: Details about analyzed *T. cristinae* populations. Population N1 is new to this study, population FHA was re-analyzed from Comeault *et al.* (2015), and all other populations were re-analyzed from Riesch *et al.* (2017). N, sample size; Lat, latitude; Long, longitude; Host, host plant (A, *Adenostoma*; C, *Ceanothus*); mm, mG, and GG provide karyotype counts assuming two main variants (m and G) and obtained by Structure analysis with $k = 2$; p, frequency of m; p-value, result from exact test for HWE of karyotype frequencies; F, fixation index for karyotype frequencies; Mela, GrSt, GrIn, and GrUn give counts of melanistic, green-striped, green-incomplete, and green-unstriped morphs; NA, samples with no phenotype information. P-values from tests for HWE remained significant at the 0.05 level for N1 and FHA after Benjamini and Hochberg adjustment or after Bonferroni correction.

Population	N	Lat	Long	Host	mm	mG	GG	p	p-value	F	Mela	GrSt	GrIn	GrUn	NA
R23A	20	34.5185	-120.0764	A	0	6	14	0.15	1.00000	-0.18	0	0	NA	0	20
R12C	20	34.5151	-120.0710	C	1	9	10	0.28	1.00000	-0.13	1	5	NA	14	0
BYA	20	34.5006	-119.8620	A	1	7	12	0.23	1.00000	0.00	1	16	NA	3	0
PRC	19	34.5333	-119.8576	C	4	9	6	0.45	1.00000	0.04	4	0	NA	15	0
OUTA	16	34.5317	-119.8435	A	1	5	10	0.22	1.00000	0.09	1	6	NA	9	0
SC	19	34.5226	-119.8318	C	5	6	8	0.42	0.16073	0.35	5	0	NA	14	0
FHA	600	34.5176	-119.8010	A	55	323	222	0.36	0.00004	-0.17	55	461	21	51	12
MA	16	34.5151	-119.7971	A	2	5	9	0.28	0.53014	0.23	2	13	NA	1	0
N1	435	34.5172	-119.7965	A C	33	223	179	0.33	0.00119	-0.16	31	137	66	175	26
OGA	17	34.5134	-119.7963	A	2	7	8	0.32	1.00000	0.06	4	11	NA	2	0
LA	20	34.5126	-119.7962	A	0	12	8	0.30	0.11878	-0.43	0	19	NA	1	0
OGC	16	34.5134	-119.7961	C	2	11	3	0.47	0.31148	-0.38	2	8	NA	6	0
HVA	20	34.4886	-119.7858	A	8	6	6	0.55	0.08546	0.39	9	10	NA	1	0
PC	20	34.4768	-119.7688	C	7	7	6	0.53	0.19793	0.30	7	3	NA	10	0
ECC35A	19	34.5062	-119.7681	A	2	12	5	0.42	0.35561	-0.30	2	16	NA	1	0
ECCCampA	10	34.5064	-119.7616	A	1	6	3	0.40	1.00000	-0.25	1	8	NA	1	0
ECC20A	19	34.5050	-119.7329	A	0	13	6	0.34	0.04743	-0.52	0	16	NA	3	0
NS1A	19	34.4884	-119.6546	A	3	12	4	0.47	0.37881	-0.27	3	0	NA	16	0
MH19.78C	10	34.5191	-119.2710	C	0	5	5	0.25	1.00000	-0.33	0	0	NA	10	0
MH29.19C	5	34.5554	-119.2632	C	2	2	1	0.60	1.00000	0.17	2	0	NA	3	0
MH25.59C	20	34.5332	-119.2431	C	0	12	8	0.30	0.11878	-0.43	0	0	NA	20	0

Table S2: Cross table summarizing the association between PCA clusters (rows) and phenotypic morphs (columns) for population N1. Mela, melanistic; GrSt, green-striped; GrIn, green-incomplete; GrUn, green-unstriped. NA indicates samples that were below 80% assignment probability for PCA clusters, or that could not unambiguously be scored as a particular morph (NAs were excluded for chi-squared test in main article).

	Mela	GrSt	GrIn	GrUn	NA
mm	25	0	0	0	2
mS	1	78	14	0	3
SS	0	35	3	2	0
US	0	4	22	46	6
mU	0	1	10	82	5
UU	0	0	0	34	2
NA	1	0	5	3	3

Table S3: Cross table summarizing the association between PCA clusters (rows) and phenotypic morphs (columns) for population FHA. Mela, melanistic; GrSt, green-striped; GrIn, green-incomplete; GrUn, green-unstriped. NA indicates samples that were below 80% assignment probability for PCA clusters, or that could not unambiguously be scored as a particular morph.

	Mela	GrSt	GrIn	GrUn	NA
mm	47	5	0	0	1
mS	5	273	1	3	0
SS	2	141	0	2	3
US	0	8	15	26	6
mU	0	3	3	18	1
NA	0	0	0	0	0

Table S4: Genetic architecture of colour as identified by multilocus genomewide association mapping in population FHA. PVE, proportion of total phenotypic variance explained by all SNPs; PGE, proportion of genetic variance that can be explained by SNPs with measurable effects on phenotype; n-SNPs, number of SNPs with a measurable effect. The median values and 95% equal-tail probability intervals are given.

Parameter	Median	2.5	97.5
PVE	0.974	0.872	0.999
PGE	0.967	0.865	0.999
n-SNPs	7	2	18

Table S5: Genetic architecture of pattern as identified by multilocus genomewide association mapping in population FHA. PVE, proportion of total phenotypic variance explained by all SNPs; PGE, proportion of genetic variance that can be explained by SNPs with measurable effects on phenotype; n-SNPs, number of SNPs with a measurable effect. The median values and 95% equal-tail probability intervals are given.

Parameter	Median	2.5	97.5
PVE	0.931	0.696	0.999
PGE	0.888	0.678	0.995
n-SNPs	5	2	16

Table S6: Candidate SNPs associated with colour. LG, linkage group; β, effect size of SNP; PIP, posterior inclusion probability. Seven SNPs with the highest PIPs are shown.

LG	Scaffold	Position	β	PIP
8	318	389554	6.34	0.352
8	2482	25206	8.59	0.300
8	2482	25214	8.11	0.269
8	2482	15150	9.17	0.234
8	318	389549	6.02	0.219
8	2482	15152	8.28	0.197
8	1061	150933	4.39	0.149

Table S7: Candidate SNPs associated with pattern. LG, linkage group; β, effect size of SNP; PIP, posterior inclusion probability. Five SNPs with the highest PIPs are shown.

LG	Scaffold	Position	β	PIP
8	523	382166	-7.36	0.945
4	482	356830	-10.92	0.793
8	1036	131605	-8.19	0.137
8	1036	131573	-5.53	0.116
8	318	207008	-7.23	0.079

Table S8: Phenotypic differences among karyotypes mm, mG, and GG in population FHA. Models either including binary colour state (M0) or karyotypic state (M1) as explanatory variable were compared by analysis-of-deviance or AIC. Sex and % striped were included as covariates in all models. BL, body length; six traits in colour channels: latGB, lateral green-blue; latRG, lateral red-green; latL, lateral luminance; dorGB, dorsal green-blue; dorRG, dorsal red-green; dorL, dorsal luminance. Models including karyotype showed a significant improvement over M0 at the 0.05 level for all traits after Benjamini and Hochberg adjustment, and for traits latRG, latL, and dorGB after Bonferroni correction.

Trait	Model	Res.Df	RSS	AIC	F	p-value	ΔAIC
BL	M0	535	951.09	1845.7			
	M1	534	942.96	1843.1	4.60	0.0323524	2.6
latGB	M0	535	4.49	-1040.4			
	M1	534	4.44	-1045.6	7.15	0.0077289	5.2
latRG	M0	535	0.45	-2284.8			
	M1	534	0.43	-2308.1	25.68	0.0000006	23.3
latL	M0	535	609790.77	5329.4			
	M1	534	597012.00	5320.0	11.43	0.0007755	9.4
dorGB	M0	535	1.40	-1670.9			
	M1	534	1.34	-1690.9	22.30	0.0000030	20.0
dorRG	M0	535	0.19	-2756.8			
	M1	534	0.18	-2758.9	4.14	0.0423909	2.2
dorL	M0	535	700907.38	5404.5			
	M1	534	691823.38	5399.4	7.01	0.0083373	5.0

Table S9: Phenotypic differences between karyotypes mG and GG for green colour morphs from population FHA. Models excluding (M0) or including (M1) karyotypic state as explanatory variable were compared by analysis-of-deviance or AIC. Sex and % striped were included as covariates in all models. BL, body length; six traits in colour channels: latGB, lateral green-blue; latRG, lateral red-green; latL, lateral luminance; dorGB, dorsal green-blue; dorRG, dorsal red-green; dorL, dorsal luminance. Models including karyotype showed a significant improvement over M0 at the 0.05 level for all traits after Benjamini and Hochberg adjustment, and for traits latRG, latL, dorGB, and dorL after Bonferroni correction.

Trait	Model	Res.Df	RSS	AIC	F	*p*-value	ΔAIC
BL	M0	488	858.95	1676.0			
	M1	487	851.01	1673.4	4.55	0.0334405	2.6
latGB	M0	488	4.27	-927.9			
	M1	487	4.21	-933.0	7.13	0.0078283	5.1
latRG	M0	488	0.37	-2123.5			
	M1	487	0.35	-2149.9	29.05	0.0000001	26.4
latL	M0	488	528358.07	4829.1			
	M1	487	515396.33	4818.9	12.25	0.0005086	10.2
dorGB	M0	488	1.33	-1502.5			
	M1	487	1.27	-1521.1	20.85	0.0000063	18.6
dorRG	M0	488	0.12	-2693.8			
	M1	487	0.12	-2698.6	6.78	0.0094900	4.8
dorL	M0	488	606202.16	4896.6			
	M1	487	596561.29	4890.7	7.87	0.0052265	5.9

Table S10: Between-species divergence times used as calibration priors in Beast 2 analyses. Split times were calibrated for clades consiting of *T. californicum* (cali), *T. landelsensis* (land), *T. knulli* (knul), *T. poppensis* (popp), and *T. cristinae* (cris), for which divergence times were estimated previously ('Estimate'; Riesch *et al.*, 2017). Prior shape and prior scale are parameters of gamma distributions fitted to these previously estimated divergence times. Effective prior indicates the intervals of the joint prior calibration distributions obtained from running Beast 2 without data to evaluate the interaction between priors. Posterior estimates are given for data set i (scaffolds 931, 318, and 1440) and data set ii (high-differentiation scaffolds common to all three pairwise combinations of homokaryotypic clusters). For all columns except prior shape and prior scale, 95% highest posterior density intervals are given and numbers are in million years. Note that the reported divergence times are coalescence times of all samples in the clade, as species were not recovered as reciprocally monophyletic in all cases.

Split	Estimate	Prior shape	Prior scale	Effective prior	Set i posterior	Set ii posterior
cali-land	2.65 - 10.37	8.4996	0.667	2.43 - 8.06	3.43 - 9.21	3.82 - 9.37
knul-popp	2.12 - 7.46	10.0184	0.4256	2.18 - 6.70	1.84 - 5.54	1.73 - 5.24
cali-land-knul-popp	5.34 - 18.01	10.8509	0.973	4.21 - 13.17	5.42 - 13.69	5.77 - 13.97
cali-land-knul-popp-cris	14.25 - 36.36	17.8501	1.341	13.34 - 32.71	12.65 - 30.83	12.62 - 30.04

A3 Supplemental Figures

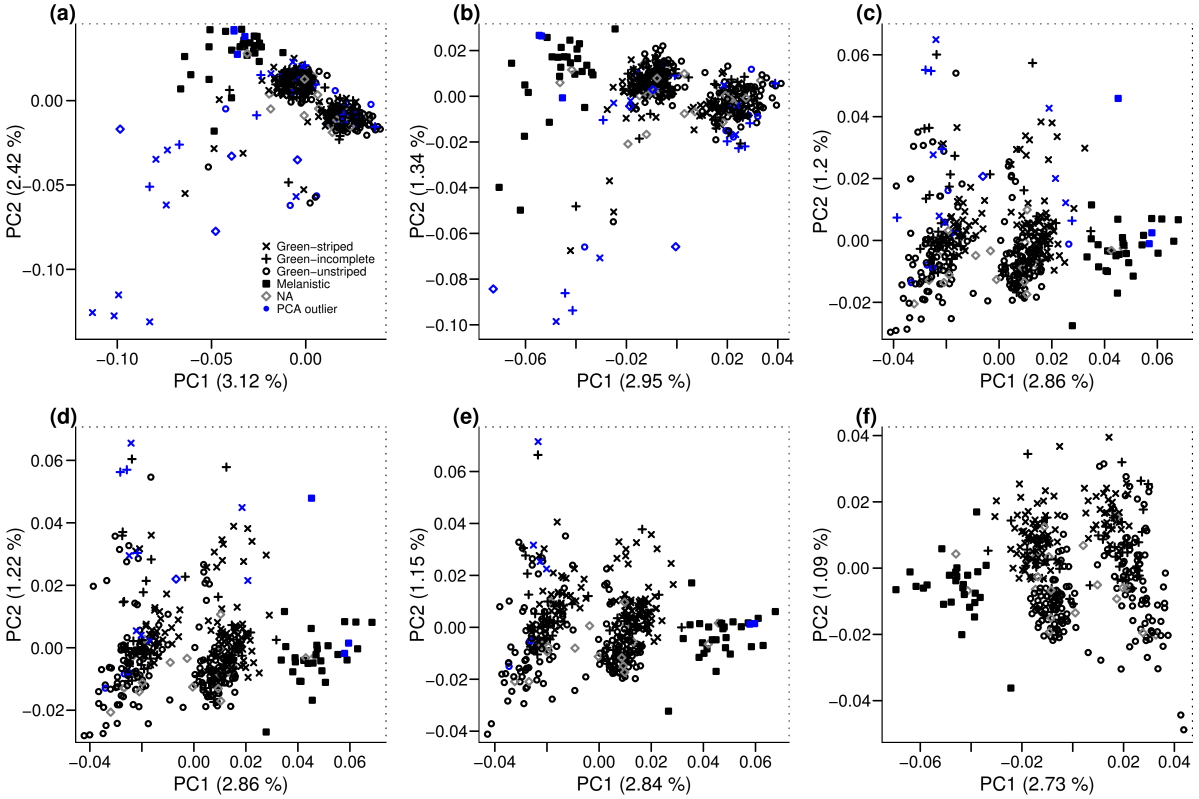

Figure S1: Genomewide structure before, during, and after sequential outlier removal in population N1. Outliers that were excluded for most analyses are indicated in blue. (a) PCA including all 435 samples. (b)–(f) PCAs after each iteration of outlier removal. (f) PCA excluding all 48 outlier samples shows clear clustering associated with colour morphs.

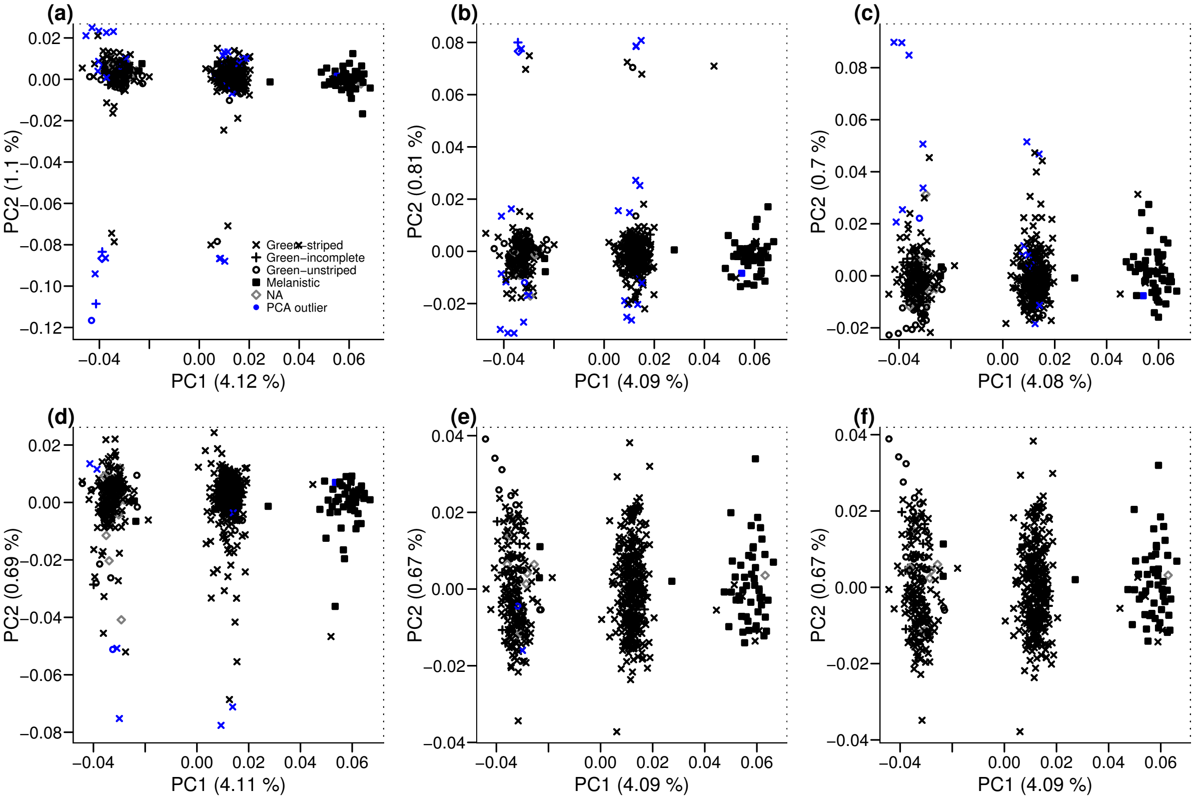

Figure S2: Genomewide structure before, during, and after sequential outlier removal in population FHA. Outliers that were excluded for most analyses are indicated in blue. (a) PCA including all 600 samples. (b)–(f) PCAs after each iteration of outlier removal. (f) PCA excluding all 37 outlier samples shows clear clustering associated with colour morphs.

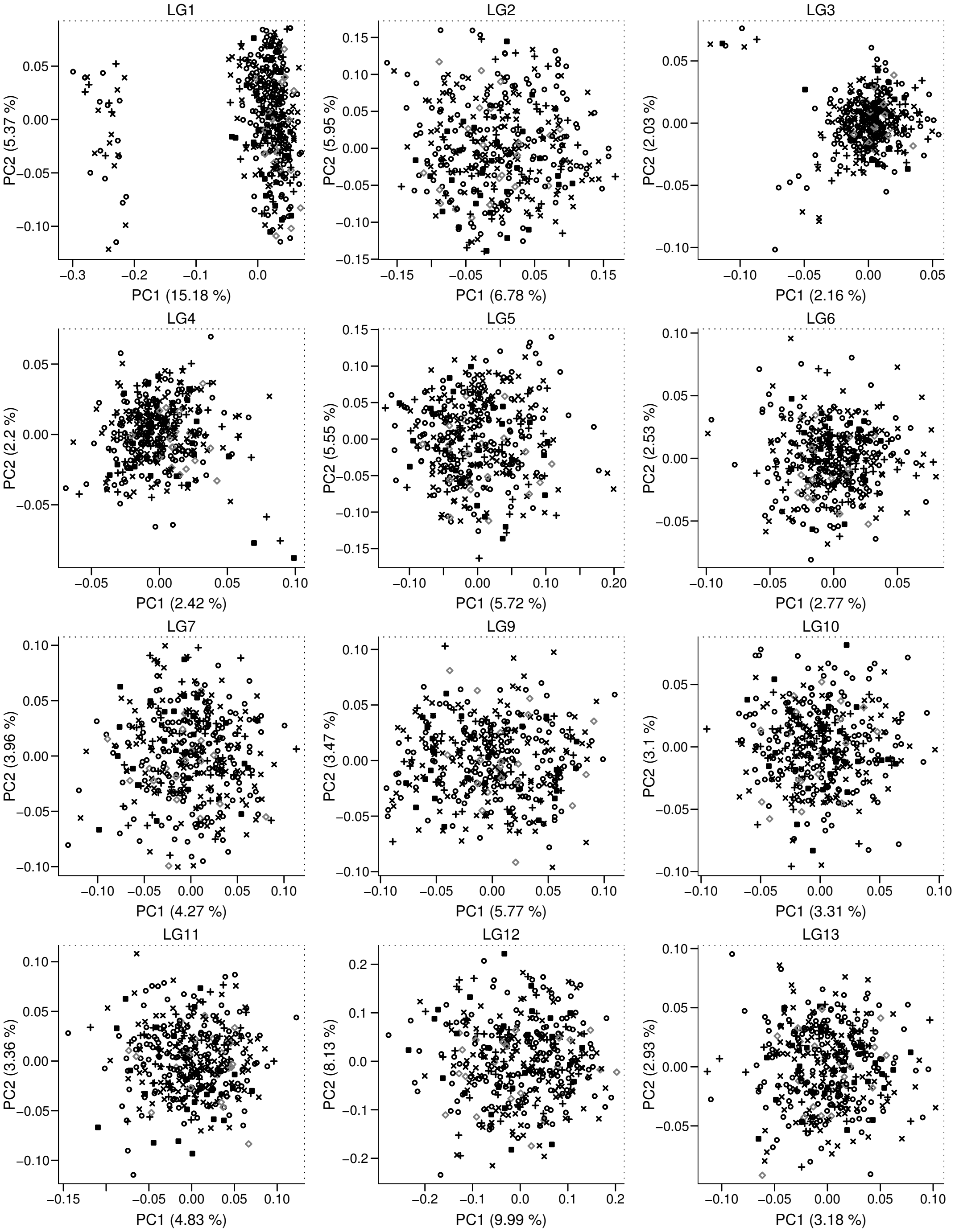

Figure S3: PCA applied to all LGs excluding LG8 in population N1 shows little genetic structure. Clustering on LG1 is not markedly associated with colour morph. Outlier samples were excluded.

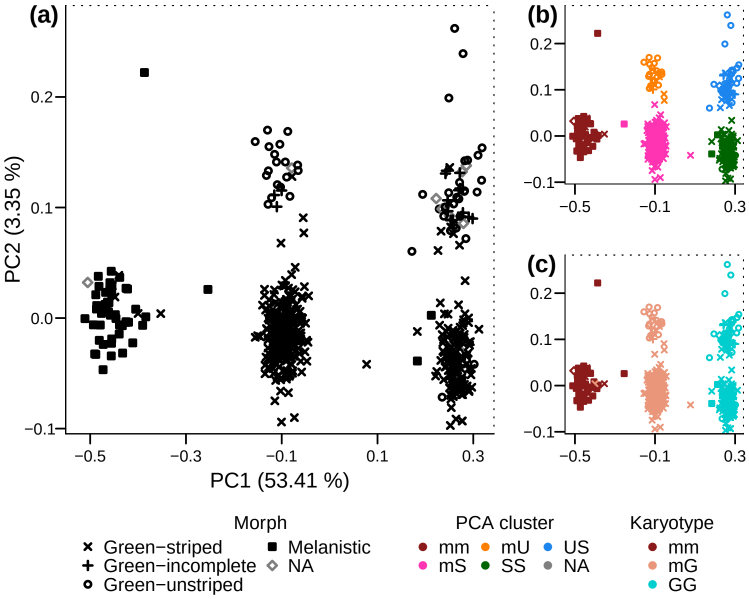

Figure S4: Genetic structure on LG8 in population FHA. (a) Principal component axis one (PC1) shows clustering by colour morph, with one cluster for the melanistic morph (filled squares) and two distinct clusters for green morphs (crosses, pluses, and circles). Principal component axis two (PC2) shows a gradient by pattern morph, from green-striped (crosses) to green-unstriped morphs (circles). (b) K-means clustering and linear discriminant analysis were used to define five PCA clusters (colours), corresponding to diploid combinations of three chromosomal variants m, S, and U. The homokaryotypic cluster UU was not identified by k-means clustering in FHA. (c) Structure with $k = 2$ identified two main chromosomal variants 'melanistic' (m) and 'green' (G), resulting in three main karyotypes by their diploid combinations. PCA outlier individuals were excluded in (a) and (b) and are not shown in (c).

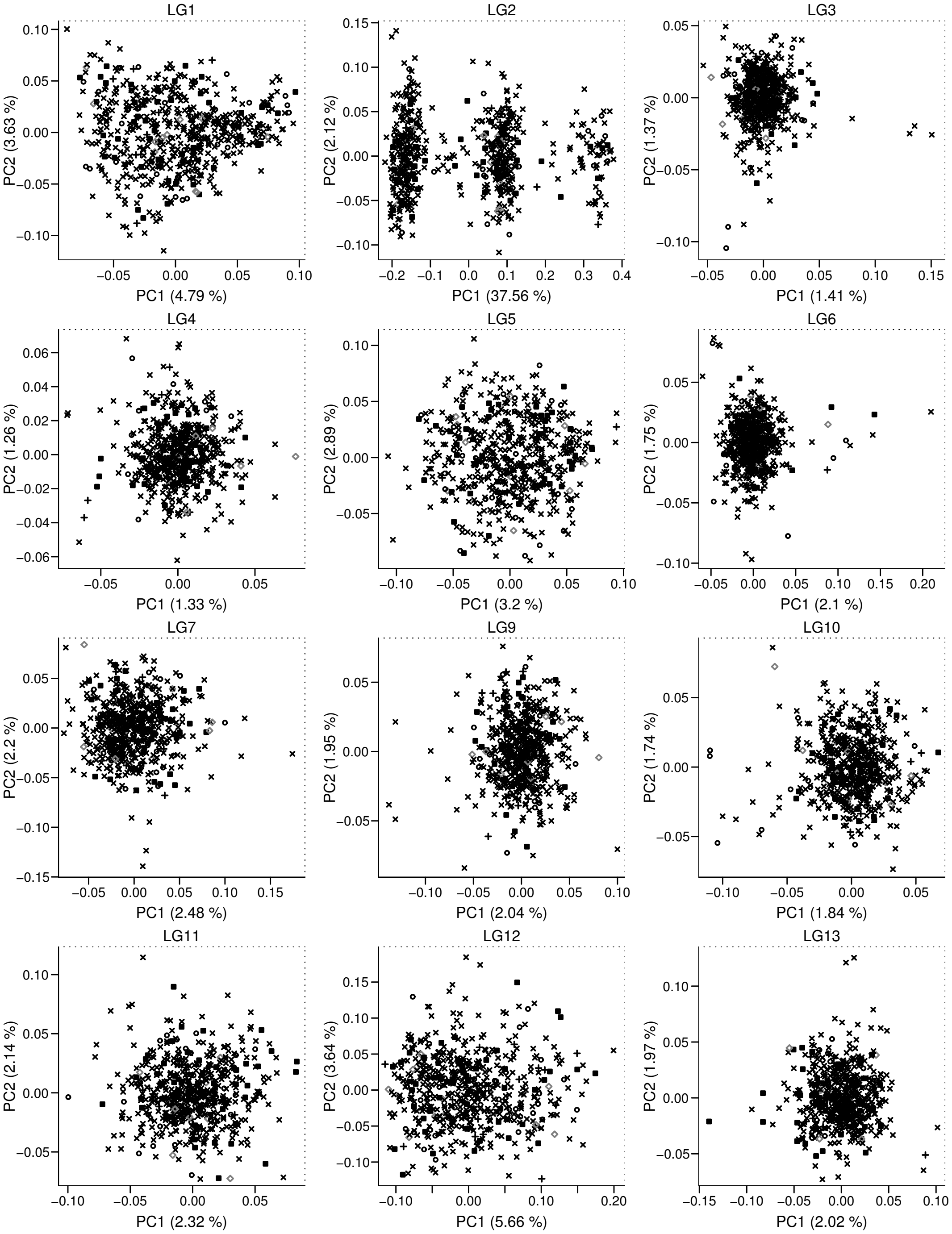

Figure S5: PCA applied to all LGs excluding LG8 in population FHA shows little genetic structure. Clustering on LG2 is not associated with colour morph. Outlier samples were excluded.

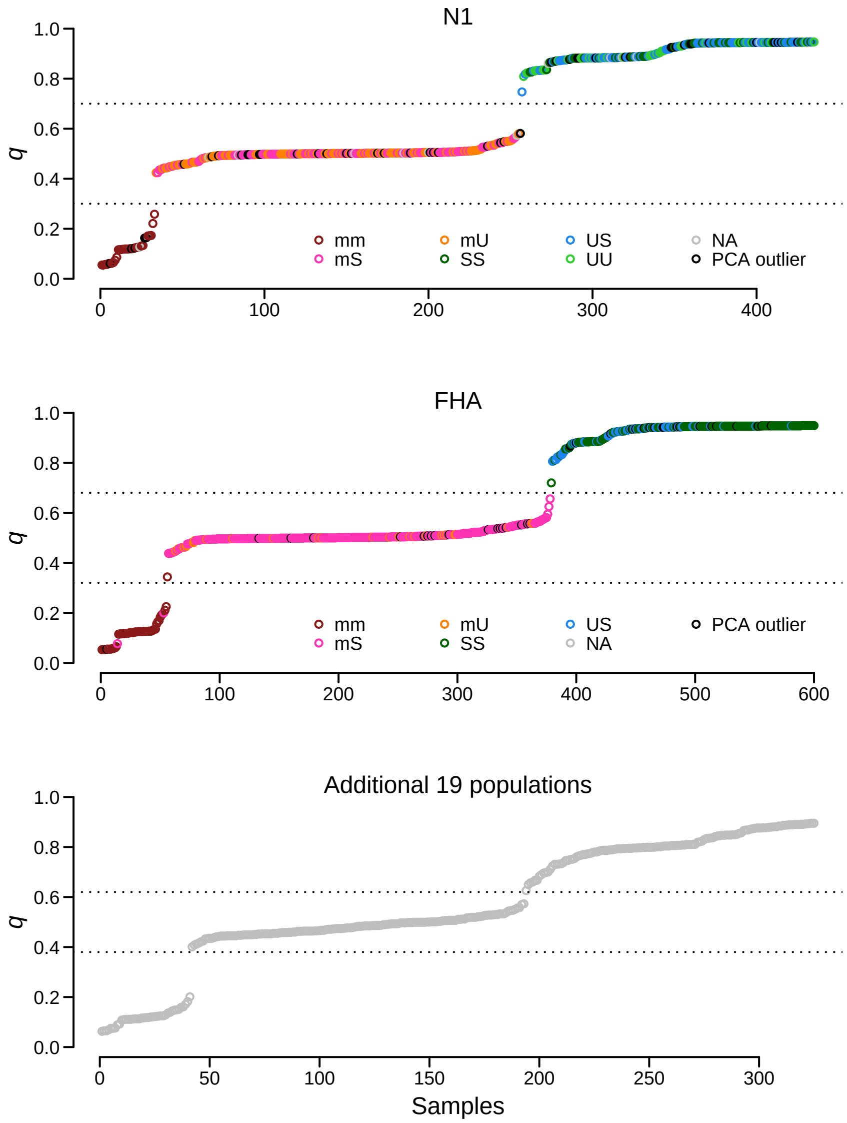

Figure S6: Structure admixture proportions (q) from the linkage model with $k = 2$ for scaffolds 931, 318, and 1440 on LG8. For each data set (N1, FHA, and additional 19 *T. cristinae* populations), homo- and heterokaryotypes of the main chromosomal variants (mm, mG, and GG) were defined by setting thresholds for q that best delimited the three distinct clusters (dotted lines). Colours indicate PCA-based cluster asignments for N1 and FHA.

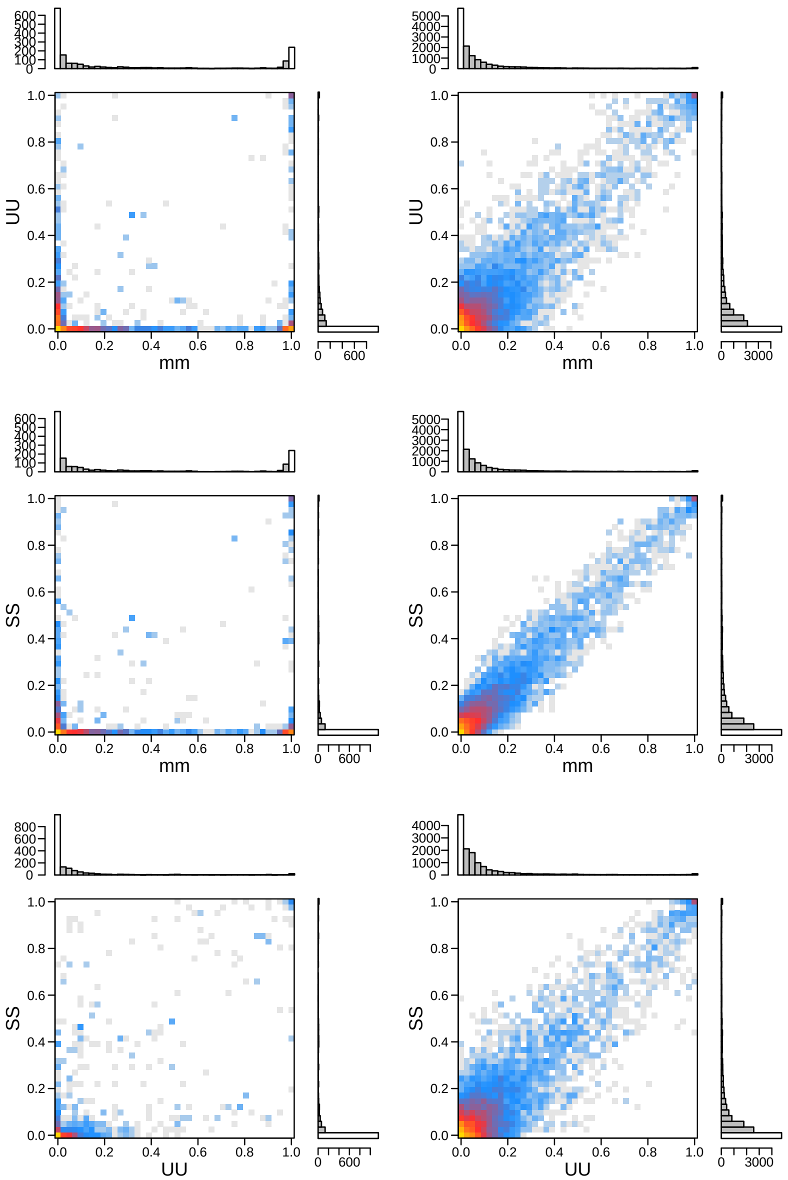

Figure S7: Joint allele frequency spectra for different sets of scaffolds on LG8 and pairs of homokaryotypic clusters in population N1. Corresponding allele frequency spectra for each cluster are shown at the sides of each panel. Left, spectra for SNPs in high-differentiation scaffolds common to all three pairwise combinations of homokaryotypic clusters. Right, spectra for SNPs in scaffolds that did not show high-differentiation in any of the three pairwise combinations of homokaryotypic clusters. Allele frequencies were inferred from genotype likelihoods using an iterative soft expectation-maximization algorithm, and binned with 0.025 increments. Colours indicate the logarithm of the number of sites in each entry of the joint frequency spectrum, ranging from one (grey) to many (gold). Fixed sites are indicated with a white bar in the marginal histograms.

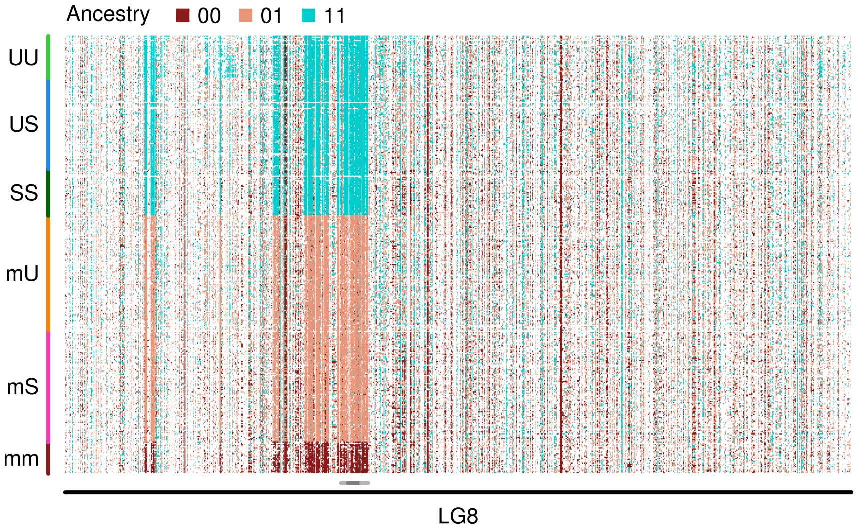

Figure S8: Locus-specific estimates of ancestry on LG8 for $k = 2$ in population N1. Individuals are ordered on the y-axis according to their PCA cluster assignments (Fig. 1c), and the x-axis shows 11 674 SNPs on LG8. Colours indicate diploid ancestry assignments to two clusters, 0 and 1, and are left white when posterior assignment probabilities were less than 90%. Grey horizontal bars on the x-axis indicate positions of three scaffolds shown in more detail in Fig. S9.

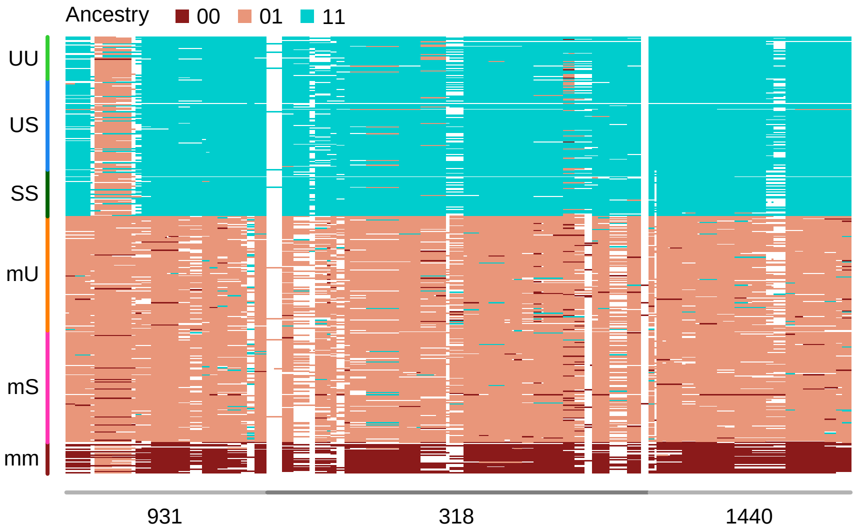

Figure S9: Locus-specific estimates of ancestry for three scaffolds on LG8 for $k = 2$ in population N1. Shown are the same data as in Fig. S8 but for a subset of 403 SNPs from high-differentiation scaffolds that also showed a particularly strong signal of genetic clustering (adjacent scaffolds 931, 318, and 1440). All other elements as in Fig. S8.

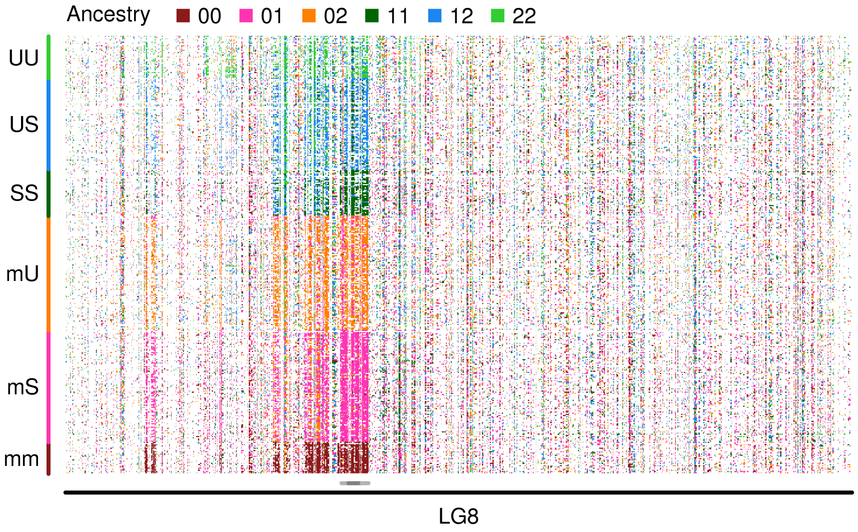

Figure S10: Locus-specific estimates of ancestry on LG8 for $k = 3$ in population N1. Individuals are ordered on the y-axis according to their PCA cluster assignments (Fig. 1c), and the x-axis shows 11 674 SNPs on LG8. Colours indicate diploid ancestry assignments to three clusters, 0, 1, and 2, and are left white when posterior assignment probabilities were less than 90%. Grey horizontal bars on the x-axis indicate positions of three scaffolds shown in more detail in Fig. S11.

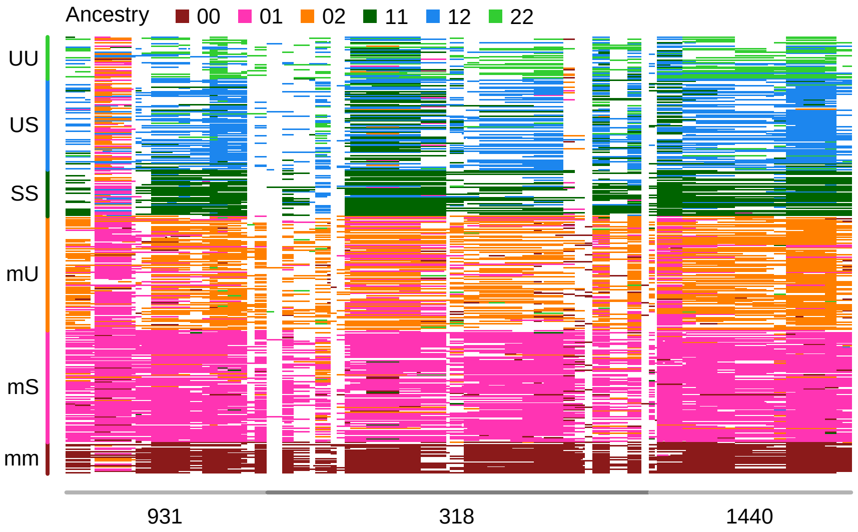

Figure S11: Locus-specific estimates of ancestry for three scaffolds on LG8 for $k = 3$ in population N1. Shown are the same data as in Fig. S10 but for a subset of 403 SNPs from high-differentiation scaffolds that also showed a particularly strong signal of genetic clustering (adjacent scaffolds 931, 318, and 1440). All other elements as in Fig. S10.

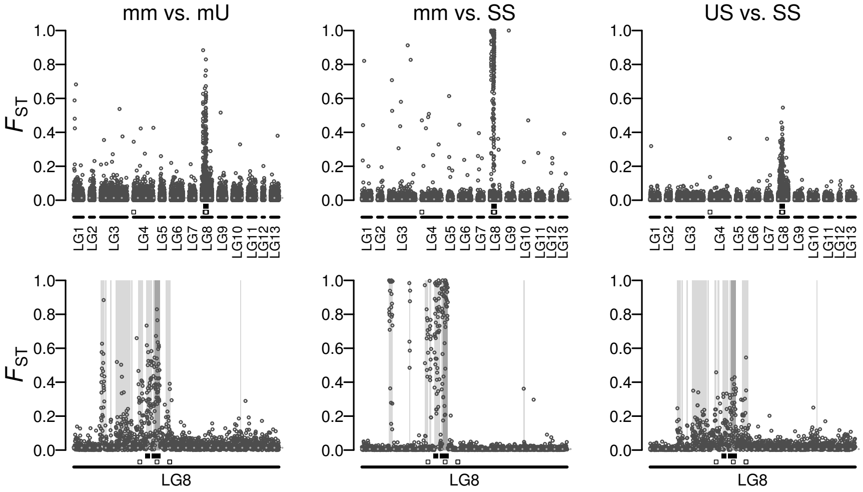

Figure S12: Genomewide differentiation between pairs of clusters in population FHA. Because the homokaryotypic cluster UU was not identified by k-means clustering in FHA, we instead used heterokaryotypic clusters mU and US to calculate F_{ST} between melanistic and green clusters, resulting in lower estimates compared to population N1. Top row, F_{ST} for non-overlapping 20-kb windows and all LGs; bottom row, LG8 only. Grey dotted lines show genomewide 50% quantiles, and squares on the x-axis indicate positions of candidate SNPs for colour (closed symbols) and pattern (open symbols). Scaffolds of high-differentiation in population N1 (Fig. 3) are highlighted in grey.

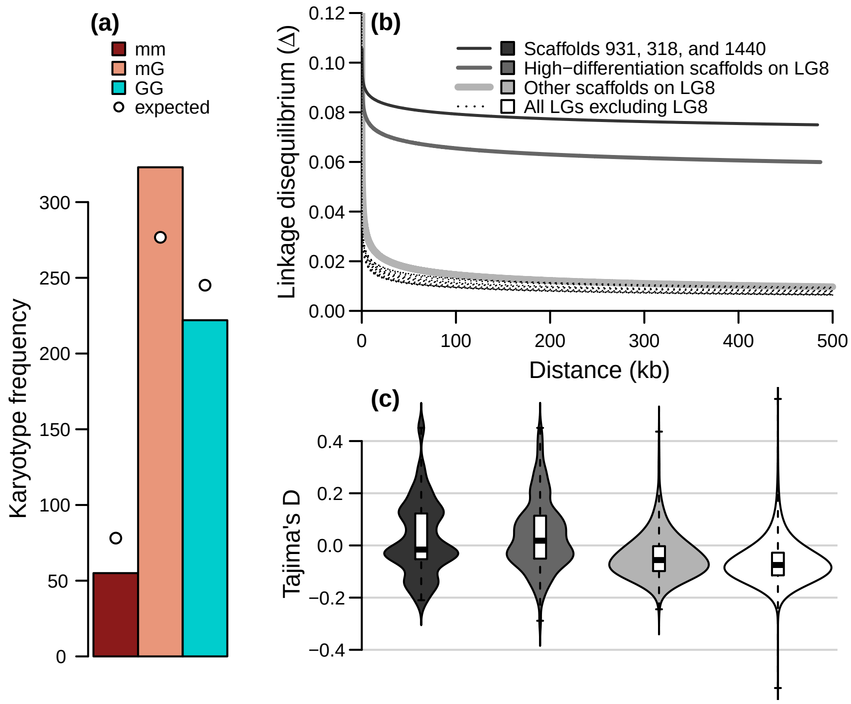

Figure S13: History of balancing selection in population FHA. (a) Karyotype frequencies. Bars show observed karyotype counts, and circles indicate expected counts for a population in HWE. (b) Decay of LD with physical distance between pairs of SNPs summarized for different sets of scaffolds on LG8 and for all other LGs. The y-axis shows Burrow's composite measure of Hardy Weinberg and LD (Δ). Lines were fitted by non-linear regression. (c) Tajima's D statistic for non-overlapping 20-kb windows for different sets of scaffolds on LG8 and for all other LGs combined. White boxes range from the first to third quartile, black horizontal bars give the median, whiskers extend to the data extremes, and shapes are Gaussian kernel densities.

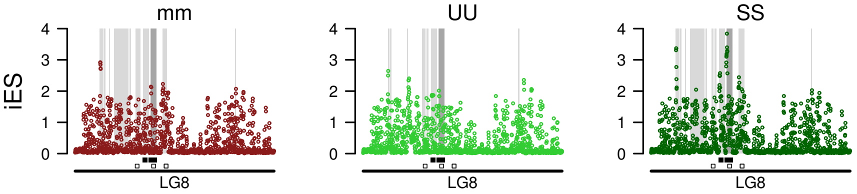

Figure S14: Integrated site-specific extended haplotype homozygosity (iES) along LG8 for each homokaryotypic cluster in population N1, calculated in non-overlapping 20-kb windows. High-differentiation scaffolds are highlighted in grey, and squares on the x-axis indicate positions of candidate SNPs for colour (closed symbols) and pattern (open symbols).

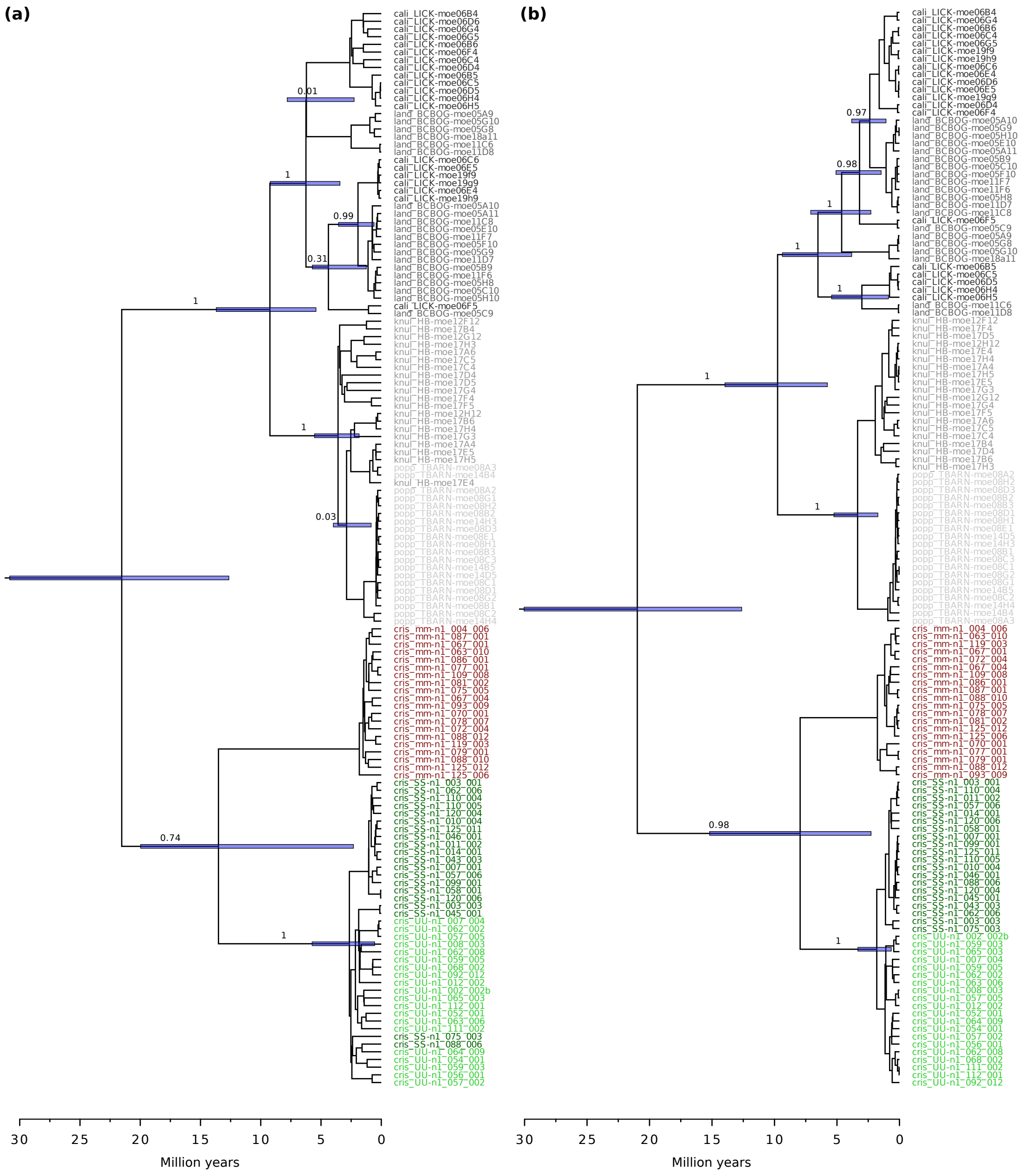

Figure S15: Phylogenetic trees show estimated split times between three chromosomal variants within *T. cristinae* (dark red, m; light green, U; dark green, S) and four species related to *T. cristinae* (greys; cali, *T. californicum*; land, *T. landelsensis*; knul, *T. knulli*; popp, *T. poppensis*). (a) Data set i (scaffolds 931, 318, and 1440). (b) Data set ii (high-differentiation scaffolds common to all three pairwise combinations of homokaryotypic clusters). Numbers on branches give Bayesian posterior probabilities for clade support, and blue bars indicate 95% high posterior density intervals for estimated node ages in million years. Trees were generated using the BEAST 2 software and drawn with FIGTREE v1.4.3 (`http://tree.bio.ed.acuk/software/figtree/`).

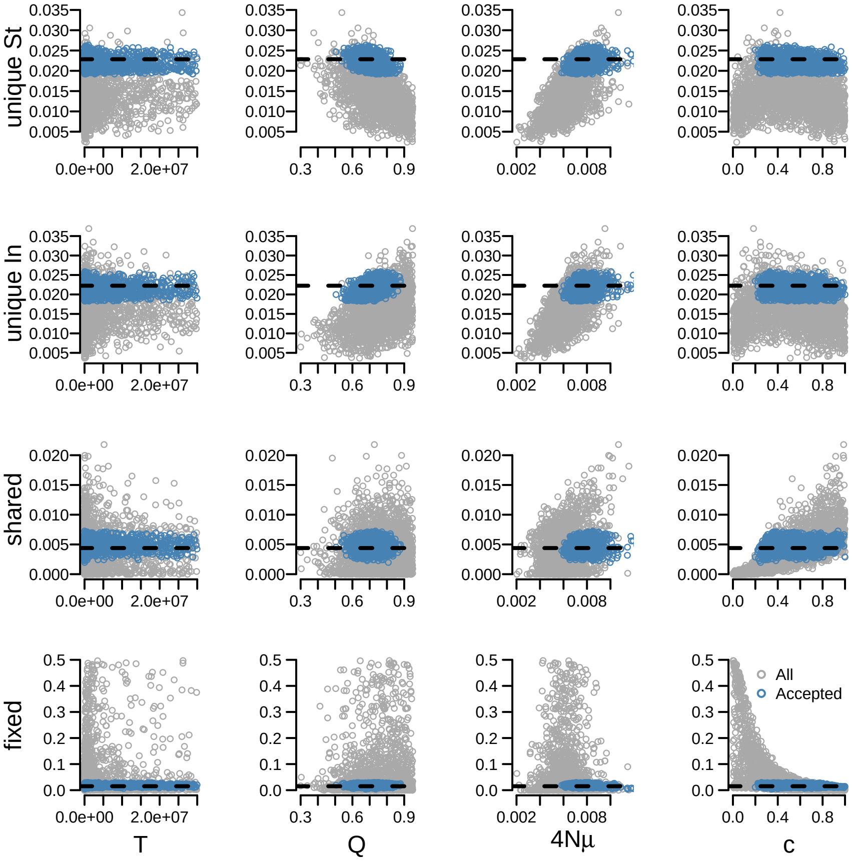

Figure S16: Scatterplots showing relationships between parameter values (x-axes) and summary statistics (y-axes) from ABC used to estimate divergence time between m and U chromosomal variants. Values from 2 000 random simulations are indicated in grey, and all accepted values are shown in blue. Black dashed lines indicate observed summary statistics in N1. T, divergence time (in generations); Q, frequency of *In* carriers; N, population size; μ, mutation rate; c, site-specific flux scaling factor.

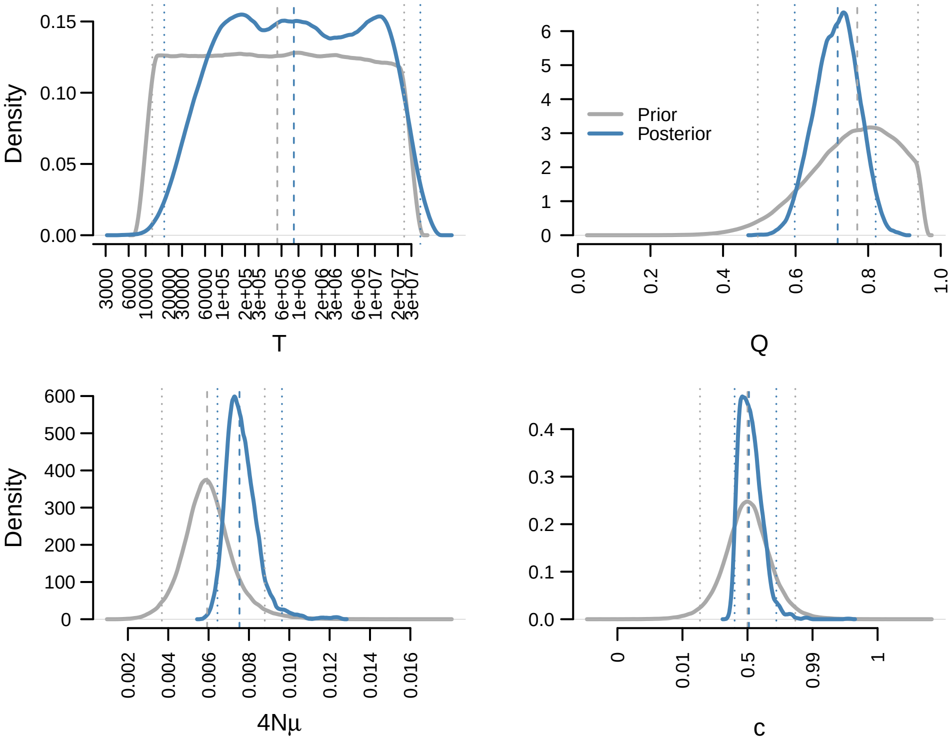

Figure S17: Prior and posterior distributions of four parameters obtained from ABC used to estimate divergence time between m and U chromosomal variants. Dashed and dotted vertical lines indicate median values and 2.5 or 97.5% quantiles of the distributions. T, divergence time (in generations); Q, frequency of *In* carriers; N, population size; μ, mutation rate; c, site-specific flux scaling factor.

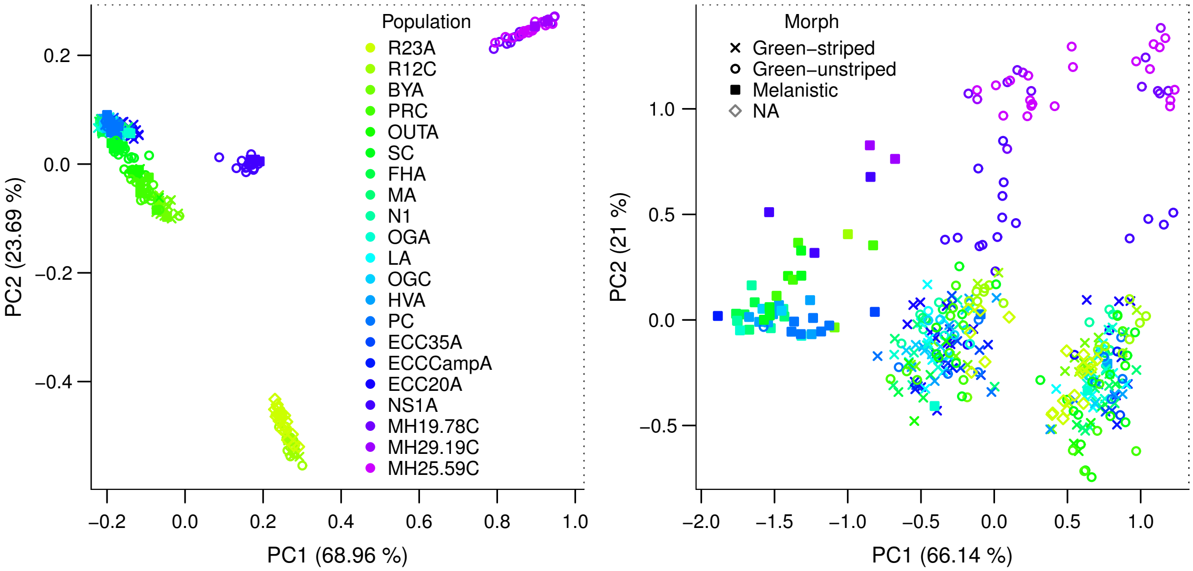

Figure S18: Genetic structure for all 21 *T. cristinae* populations. Populations are coloured by sampling location, from west (green-yellow) to east (purple). Left, PCA applied to all LGs excluding LG8 shows an isolation-by-distance pattern. An approximately horseshoe-shaped pattern for biplots of PC1 versus PC2 is expected for uniform sampling along a one-dimensional habitat (Novembre & Stephens, 2008), roughly corresponding to our sampling scheme (Fig. 2). Right, PCA applied to high-differentiation scaffolds on LG8 shows clustering by colour morph on PC1, with one melanistic and two distinct green clusters. PC2 reflects the west-east gradient of sampling locations. For both PCAs, 20 individuals from each N1 and FHA were included as reference samples.

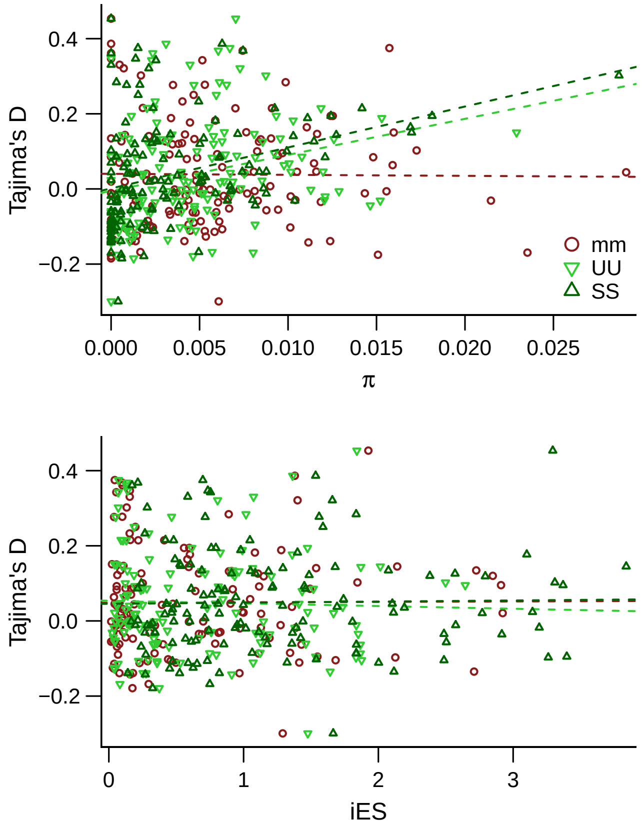

Figure S19: Tajima's D for non-overlapping 20-kb windows within high-differentiation scaffolds as a function of π (top) or iES (bottom) in population N1. Symbols and regression lines are plotted for π and iES computed in each homokaryotypic cluster (mm, UU, or SS). As Tajima's D was estimated for the whole population, each value on the y-axis has three corresponding values on the x-axis. Although we found statistically significant effects of π in clusters UU and SS on Tajima's D, the slope was positive, the opposite of what should be expected if recent selection on one chromosomal variant caused an excess of intermediate-frequency alleles in the whole population that could increase Tajima's D. Effects of π: mm: β = -0.27, p-value = 0.91, adjusted R^2 = -0.007; UU: β = 9.60, p-value = 0.00097, adjusted R^2 = 0.07; SS: β = 10.90, p-value = 0.00002, adjusted R^2 = 0.11. Effects of iES: mm: β = 0.001, p-value = 0.95, adjusted R^2 = -0.007; UU: β = -0.007, p-value = 0.71, adjusted R^2 = -0.006; SS: β = 0.003, p-value = 0.83, adjusted R^2 = -0.007.

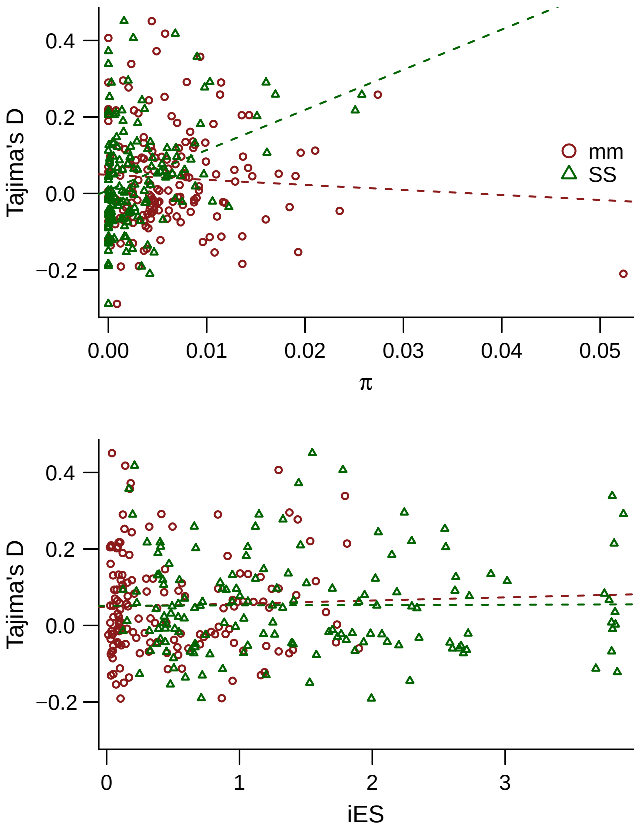

Figure S20: Tajima's D for non-overlapping 20-kb windows within high-differentiation scaffolds as a function of π (top) or iES (bottom) in population FHA. Symbols and regression lines are plotted for π and iES computed in each homokaryotypic cluster (mm or SS). As Tajima's D was estimated for the whole population, each value on the y-axis has two corresponding values on the x-axis. Although we found a statistically significant effect of π in cluster SS on Tajima's D, the slope was positive, the opposite of what should be expected if recent selection on one chromosomal variant caused an excess of intermediate-frequency alleles in the whole population that could increase Tajima's D. Effects of π: mm: β = -1.31, p-value = 0.44, adjusted R^2 = -0.002; SS: β = 10.50, p-value = 0.00001, adjusted R^2 = 0.11. Effects of iES: mm: β = 0.008, p-value = 0.70, adjusted R^2 = -0.006; SS: β = 0.0009, p-value = 0.93, adjusted R^2 = -0.007.

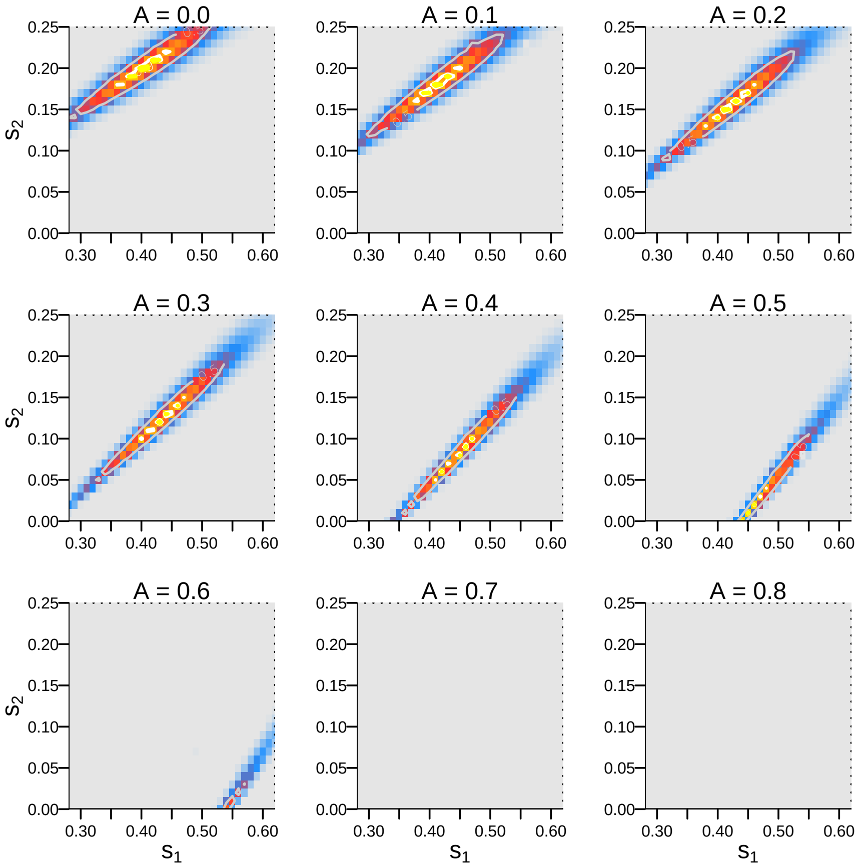

Figure S21: Probabilities of observed karyotype frequencies in population N1 with negative assortative mating between colour morphs. Panels show model results for different strengths of negative assortative mating (A). Strengths of selection against mm homokaryotypes (s_1) and GG homokaryotypes (s_2) are given on the x- and y-axis, respectively. Colours indicate the standardized probabilities of obtaining the observed genotypes with the model parameters, ranging from 0 (grey) to 1 (gold). Contour lines are drawn below probabilites of 0.5 (grey), 0.9 (white), and 0.95 (yellow). Note that the axes are on different scales, and that the x-axis does not start at zero. Results based on observed karyotype frequencies in population FHA were very similar to those shown here.

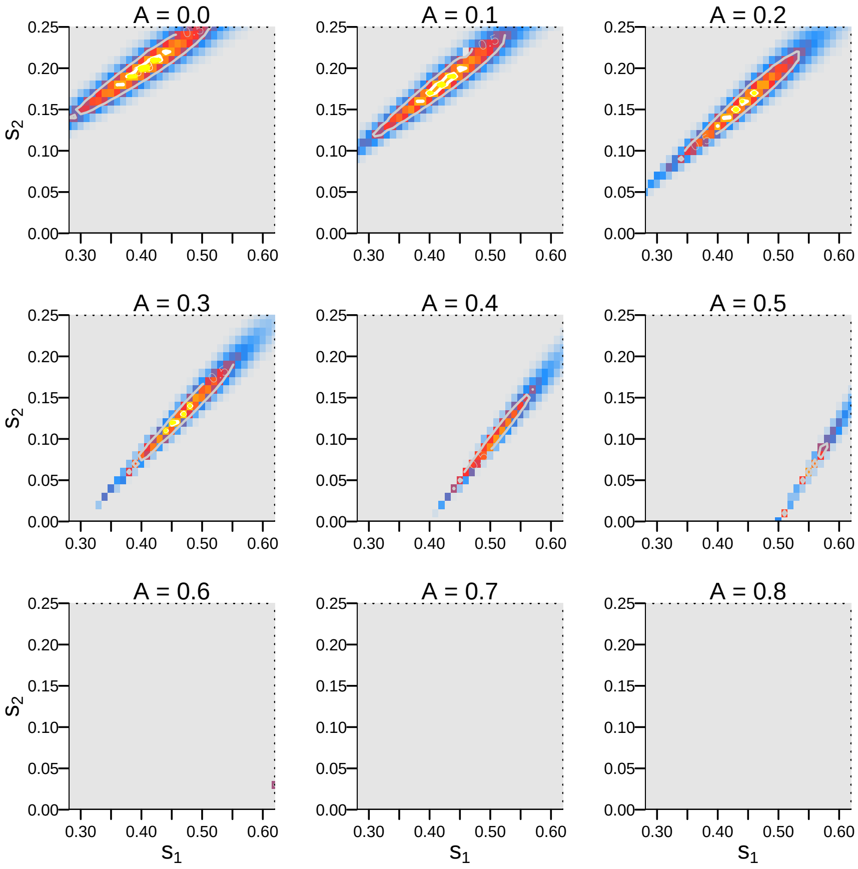

Figure S22: Probabilities of observed karyotype frequencies in population N1 with universal mating advantage of melanistic morphs. Panels show model results for different strengths of universal mating advantage of the melanistic morph (A). All graphical elements are as in Fig. S21. Given results from mating trials in Comeault *et al.* (2015), showing a 16% and 58% increase in mating success for pairs where one or both individuals were melanistic, values of A between 0.20 and 0.27 should provide the best fit. Results based on observed karyotype frequencies in population FHA were very similar to those shown here.

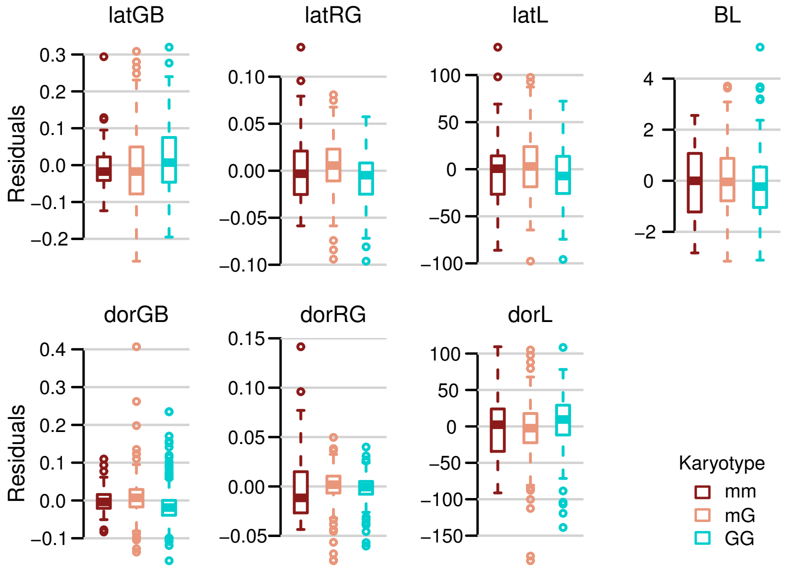

Figure S23: Phenotypic variation among homo- and heterokaryotypes of the main chromosomal variants (mm, mG, and GG). Shown are residual phenotypes from linear models that included binary colour state, sex, and % striped as covariates. The following six traits in colour channels were studied: latGB, lateral green-blue; latRG, lateral red-green; latL, lateral luminance; dorGB, dorsal green-blue; dorRG, dorsal red-green; dorL, dorsal luminance. BL, body length. See Tables S8 and S9 for model results.

A4 Supplemental References

Becquet C, Przeworski M (2007) A new approach to estimate parameters of speciation models with application to apes. *Genome Research*, **17**, 1505–1519.

Bhatia G, Patterson N, Sankararaman S, Price AL (2013) Estimating and interpreting F_{ST}: The impact of rare variants. *Genome Research*, **23**, 1514–1521.

Bouckaert R, Alvarado-Mora MV, Rebello Pinho JR (2013) Evolutionary rates and HBV: issues of rate estimation with Bayesian molecular methods. *Antiviral Therapy*, **18**, 497–503.

Coop G, Ralph P (2012) Patterns of neutral diversity under general models of selective sweeps. *Genetics*, **192**, 205–224.

Csillery K, Francois O, Blum MG (2012) abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, **3**, 475–479.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biology*, **4**, 699–710.

Galassi M, Davies J, Theiler J, *et al.* (2009) *GNU Scientific Library: Reference Manual*. Network Theory Ltd.

Gautier M, Vitalis R (2012) *rehh*: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*, **28**, 1176–1177.

Heled J, Bouckaert RR (2013) Looking for trees in the forest: summary tree from posterior samples. *BMC Evolutionary Biology*, **13**, 221.

Heled J, Drummond AJ (2008) Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology*, **8**, 289.

Hermisson J, Pennings P (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*, **169**, 2335–2352.

Leman S, Chen Y, Stajich J, Noor M, Uyenoyama M (2005) Likelihoods from summary statistics: recent divergence between species. *Genetics*, **171**, 1419–1436.

Nei M, Li W (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, **76**, 5269–5273.

Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, **40**, 646–649.

Pennings PS, Hermisson J (2006) Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS Genetics*, **2**, 1998–2012.

Przeworski M, Coop G, Wall J (2005) The signature of positive selection on standing genetic variation. *Evolution*, **59**, 2312–2323.

Savitzky A, Golay M (1964) Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, **36**, 1627–1639.

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, **78**, 629–644.

Scheet P, Stephens M (2008) Linkage disequilibrium-based quality control for large-scale genetic studies. *PLoS Genetics*, **4**, e1000147.

Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting natural selection in genomic data. *Annual Review of Genetics*, **47**, 97–120.

Wakeley J, Hey J (1997) Estimating ancestral population parameters. *Genetics*, **145**, 847–855.

Warnock RCM, Yang Z, Donoghue PCJ (2012) Exploring uncertainty in the calibration of the molecular clock. *Biology Letters*, **8**, 156–159.

Weir B, Cockerham C (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Wilfert L, Gadau J, Schmid-Hempel P (2007) Variation in genomic recombination rates among animal taxa and the case of social insects. *Heredity*, **98**, 189–197.