Variation and constraints in hybrid genome formation

Anna Runemark[1]*, Cassandra N. Trier[1], Fabrice Eroukhmanoff[1], Jo S. Hermansen[1], Michael Matschiner[1], Mark Ravinet[1], Tore O. Elgvin[1], and Glenn-Peter Sætre[1]

Affiliations:

[1]Department of Biosciences, Centre for Ecological and Evolutionary Synthesis, University of Oslo, PO Box 1066, N-0316, Oslo, Norway.

*Corresponding author. Email: anna.runemark@ibv.uio.no (A.R.).

Summary

Recent genomic investigations have revealed hybridization to be an important source of variation, the working material of natural selection[1,2]. Hybridization can spur adaptive radiations[3], transfer adaptive variation across species boundaries[4], and generate species with novel niches[5]. Yet, the limits to viable hybrid genome formation are poorly understood. Here we investigated to what extent hybrid genomes are free to evolve or whether they are restricted to a specific combination of parental alleles by sequencing the genomes of four isolated island populations of the homoploid hybrid Italian sparrow *Passer italiae*[6,7]. Based on 61 Italian sparrow genomes from Crete, Corsica, Sicily and Malta, and 10 genomes of each of the parent species *P. domesticus* and *P. hispaniolensis,* we report that a variety of novel and fully functional hybrid genomic combinations have arisen on the different islands, with differentiation in candidate genes for beak shape and plumage colour. There are limits to successful genome fusion, however, as certain genomic regions are invariably inherited from the same parent species. These regions are overrepresented on the Z-chromosome and harbour candidate incompatibility loci, including DNA-repair and mito-nuclear genes; loci that may drive the general reduction of introgression on sex chromosomes[8]. Our findings demonstrate that hybridization is a potent process for generating novel variation, but variation is limited by DNA-repair and mito-nuclear genes, which play an important role in reproductive isolation and thus contribute to speciation.

Introgressive hybridization can transfer adaptive genetic variation across species boundaries[4] and generate new species[2,4,5]. Although historically thought to be unimportant in animals, the genomic revolution has shown hybridization to be a pervasive evolutionary force, common in both plants and animals,[1,2] and one that has even shaped the genome of our own species[8]. Portions of the genome vary in extent of introgression[4,9], and homoploid hybrid genomes are predicted to have unequal parental contributions[10]. However, how hybridization, and subsequent recombination and selection mold genomes is not well understood. For instance, it is not known what determines the relative contribution of each parent and whether the genomic locations of introgression are subject to constraints or are largely stochastic. Moreover, genomic constraints in hybridizing taxa can be used to identify genes involved in reproductive isolation between species as they do not introgress[7] , and hence are important for our understanding of speciation and how biodiversity arises. One well-supported finding is reduced introgression on sex chromosomes, which is common in species where one sex is heterogametic[8,11]. To what extent this pattern is caused by specific types of genes on sex chromosomes is debated[12], and knowledge of the genes causing incompatibilities and reproductive isolation is needed to resolve this.

To investigate whether certain genes must be inherited from a specific parent species to form a stable and functional hybrid genome and whether divergent genomes can arise from hybridization, genome-wide data from multiple independent hybrid populations are required. We studied isolated island populations of the homoploid hybrid Italian sparrow *Passer italiae*[6,7] (Fig. 1a) to determine the constraints on, and variability of, genome regions and gene categories. The Italian sparrow genome is a mosaic of the genomes of its parent species, the house *P. domesticus* and Spanish *P.*

hispaniolensis[13] sparrows. The species is thought to have originated when house sparrows colonized the Mediterranean less than 10,000 years ago[6,14]. Morphologically divergent populations of Italian sparrows are found on the Mediterranean islands of Crete, Corsica, Sicily and Malta (Fig 1a; [15]) and we used these isolated island populations to investigate genome-wide patterns of divergence and differentiation from the parent species and within the hybrid species (Supplementary Tables 1-2). We sequenced 10-21 Italian sparrow genomes from each island and 10 genomes from each of the parent species to 6-16X coverage as well as a tree sparrow (*P. montanus*) as an outgroup and aligned them to the recently *de novo* assembled house sparrow reference genome[13].

To determine if these Italian sparrow island populations were genomically differentiated, we first used a principal component analysis. Our results show strong support for differentiation among the hybrid island populations and from the parent species (Fig 1b; Supplementary Fig. 1; Supplementary Tables 3-4). The Italian sparrow populations differ in position along the axis of divergence of the parent species, with Crete and Corsica closer to house sparrow and Sicily and Malta closer to Spanish sparrow (Supplementary Table 5). To further investigate population differentiation, we assessed the most likely number of genetic clusters in the data, and the individual probability of belonging to these clusters using a structure analysis. We found support for two clusters (Supplementary Table 6) corresponding to the parent species and intermediate assignment probabilities for the Italian populations, with Crete and Corsica having the highest probabilities of clustering with house sparrow and Sicily and Malta the highest probabilities of clustering with Spanish sparrow (Fig. 1c). The assignment probabilities differed significantly among populations (ANOVA

$F_{3,47}$=736.54, *P*=2.2e-16; Supplementary Table 7), and across the genome (Supplementary Fig. 2). Moreover, in the most frequent phylogenetic tree topology, house sparrow, Crete and Corsica form one cluster and Spanish sparrow, Malta and Sicily the other, with Crete and Malta clustering most closely with parent species (Fig. 1d and Supplementary Fig. 3). The phylogenetic clustering varied spatially over the genome (Fig. 1e; Supplementary Table 8). We also found significantly higher Spanish sparrow introgression in the Sicily and Malta populations (Patterson's D: ANOVA $F_{3,115}$=22.52, *P*<0.001; Supplementary Table 9). The non-recombining mitochondrial DNA was similar to that of house sparrow for all Italian sparrows, with the exception of one Corsican individual having Spanish sparrow mitochondrial DNA, and two Maltese individuals appearing to have both house and Spanish sparrow mitochondria (Fig. 1f). This is consistent with heteroplasmy, as previously reported for mainland Italian sparrows[13].

To address whether differentiation within a hybrid species is likely to result from positive selection or if differentiated areas mainly arise as a by-product of long-term linked background selection in low recombination areas[16], we first investigated whether differentiation among Italian populations correlates with that between the parent species. We then tested whether variation in differentiation reflects variation in recombination rate, as has previously been shown in flycatchers[16]. Differentiation among the Italian sparrow populations was not strongly correlated with differentiation between the parents, (Supplementary Fig. 4). The standardized slope (beta) of the relationship between within-Italian differentiation and recombination rate was much shallower than that between the parent species differentiation and recombination rate (Supplementary Fig. 4; Tables S10-S11), and we therefore find no strong evidence for

111 house sparrow recombination rate accounting for the genomic heterogeneity observed
112 within this hybrid species. Sorting of parental variants, potentially due to Hill-
113 Robertson effects[17], may have contributed to this reduction in differentiation in low
114 recombination regions compared to that of the parent species. Regions with high
115 differentiation may instead reflect divergent selection, or, alternatively, the
116 recombination landscape may be altered following hybridization and spur
117 differentiation in other areas of the genome in the hybrid than in its parents.
118
119 To identify regions under divergent selection in the hybrid Italian sparrow we
120 extracted the windows where the island populations were most divergent (measured
121 as relative similarity to the parent species in terms of F_{ST}). Among the genes found in
122 the 1% windows that are most divergent between the island populations, 43 different
123 gene ontology (GO) categories were overrepresented relative to the rest of the
124 genome (Supplementary Table 12). Among these categories were genes related to
125 neuron function, including nervous system development, transmission of nerve
126 impulse, and synaptic transmission (Supplementary Table 12), suggesting that these
127 categories of genes have been under divergent selection. Interestingly, genes related
128 to neuron function have also been targets of recent positive selection in great tits[18].
129 The outliers also included FGF10, a gene explaining beak shape divergence between
130 Darwin's finches[19]. Sicily and Crete were strongly divergent at this beak shape
131 candidate gene and the Italian sparrow has previously been shown to exhibit adaptive
132 beak shape divergence[20] (Fig. 2a-c). A gene involved in feather development[21] and
133 melanogenesis[22], wnt7A, was also a highly differentiated outliers among the
134 otherwise genetically very similar (mean F_{ST}=0.016) but plumage-wise divergent
135 Sicilian and Maltese populations[15] (Fig. 2d-f and Supplementary Table 12). These two

136 examples underscore that repeated hybridization between the same parental species
137 can generate locally adapted populations through reshuffling of parental alleles at
138 biologically important loci.
139
140 To identify areas of unique Italian sparrow evolution, we targeted regions in which
141 the Italian sparrow populations have diverged from both parents by extracting the 1%
142 of windows exhibiting the largest differences in F_{ST} between each Italian/parent
143 comparison, only keeping windows overlapping between hybrid/parent in both
144 comparisons. We found 12 overrepresented GO categories (Supplementary Table 13)
145 including circadian rhythm, entrainment of circadian clock and rhythmic process:
146 these genes showed strong signals of stabilizing selection (Fig. 2g), and elevated
147 linkage disequilibrium (Supplementary Table 14). These results illustrate how
148 population specific selection in concert with the parental mosaic is able to form
149 unique features in the genomes of hybrid populations.
150
151
152 The level of differentiation between the Italian sparrow and the parent species is
153 elevated on the Z-chromosome compared to autosomes (Paired t-test; t_3=-8.40;
154 P=0.004; Supplementary Fig. 5-6). This is expected based on the lower effective
155 population size of this sex chromosome and hence elevated rates of genetic drift[23].
156 However, increased differentiation on the Z-chromosome is also expected from the
157 faster X(Z) effect of elevated rates of adaptive evolution on the macro sex
158 chromosome due to hemizygous exposure[24]. Outlier loci are strongly overrepresented
159 on the Z-chromosome (Fig. 2h; all P's<0.001; Supplementary Table 15), except for
160 the outliers in the category where Italian populations invariably had inherited Spanish

161 sparrow alleles, none of which resided on the Z. Interestingly, Tajima's D estimates
162 for the Z-chromosome have significantly higher variance than those for the autosomes
163 (Fig 2c; Repeated measures ANOVA $F_{1,3}$=56.94; P=0.005; Supplementary Fig. 7-8)
164 and dn/ds was higher on the Z-chromosome compared to autosomes (goodness of fit
165 P<0.001 for fixed differences against both house and Spanish sparrows;
166 Supplementary Table 16), supporting a role for selection driving strong Z-
167 chromosome divergence.
168
169 Across taxa with heteromorphic sex chromosomes, introgression on sex chromosomes
170 is reduced[8,11]. To detect loci potentially important in causing such reduction in
171 introgression on the sex chromosomes, we identified regions invariably inherited from
172 a specific parent across all populations. We summed the F_{ST} against house sparrow
173 across island populations and subtracted the summed F_{ST} against Spanish sparrow
174 before extracting the extremes at both ends of the distribution (the 2% of the windows
175 with squared values most diverged from 0; Supplementary Table 17). We found
176 strong evidence for genes invariably inherited from house sparrow, especially on the
177 Z-chromosome. DNA damage stimuli were significantly overrepresented among these
178 genes ($P_{DNArepair}$=0.026). There were 11 mitonuclear loci and although these were not
179 generally overrepresented ($P_{mitonuclear}$=0.11), 7 were found in the areas on the Z-
180 chromosome strongly constrained to house sparrow inheritance, including the
181 previously identified candidate incompatibility gene HSDL2[7] (Supplementary Table
182 18; Fig. 3). There were also 6 DNA mismatch-repair genes on the Z-chromosome,
183 among them the candidate incompatibility gene GTF2H2[7] which is involved in
184 nucleotide excision repair (Fig. 3). This suggests that hybrid genome formation is
185 restricted to uniparental inheritance for these gene classes.

Whereas mito-nuclear genes are known as drivers of reproductive isolation[7] and are expected to be under selection to interact with the frequent house sparrow-like mitochondria, the role for DNA repair genes is less established. However, reduced DNA repair functioning[25] has been found in *Xiphophorus* fish hybrids, and the mismatch repair systems have been shown to contribute to meiotic sterility and cause incompatibilities in yeast[26]. Hence, multigenic DNA repair pathways may need parent specific inheritance to function. As most of these outlier genes were located on the Z-chromosome, they may contribute to the pattern of reduced introgression on sex chromosomes[7].

Our comparison of isolated homoploid hybrid populations formed from the same parent species combination reveals that hybridization can produce diverged genomes with a range of different proportions of parental contribution. More outlier genes exhibited invariable inheritance from house sparrows than from Spanish sparrows. This suggests that parts of the genome must be inherited exclusively from one of the parent species, while the rest of the genome may vary with respect to parent species inheritance. Purging of Dobzhansky-Muller incompatibilities[27] has been suggested to be important for shaping hybrid genomes. We find that house sparrow DNA-repair genes and mito-nuclear genes are necessary for ”escaping the mass of unfit recombinants”[28], most likely due to epistatic interactions. Genes invariably inherited from the Spanish sparrow are fewer but include a candidate pigmentation gene WNT4[29] and a gene involved in vision, OLFML2B[30], hence affecting external phenotype rather than genome function. Hence, both genome and organismal function

can constrain hybrid genome formation, and the relative importance of the two may vary both quantitatively and qualitatively with the parent species.

Our data suggests that hybridization is a more potent force for creating novel variation than previously recognised, as many different combinations of the parental genomes can arise in hybrids and allow for adaptive divergence between isolated populations of the hybrid species. Importantly, we show that the variation is limited for DNA-repair and mitochondrial genes. These may contribute to the general pattern of reduced introgression on sex chromosomes, and are candidate loci for reproductive loci that may be important in speciation.

Methods

Field sampling

Italian sparrows were caught on Crete, Corsica and Sicily in 2013 and on Malta in 2014, while Spanish sparrows were caught in Lesina, Italy, 2008. House sparrows were caught in Northern Norway between 2007 and 2013, and the outgroup tree sparrow was caught in Sicily during 2008 (Supplementary Table 1). All sparrows were caught using mist nets, and blood was sampled from the brachial vein and immediately stored in Queens lysis buffer. Sparrows were released immediately after blood sampling to minimize stress. All permits were obtained from appropriate authorities prior to sampling.

Whole genome resequencing, data processing and analysis

DNA extraction and sequencing

DNA was extracted from blood samples stored in Queens lysis buffer using Qiagen DNeasy Blood and Tissue Kits (*Qiagen Corp*., Valencia, CA), and stored in Qiagen Elution Buffer (*Qiagen Corp*., Valencia, CA). Whole genome re-sequencing was performed with Illumina sequencing technology. An Illumina TruSeq gDNA 180 bp library was created and sequenced on the Illumina HiSeq 2000 platform with 100 bp read length and three individuals per lane for the parent species, the tree sparrow and the Italian sparrows from Malta; sparrows from Crete, Corsica and Sicily were sequenced with four individuals per lane. All re-sequencing was performed by Genome Quebec at McGill University (Montreal, Canada)

(http://www.genomequebec.com/en/home.html). Raw data have been deposited at the NCBI Sequence Read Archive under (Accessions YYYY).

Variant calling and filtering

All raw sequence reads were mapped to a repeat-masked version of the house sparrow genome[13] using BWA 0.7.8[31] with the mem, -M and -R options. A sorted BAM file was produced using SAMTOOLS version 1.0[32] using view with the -b, -U and -s options and a pipe to the sort command. Duplicates were identified and filtered out using MARKDUPLICATES from PICARD-TOOLS version 1.107 (http://broadinstitute.github.io/picard/) using the options validation stringency = lenient, assume sorted = true and index = true. Indels were identified using RealignerTargetCreator and local realignments around these were performed with IndelRealigner. Standard settings were used for both these tools which are components of GATK 3.3.0[33,34]. Final sequencing coverage of these final BAM files (excluding duplicates) was 8x per individual (min. 5.99x, max. 15.8x; Supplementary Table 2). Variants were then called using the GATK HaplotypeCaller. First, HaplotypeCaller was run separately for each individual to create single-sample gVCFs using the --emitRefConfidence GVCF, --variant_index_type LINEAR and --variant_index_parameter 128000 options, and then GATK GenotypeGVCFs tool was run using standard settings to achieve joint genotyping. Two different versions of the VCF file were created one with variable sites only (49 237 560 SNPs), and one where all sites were called as specified by the --includeNonVariantSites option (1 040 518 317 SNPs).

For both VCF files, indels were first filtered out using VCFtools version 0.1.12b[35],

and hard filtering according to the Broad Institute's recommendations was performed with bcftools-1.2[32] in order to filter out sequencing artefacts. This included requiring a QualByDepth of at least 2.0 and a FisherStrand phred-scaled p-value of less than 60.0, based on Fisher's Exact Test to detect strand bias in the reads, which may be indicative of false positive calls. Hard filtering also required a RMSMappingQuality of at least 40.0 to ensure high mapping quality of the reads across all samples, a MappingQualityRankSumTest value of -12.5 to exclude reads where the alternate alleles has a lower mapping quality than reads with the reference allele, and finally a ReadPosRankSumTest value of less than -8.0 was required to ensure that reads with the alternate allele were not shorter those with the reference allele, potentially indicating sequencing artefacts. In addition, we filtered out sites with mean number of reads per individual lower than 3, with a genotype quality lower than 20 and with a mapping quality lower than 20 using VCFtools version 0.1.12b[35]. To avoid paralogs, we also excluded sites with read depths above 5 times the variance of coverage. This left a total of 38 341 426 sites for further analysis.

PCA

Principal component analysis was performed using ANGSD[36] version 0.911 and ngsTools version 1.0.1[37]. This pipeline was chosen because it does not rely on genotype calls but instead takes allele frequency likelihoods and genotype probabilities into account[38]. We first estimated genotype probabilities from BAM files with ANGSD, including bi-allelic sites only and allowing a minimum mapping and site quality of 20 (Phred score) and a minimum coverage of 30x across all individuals. PCA was then performed using ngsTools on genotype probabilities. Allele frequencies were normalized and genotypes were not explicitly called, as specified by setting the options -norm 0 and -call 0 options. Eigenvalues for each PC were then

292 estimated from the covariance matrix produced by ngsTools and data was plotted
293 using a custom R script. We used the broken stick criteria to assess which PC axes
294 were biologically informative from a simple scree plot and then extracted the
295 covariance from these axes (Supplementary Table 3). The analysis was performed on
296 three datasets: all sites, sites from the Z chromosome only, and autosomal sites only.

297 **Population genomic analysis**

298 Population genetic parameters were estimated for non-overlapping 100 kb windows
299 along the genome, as this window size was larger than the distance of LD-decay[13].
300 Population genetic inference was based on genotype likelihoods whenever possible.
301 ANGSD[36] version 0.911 was used to estimate allele frequency likelihoods and to
302 obtain a maximum likelihood estimate of the unfolded site frequency spectrum (SFS)
303 for estimation of Tajima's D. Nucleotide diversity was estimated by dividing the
304 pairwise Theta (population scaled mutation rate) estimates by the number of variable
305 sites per window. The ancestral sequence was reconstructed using genotypes from the
306 outgroup tree sparrow. A Fasta file of the tree sparrow genome was obtained by using
307 the -doFasta 2 command with the -GL 1 -doCounts 1, -setMinDepth 3 and -
308 setMaxDepth 65 options in ANGSD[36] version 0.911. Here, BAM files from eight
309 additional tree sparrows, sequenced to 8-12x depth and processed as described for the
310 other samples above, were used. Genetic differentiation (F_{ST}) based on genotype
311 likelihood was estimated based on the two-dimensional SFS using ngsTools[38]. All
312 these analyses were performed on the final BAM files. Sequence divergence (d_{xy}) was
313 calculated from the VCF file with all positions called using the script developed in[39]
314 (https://github.com/johnomics/Martin_Davey_Jiggins_evaluating_introgression_statis
315 tics/blob/master/egglib_sliding_windows.py; version August 2014).

316 **Admixture analysis**

317 Genetic admixture was estimated using ADMIXTURE[40] version 0.911. The VCF-file
318 was converted to plink's PED format using VCFtools version 0.1.12b[35] and plink
319 version 1.07[41]. Log likelihood values for K, the number of genetic clusters in the
320 datasets, between K=1 and K=8 were estimated (Supplementary Table 4), and
321 admixture analyses were run for the most appropriate value of K. Analyses were first
322 run for a LD-pruned whole genome dataset (sites within a 50 SNP stepping window
323 with a correlation coefficient higher than 0.1 were omitted; pruning of the BED file
324 was performed with plink version 1.07 using the --indep-pairwise command; this left
325 438,443 sites for analysis). Sliding window analyses with 100 kb windows were then
326 carried out to investigate variability in probability of parental inheritance across the
327 genome. This analysis was performed individually for each island population together
328 with the two parent species. The VCF-file was then split into individual VCF files per
329 100 kb using the –L option in GATK 3.3.0[33,34]. Individual ADMIXTURE analyses for
330 K=2 were then performed for each 100 kb BED file. A mean estimated cluster
331 assignment probability for all individuals per population was computed for each
332 analysis using a custom python script.

333 **Phylogenetic analyses**

334 To investigate whether phylogenetic relationships varied across the genome due to
335 introgression or incomplete lineage sorting, a machine-learning approach
336 implemented in SAGUARO version 0.1[42] was used to identify genomic regions
337 characterized by distinct similarity matrices. To focus on breakpoints at which
338 recombination led to changes in the topology of populations or species, rather than
339 topological changes within populations, one representative high-coverage individual
340 per Italian sparrow population or sparrow species was used in these analyses. These

were the house sparrow 8L19786, the Spanish sparrow Lesina_280 and the Italian sparrows C081 (Crete), K035 (Corsica), S059 (Sicily) and M036 (Malta). The program VCF2HMMFeature (included in the SAGUARO package) was used to convert the VCF-file to the HMMFeature format required by SAGUARO, and SAGUARO was run using default parameters. Analysis was performed jointly for all chromosomes, as the same similarity matrices are expected to occur on multiple chromosomes. A total of 797 contiguous regions of up to 54,398 bp were identified and assigned to one out of 41 similarity matrices, of which the most common similarity matrix characterized 79.08% of the genome. Similarity matrices represented by more than 2% of the genome are depicted in Figure 1.

The genomic regions identified by SAGUARO were subsequently used to infer differences in phylogenetic relationships more thoroughly with the Bayesian software BEAST version 2.2.0[43]. To this end, chromosome-length alignments were first phased using SHAPE-IT version 2[44]. To improve phasing, this analysis was conducted with a subset consisting of the six individuals with the highest coverage per population (24 individuals in total) rather than just the six individuals used for SAGUARO analyses, however, the six focal individuals were extracted from the alignments following phasing. The phased chromosome-length alignments were then used to extract 38,964 non-overlapping blocks of 25,000 bp from the 797 contiguous regions identified with SAGUARO. For each of the 38,964 blocks, one of the two phased sequences per individual was excluded at random, so that each alignment contained a single sequence per population or species. To identify alignments particularly suitable for Bayesian phylogenetic analysis, we quantified, for each alignment, the proportion of missing data, the number of parsimony-informative sites, the proportion of heterozygous sites, the mean bootstrap support of maximum-likelihood trees

366 generated with RAXML version 8.2.4[45], and the probability that the alignment is free
367 of recombination determined with the Phi test[46]. We assumed that alignments with a
368 low proportion of heterozygous sites are less likely to contain paralogous sequences,
369 and that alignments with many parsimony-informative sites and high mean bootstrap
370 support contain strong phylogenetic signal. Thus, alignments were selected according
371 to the following "relaxed" and "strict" filters: a proportion of missing data below 0.2
372 (relaxed) or 0.1 (strict), at least 75 (relaxed) or 100 (strict) parsimony-informative
373 sites, a proportion of heterozygous sites below 0.005 (relaxed) or 0.0025 (strict), a
374 mean bootstrap support of at least 90 (both relaxed and strict), and a Phi test *p*-value
375 above 0.005 (relaxed) or 0.01 (strict). A total of 1,234 and 116 alignments were
376 selected with these relaxed and strict filters, respectively. To include an outgroup for
377 phylogenetic analyses with BEAST, consensus sequences of tree sparrow reads from
378 the Naxos1 individual mapped to the house sparrow reference genome[13] were added
379 to each selected alignment. To avoid bias towards the reference, missing data were
380 not replaced by the reference alleles. The phylogeny of each alignment was then
381 inferred with BEAST, using the GTR model of sequence evolution with estimated
382 base frequencies, a Yule tree prior[47], and 50 million Markov chain Monte Carlo
383 iterations. The ingroup, combining house sparrow, Spanish sparrow, and the
384 representatives of Italian sparrow populations was constrained to be monophyletic. In
385 the absence of a reliable absolute time line for sparrow divergences, the time of
386 divergence of ingroup and outgroup was fixed at 1 time unit, so that all divergence
387 ages within the ingroup are estimated relative to this initial split. Convergence of all
388 MCMC chains was confirmed by effective samples sizes greater than 500 for all
389 model parameters.

390 **Introgression**

Presence of introgression was estimated using Patterson's D[48,49] calculations, using the scripts provided in[39]. ABBA-BABA estimates were calculated using a minimum coverage of 3, a 100 kb window size and 1000 informative sites using the egglib_sliding_windows.py script. The test was set up to estimate Spanish sparrow introgression into a house sparrow background, with tree sparrow as an outgroup.

Mitochondrial DNA

Mitochondrial DNA gvcfs were called separately with haploid settings using the -ploidy argument in HaplotypeCaller, jointly genotyped, and filtered as described above using GATK 3.3.0[33,34]. Fitchi version 1.1.4 [50] was used to reconstruct a haplotype genealogy based on Fitch distances.

Recombination rate and common differentiation

Genome-wide recombination rates were estimated using a house sparrow linkage map. As the recombination map was produced using SNP chip data, recombination distance estimates were first averaged using a sliding window approach and then a loess fit of mean recombination rate against physical distance was performed in order to interpolate fine scale variation across non-overlapping 100 kb windows. Since recombination data were not available for the Z chromosome, this interpolation was performed on autosomes only.

To test whether there was a relationship between recombination rate variation and relative genomic differentiation, either F_{ST} from 100 kb non-overlapping windows, which is a direct measure of relative differentiation, or the common differentiation axis – i.e. shared differentiation amongst groups of populations – were used. The latter was calculated by performing PCA on multiple pairwise comparisons of differentiation featuring the same focal species following[51]. Common differentiation

was estimated amongst 1) all Italian populations, 2) between all Italian populations and the house sparrow and 3) between all Italian populations and the Spanish sparrow. In each case, all pairwise F_{ST} comparisons including these species were included, and the first principal component extracted.

Outlier gene analysis

Disparities in F_{ST} values between lineages were used to identify genomic regions in which the Italian sparrow populations display elevated divergence from either or both of their parents or other Italian sparrow populations. This approach is reminiscent to population branch statistics. Three categories of outliers were of interest: Between island outliers, where island populations differed strongly in parental resemblance, were selected as these are informative of how Italian sparrow populations are differentially adapted. Private outliers, windows in which Italian sparrows are diverged from both parent species, show where unique adaptation is putatively strongly selected for. Finally, portions of the genome invariably inherited from the same parent species for all populations are informative of parts that are under strong selection to resemble a specific parent species and may reveal constraints on hybrid speciation.

<u>Between island outliers:</u>

The 1% of 100 kb regions which differed most with respect to F_{ST} against the parental species between two islands were selected for all possible island-island combinations. All genes within or partly within these regions were then extracted. As historical effects such as ancestral polymorphism and selection prior to the parental split can be assumed to be constant across Italian populations, using the difference in F_{ST} against the same parent species will yield results which are not dependent on these factors.

Furthermore, F_{ST} was not strongly dependent on recombination rate between Italian sparrow populations (Supplementary Fig. 4).

Private Outliers:

Outliers in which Italian sparrows were differentiated from both parent species, hereafter private outliers, were extracted from the 1% of windows exhibiting the largest difference in F_{ST} between each hybrid/parent comparison, only keeping windows overlapping between both hybrid/parent comparisons. This was done separately for each Italian population, and all outliers detected across the populations were then merged for gene ontology analyses.

Outliers invariably resembling one parent species:

To limit historical effects, due to for instance ancestral polymorphism and selection prior to the parental split, we used the 100 kb windows in which the Italian sparrow had the cumulative largest difference in F_{ST} value between one parent and the other. This was achieved by summing the F_{ST} values between all Italian sparrow populations and house sparrow, and subtract the sum of the F_{ST} values between these populations and Spanish sparrow. As the resulting distribution was skewed, using a percentage at each tail would not have captured the biological pattern where house sparrow inheritance across populations was more common than Spanish sparrow inheritance across all populations. Therefore, we squared the summed values and extracted the 2% of the windows that deviated most strongly from 0 (Supplementary Fig. 8B), which yielded more invariably house sparrow like outliers than invariably Spanish sparrow like outliers.

For all outlier windows in each of the three categories above, annotated genes that resided completely or partially within them were extracted for separate gene ontology

463 analysis. One analysis was performed on all private outliers identified across
464 populations, one on all outliers between populations, including all combinations of
465 populations, one on outliers that resembled house sparrow across all populations, and
466 finally one on outliers that resembled Spanish sparrow across all populations. As only
467 14 outliers resembling Spanish sparrow across all populations were found, no
468 significant GO-terms were found for this analysis. Therefore, we do not provide a
469 table of significant terms for this analysis. These analyses were performed using GO-
470 stat[52], with a human reference base. We implemented standard settings for GO
471 analyses, i.e. a values of 3 as the minimal length of considered GO paths and no
472 merging of GOs if gene lists overlap. Mito-nuclear genes were identified using
473 MITOMINER 4.0[53] with a human reference database and standard settings.
474 Overrepresentation of mitonuclear genes was subsequently tested using a Chi-square
475 test. Corrections for multiple testing were performed with the Benajmini method.
476
477 **DN/DS analyses**
478 To test if the Z chromosome was under stronger selection, synonymous and non-
479 synonymous fixed differences within genes against each parent species for all
480 autosomes and the Z chromosome were estimated for each Italian sparrow population.
481 A goodness of fit test was applied to test if the number of nonsynonymous
482 substitutions on the Z chromosome was higher than expected for each parent species.
483 To this end, the R package PopGenome[54] was used. The splitting,data command was
484 used to extract genes and fixed sites were extracted. Synonymous and
485 nonsynonymous sites were then identified using the options subsites="syn" and
486 subsites="nonsyn", respectively.
487

Linkage disequilibrium decay

To address whether linkage disequilibrium (LD) was higher and LD decayed more slowly in outlier windows than in randomly selected windows, plink version 1.90b3b[41] was used. Using --const-fid --ld-window 1000 --ld-window-kb 100 --r2 and –ld-window-r2 0.0, linkage disequilibrium in the 100 kb outlier windows was estimated within each of the Italian populations. In addition, we randomly selected 1,000 100 kb windows spread across the chromosomes in proportion with chromosome size and estimated linkage disequilibrium for these in the same manner for comparison. A linear model was fitted for each outlier window, and intercept and slope were recorded and used in glm's to test whether LD was higher and LD decay was slower in outlier windows than in randomly selected windows.

References

1. Mallet, J. Hybridization as an invasion of the genome. Trends Ecol. Evol. **20**, 229–237 (2005).

2. Abbott, R. et al. Hybridization and speciation. J. Evol. Biol. **26**, 229–246 (2013).

3. Seehausen, O. Hybridization and adaptive radiation. Trends Ecol. Evol. **19**, 198–207 (2004).

4. The Heliconius Genome Sequencing Consortium. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature **487**, 94–98 (2012).

5. Rieseberg, L. H. Major Ecological Transitions in Wild Sunflowers Facilitated by Hybridization. Science **301**, 1211–1216 (2003).

6. Hermansen, J. S. et al. Hybrid speciation in sparrows I: phenotypic intermediacy, genetic admixture and barriers to gene flow. Mol. Ecol. **20**, 3812–3822 (2011).

7. Trier, C. N., Hermansen, J. S., Sætre, G.-P. & Bailey, R. I. Evidence for Mito-Nuclear and Sex-Linked Reproductive Barriers between the Hybrid Italian Sparrow and Its Parent Species. PLoS Genet **10**, e1004075–10 (2014).

8. Sankararaman, S. et al. The genomic landscape of Neanderthal ancestry in present-day humans. Nature **507**, 354–357 (2014).

9. Fontaine, M. C. et al. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science **347**, 1258524–1258524 (2015).

10. Baack, E. J. & Rieseberg, L. H. A genomic view of introgression and hybrid speciation. Curr. Opin. Genet. Dev. **17**, 513–518 (2007).

523 11. Martin, S. H. et al. Genome-wide evidence for speciation with gene flow in
524 Heliconius butterflies. Genome Res. **23**, 1817–1828 (2013).
525 12. Qvarnström, A. & Bailey, R. I. Speciation through evolution of sex-linked
526 genes. Heredity **102**, 4–15 (2008).
527 13. Elgvin, T. O. et al. The genomic mosaicism of hybrid speciation. In review,
528 Science Advances. *MS attached in submission*.
529 14. Saetre, G. P. et al. Single origin of human commensalism in the house
530 sparrow. J. Evol. Biol. **25**, 788–796 (2012).
531 15. Bache-Mathiesen, L. The Evolutionary Potential of Male Plumage Color in a
532 Hybrid Sparrow Species. UiO DUO vitenarkiv
533 https://www.duo.uio.no/handle/10852/45473
534 16. Burri, R. et al. Linked selection and recombination rate variation drive the
535 evolution of the genomic landscape of differentiation across the speciation continuum
536 of Ficedula flycatchers. Genome Res. **25**, 1656–1665 (2015).
537 17. Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial
538 selection. Genet. Res. **8**, 269–294 (1966).
539 18. Laine, V. N. et al. Evolutionary signals of selection on cognition from the
540 great tit genome and methylome. Nat. Commun. **7**, 1–9 (2016).
541 19. Lamichhaney, S. et al. Evolution of Darwin's finches and their beaks revealed
542 by genome sequencing. Nature **518**, 371–375 (2015).
543 20. Eroukhmanoff, F., Hermansen, J. S., Bailey, R. I., Sæther, S. A. & Sætre, G.-
544 P. S. Local adaptation within a hybrid species. Heredity **111**, 286–292 (2013).
545 21. Noramly, S., Freeman, A. & Morgan, B. A. Beta-catenin signaling can initiate
546 feather bud development. Development **126**, 3509–3521 (1999).

547 22. Guo, H. et al. Wnt/beta-catenin signaling pathway activates melanocyte stem
548 cells in vitro and in vivo. J. Dermatol. Sci. **83**, 45–51 (2016).
549 23. Mank, J. E., Nam, K. & Ellegren, H. Faster-Z evolution is predominantly due
550 to genetic drift. Mol. Biol. Evol. **27**, 661–670 (2010).
551 24. Charlesworth, B., Coyne, J. A. & Barton, N. H. The Relative Rates of
552 Evolution of Sex Chromosomes and Autosomes. Am. Nat. **130**, 113–146 (2016).
553 25. David, W. M., Mitchell, D. L. & Walter, R. B. DNA repair in hybrid fish of
554 the genus Xiphophorus. Comp. Biochem. Physiol. C Toxicol. Pharmacol. **138**, 301–
555 309 (2004).
556 26. Greig, D., Travisano, M., Louis, E. J. & Borts, R. H. A role for the mismatch
557 repair system during incipient speciation in Saccharomyces. J. Evol. Biol. **16**, 429–
558 437 (2003).
559 27. Schumer, M. & Brandvain, Y. Determining epistatic selection in admixed
560 populations. Mol. Ecol. **25**, 2577–2591 (2016).
561 28. Barton, N. H. The role of hybridization in evolution. Mol. Ecol. **10**, 551–568
562 (2001).
563 29. Poelstra, J. W., Vijay, N., Hoeppner, M. P. & Wolf, J. B. W. Transcriptomics
564 of colour patterning and coloration shifts in crows. Mol. Ecol. **24**, 4617–4628 (2015).
565 30. Tomarev, S. I. & Nakaya, N. Olfactomedin Domain-Containing Proteins:
566 Possible Mechanisms of Action and Functions in Normal Development and
567 Pathology. Mol. Neurobiol. **40**, 122–138 (2009).

Additional References for Methods Section

31. Li, H. & Durbin, R. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. Bioinformatics **25**, 1754–1760 (2009).

32. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**, 2078–2079 (2009).

33. McKenna, A. et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. **20**, 1297–1303 (2010).

34. Van der Auwera, G. A. et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr. Protoc. Bioinformatics. **11**, 1–33 (2013).

35. Danecek, P. et al. The variant call format and VCFtools. Bioinformatics **27**, 2156–2158 (2011).

36. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics **15**, 443–13 (2014).

37. Fumagalli, M., Vieira, F. G., Linderoth, T. & Nielsen, R. ngsTools: methods for population genetics analyses from next-generation sequencing data. Bioinformatics **30**, 1486–1487 (2014).

38. Fumagalli, M. et al. Quantifying Population Genetic Differentiation from Next-Generation Sequencing Data. Genetics **195**, 979–992 (2013).

39. Martin, S. H., Davey, J. W. & Jiggins, C. D. Evaluating the Use of ABBA-BABA Statistics to Locate Introgressed Loci. Mol. Biol. Evol. **32**, 244–257 (2014).

40. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. **19**, 1655–1664 (2009).

592 41. Purcell, S. et al. PLINK: A Tool Set for Whole-Genome Association and
593 Population-Based Linkage Analyses. Am. J. Hum. Genet. **81**, 559–575 (2007).
594 42. Zamani, N. et al. Unsupervised genome-wide recognition of local relationship
595 patterns. BMC Genomics **14**, 1-11 (2013).
596 43. Bouckaert, R. et al. BEAST 2: A Software Platform for Bayesian Evolutionary
597 Analysis. PLoS Comput. Biol. **10**, e1003537–6 (2014).
598 44. O'Connell, J. et al. A General Approach for Haplotype Phasing across the Full
599 Spectrum of Relatedness. PLoS Genet. **10**, e1004234–21 (2014).
600 45. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-
601 analysis of large phylogenies. Bioinformatics **30**, 1312–1313 (2014).
602 46. Bruen, T. C., Philippe, H. & Bryant, D. A simple and robust statistical test for
603 detecting the presence of recombination. Genetics **172**, 2665–2681 (2006).
604 47. Yule, U. G. A Mathematical theory of evolution, based on the conclusions of
605 Dr. J. C. Willis, F.R.S. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **213**, 21–87 (1925).
606 48. Green, R. E. et al. A draft sequence of the Neandertal genome. Science **328,**
607 710–722 (2010).
608 49. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for Ancient
609 Admixture between Closely Related Populations. Mol. Biol. Evol. **28**, 2239–2252
610 (2011).
611 50. Matschiner, M. Fitchi: haplotype genealogy graphs based on the Fitch
612 algorithm. Bioinformatics **32**, 1250–1252 (2016).
613 51. Burri, R. et al. Linked selection and recombination rate variation drive the
614 evolution of the genomic landscape of differentiation across the speciation continuum
615 of Ficedulaflycatchers. Genome Res. **25**, 1656–1665 (2015).

616 52. Beissbarth, T. & Speed, T. P. GOstat: find statistically overrepresented Gene
617 Ontologies within a group of genes. Bioinformatics **20**, 1464–1465 (2004).
618 53. Smith, A. C., Blackshaw, J. A. & Robinson, A. J. MitoMiner: a data
619 warehouse for mitochondrial proteomics data. Nucleic Acids Res. **40**, D1160–D1167
620 (2011).
621 54. Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E. & Lercher, M. J.
622 PopGenome: an efficient Swiss army knife for population genomic analyses in R.
623 Mol. Biol. Evol. **31**, 1929–1936 (2014).
624

625 **Supplementary Information** is linked to the online version of the paper at

626 www.nature.com/nature.

627

628 **Acknowledgements**

629 We thank Maria Tesaker and BirdLife Malta for help with field work, Laura Piñeiro

630 and Lena Bache-Mathiesen for providing morphological data, and Anna Nilsson for

631 comments on the manuscript. This work was funded by a Swedish Research Council

632 post doctoral grant and a Wenner-Gren Fellowship to A.R. and a Norwegian Science

633 Foundation grant to G-P.S. and A.R.

634

635 **Author Contributions**

636 A.R. conceived the study, carried out field work, lab work, designed analyses,

637 analysed data and wrote the manuscript. C.N.T. helped design analyses, and provided

638 example scripts for many analyses, F.E. carried out field work and the gene ontology

639 analyses, J.S.H. carried out field work and the final touches in figure preparation,

640 M.M. did the BEAST and Saguaro analyses and M.R. performed the recombination

641 rate analyses and PCA. T.E. provided the house sparrow reference genome, and

642 G.P.S. identified the study system, designed the sampling strategy and carried out

643 field work. All co-authors commented on the manuscript.

644

645 Data is deposited at YYYY. Reprints and permissions information is available at

646 www.nature.com/reprints. We declare no competing financial interests.

647 Correspondence and requests for materials should be addressed to

648 anna.runemark@ibv.uio.no

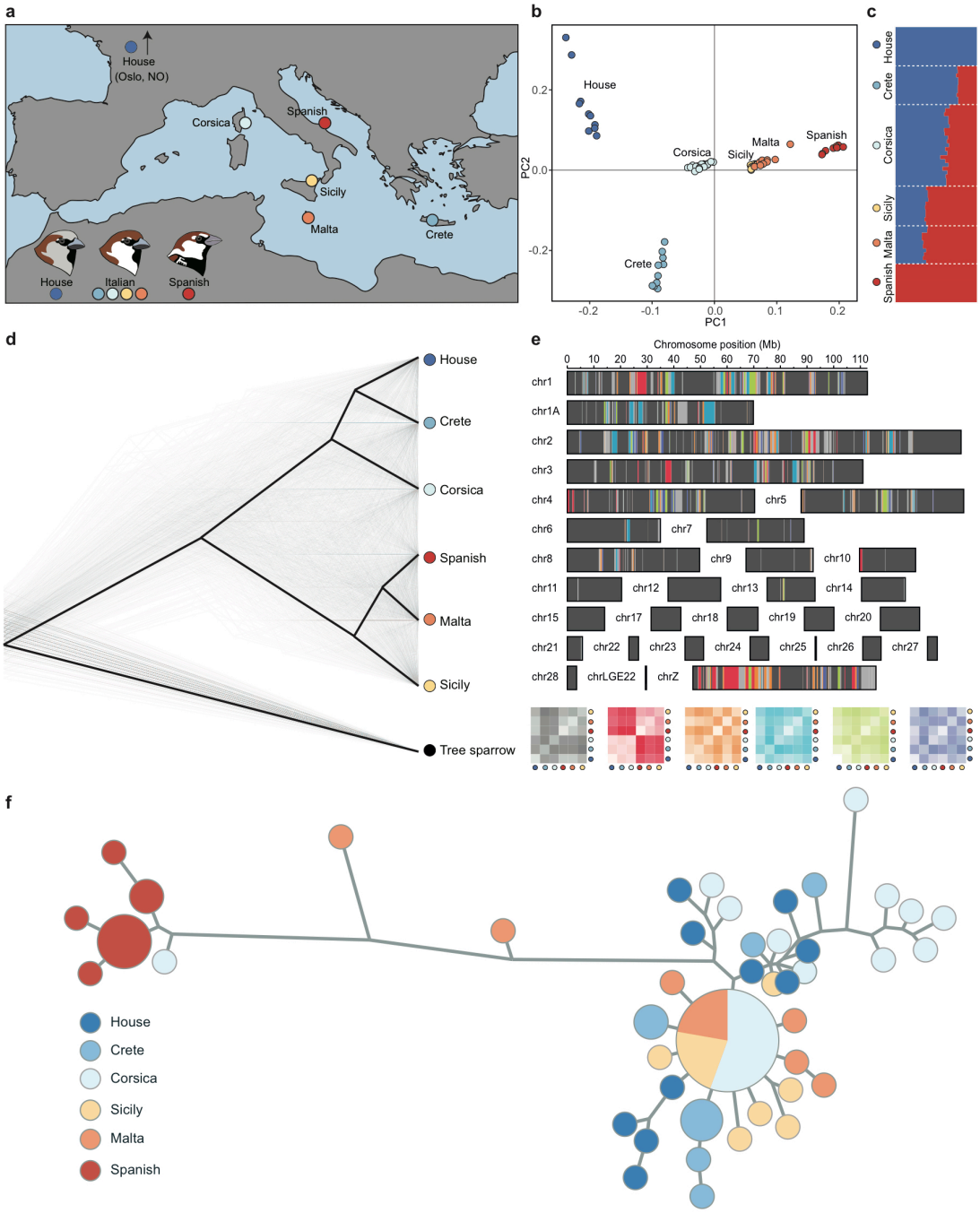
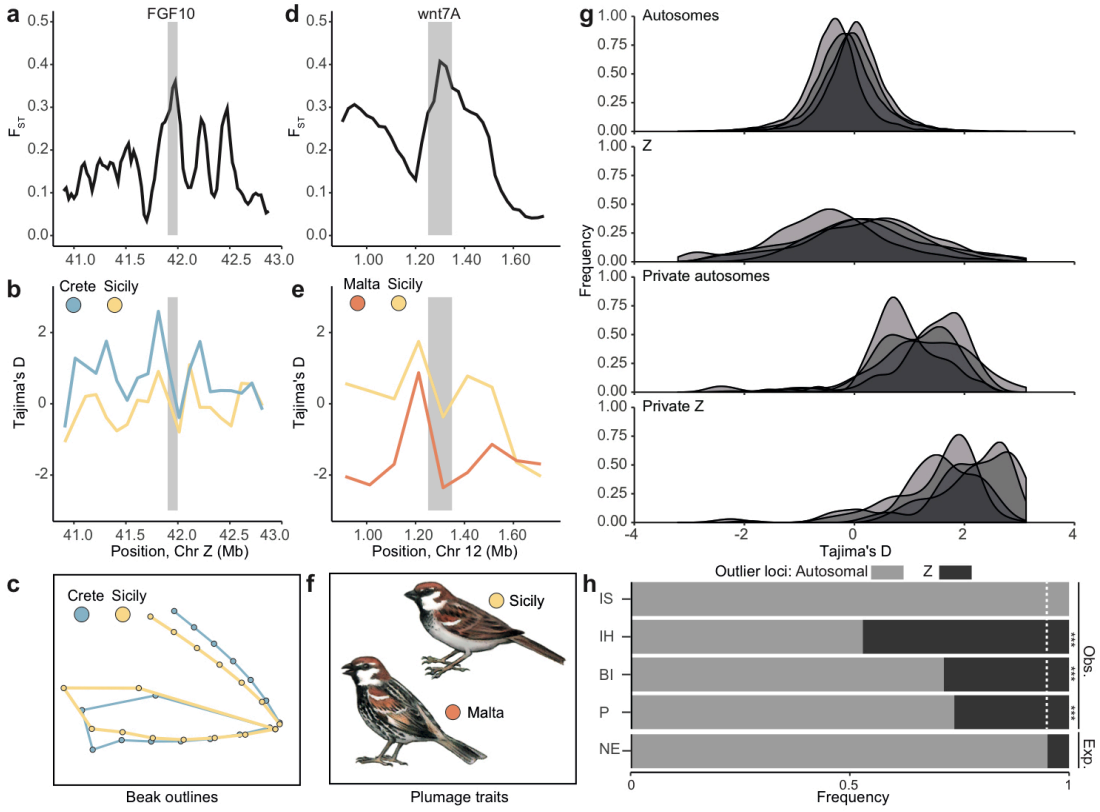
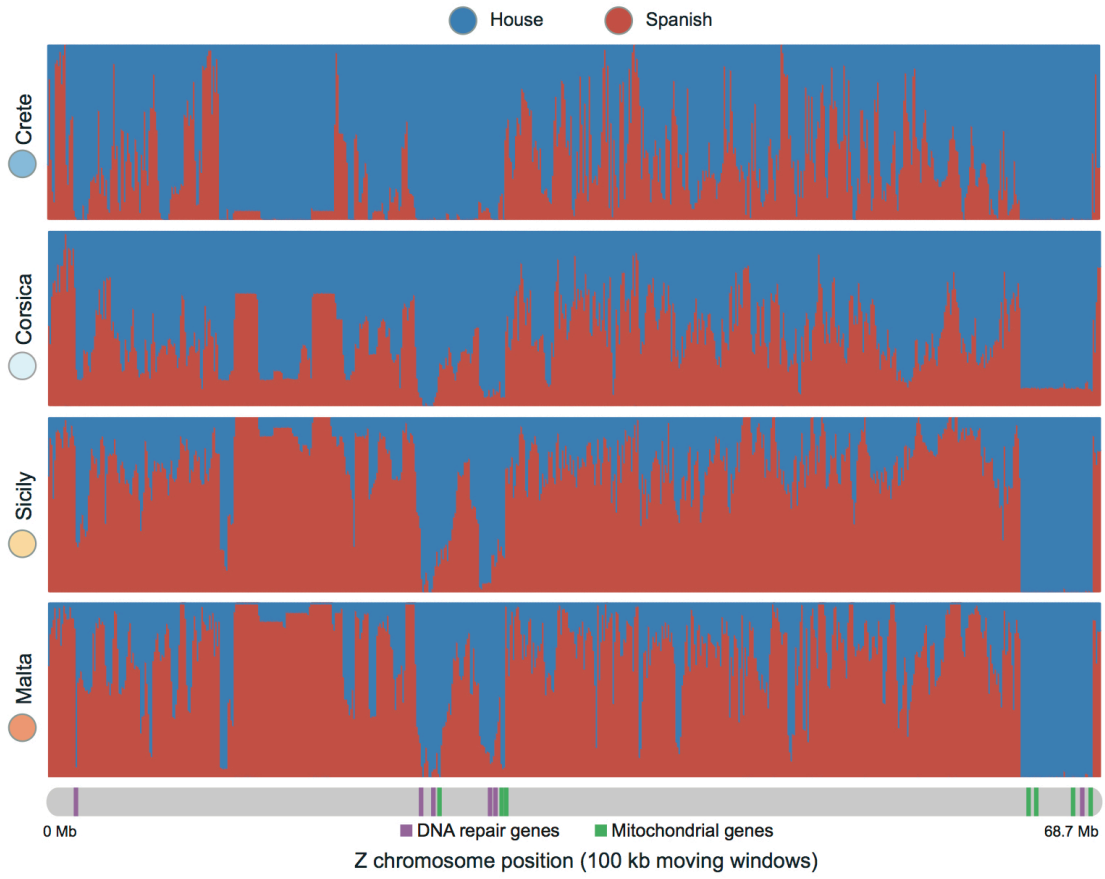


Figure 1. Population structuring of the focal taxa. a, Map showing the location of the island populations and the reference parent species populations, with examples of male plumage patterns. **b**, PCA of LD-pruned high quality SNP set with eigenvector 1 on the x-axis and eigenvector 2 on the y-axis. **c**, Population structuring based on Admixture analysis for house sparrow, Italian and Spanish sparrow populations. **d**, BEAST trees illustrating genome-wide variation in phylogenetic clustering between the taxa. **e**, SAGUARO plot illustrating the distribution of the six most common relatedness matrixes over the genome. **f**, Haplotype genealogy graph of mitochondrial sequences.



659

660 **Figure 2. Local adaptation, private variation, and strong selection on the Z-**
661 **chromosome. a**, The beak shape candidate gene FGF10 is differentiated between
662 Italian sparrows on Crete and Sicily. **b**, Tajima's D for the FGF10 region. **c**, Beak
663 shape differences between Cretan and Sicilian sparrows. **d**, The plumage candidate
664 gene wnt7a is differentiated between genetically similar sparrow populations on
665 Sicily and Malta. **e**, Tajima's D for the wnt7a region. **f**, Schematic illustration of
666 plumage differences between the populations. **g**, Strength of selection on the
667 autosomes, the Z-chromosome and on private outlier loci where the Italian sparrow is
668 differentiated from both parent species. **h**, Distribution of outlier genes between the
669 Z-chromosome and the autosomes. IS denotes invariably Spanish, IH invariably
670 house, BI between islands, P private and NE neutral expectations.



671

672 **Figure 3. Parental similarity across the Z-chromosome**. Sliding window

673 ADMIXTURE analysis of probability of house sparrow (blue) or Spanish sparrow

674 (red) inheritance over the Z-chromosome for the four Italian sparrow populations.

675 Areas highly constrained to house sparrow inheritance harbour a significantly higher

676 proportion DNA repair genes (green lines) and many mito-nuclear (purple lines)

677 genes.