
nature
ecology &
evolution
Endless forms

nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-017-0437-7

Variation and constraints in hybrid genome formation

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

Hybridization is an important source of variation; it transfers adaptive genetic variation across species boundaries and generates new species. Yet, the limits to viable hybrid genome formation are poorly understood. Here we investigated to what extent hybrid genomes are free to evolve by sequencing the genomes of four island populations of the homoploid hybrid Italian sparrow *Passer italiae*. We report that a variety of novel and fully functional hybrid genomic combinations are likely to have arisen independently on Crete, Corsica, Sicily and Malta, with differentiation in candidate genes for beak shape and plumage colour. However, certain genomic regions are invariably inherited from the same parent species, limiting variation. These regions are over-represented on the Z chromosome and harbour candidate incompatibility loci, including DNA-repair and mitonuclear genes. These gene classes may contribute to the general reduction of introgression on sex chromosomes. This study demonstrates that hybrid genomes may vary, and identifies new candidate reproductive isolation genes.

Introgressive hybridization is increasingly recognized as a potent evolutionary process[1,2]. Hybridization can spur adaptive radiations[3], transfer adaptive variation across species boundaries[4] and generate species with novel niches[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[6]. Portions of the genome vary in extent of introgression[4,7], and homoploid hybrid genomes are predicted to have unequal parental contributions[8]. However, how hybridization and subsequent recombination and selection mould 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[9], and hence are important for our understanding of the origin of species. One well-supported finding is reduced introgression on sex chromosomes, which is common in species where one sex is heterogametic[6,10]. To what extent this pattern is caused by specific types of gene on sex chromosomes is debated[11], and knowledge of the genes causing incompatibilities and reproductive isolation is needed to resolve this.

To investigate whether certain genes are strongly selected to 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*[9,12] (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 sparrow *P. domesticus* and the Spanish sparrow *P. hispaniolensis*[13]. It is reproductively isolated from both parent species[9], and the barriers isolating it from each parent species are a subset of those isolating the parent species[14]. The species is thought to have originated when house sparrows colonized the Mediterranean less than 10,000 years ago[12,15]. Morphologically divergent populations of Italian sparrows are found on the Mediterranean islands of Crete, Corsica, Sicily and Malta[16] (Fig. 1a), 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 and 2). We sequenced 10–21 Italian sparrow genomes from each island and 10 genomes from each of the parent species to 6–16× coverage as well as a tree sparrow (*P. montanus*) outgroup, and aligned them to the recently de novo assembled house sparrow reference genome[13].

Results

To determine whether the island populations of Italian sparrows 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 and Supplementary Tables 3 and 4). The Italian sparrow populations differ in position along the axis of differentiation of the parent species, with Crete and Corsica closer to the house sparrow and Sicily and Malta closer to the 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 admixture proportions for the Italian populations, with Crete and Corsica having the highest probabilities of clustering with the house sparrow and Sicily and Malta having the highest probabilities of clustering with the Spanish sparrow (Fig. 1c). The admixture proportions differed significantly among populations (analysis of variance (ANOVA) $F_{3,47}=736.54$, $P=2.2\times10^{-1}$; Supplementary Table 7), and between 100-kilobase (kb) sliding windows across the genome (Supplementary Fig. 2). Moreover, in the most frequent phylogenetic tree topology, the house sparrow, Crete and Corsica form one cluster and the Spanish sparrow, Malta and Sicily form the other, with Crete and Malta clustering most closely with parent species (Fig. 1d and Supplementary Fig. 3).

Department of Biosciences, Centre for Ecological and Evolutionary Synthesis, University of Oslo, Oslo, Norway. *e-mail: anna.runemark@ibv.uio.no

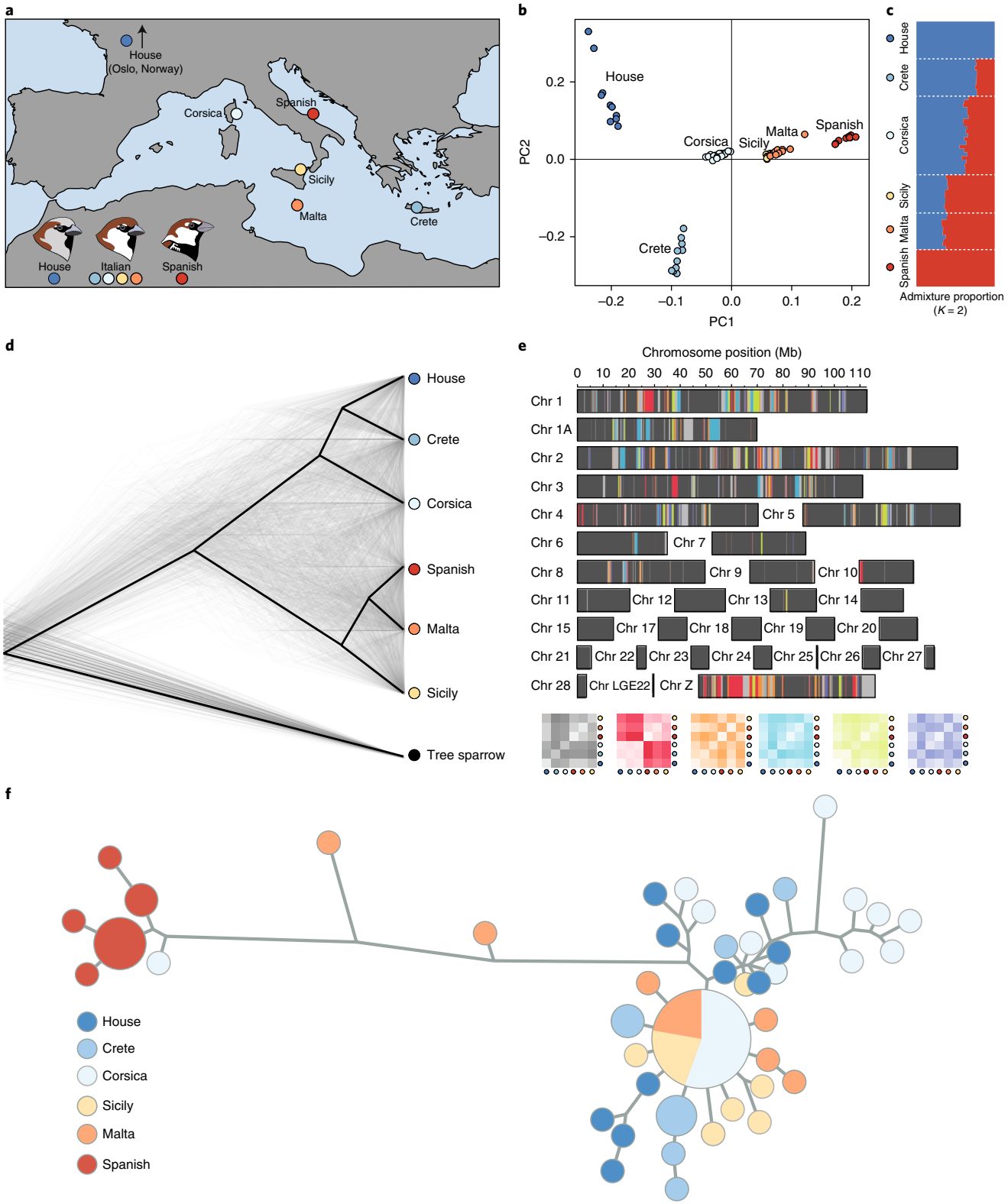

Fig. 1 | Population structuring of the focal taxa. a, Map showing the location of the Italian sparrow populations from Crete ($n = 10$), Corsica ($n = 21$), Sicily ($n = 10$) and Malta ($n = 10$) and the reference parent species populations ($n = 10$ for each parent species), with examples of male plumage patterns. **b**, Principal component analysis of linkage-disequilibrium-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, 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 matrices over the genome. The intensity of the colour in the matrix corresponds to similarity between populations indicated on the right and bottom of the matrix. Populations are denoted with circles of the same colours as in **a**. The chromosomes are coloured by the most common (over 2%) relatedness matrices. **f**, Haplotype genealogy graph of mitochondrial sequences. The size of the circles indicates the number of individuals with the specific haplotype, with the smallest circles corresponding to one individual.

The phylogenetic clustering varied spatially over the genome (Fig. 1e and Supplementary Table 8). As loci that are differentiated between Italian sparrow populations are under stronger selection than a random sample (Supplementary Table 10), the variation in phylogenetic signal is likely to at least partly be influenced by selection. We also found significantly higher Spanish sparrow introgression in the Sicily and Malta populations, compared with the populations from Crete and Corsica (Patterson's D: ANOVA $F_{3,115}=22.52$, $P<0.001$; Supplementary Table 9). Ancestry painting, based on loci fixed for different alleles in the parent species, shows differences in heterozygosity among the Italian sparrow populations, with Crete having the highest and Malta having the lowest level of heterozygosity (ANOVA; $F_{3,46}=9.34$, $P=6.2\times10^{-5}$; Supplementary Fig. 4 and Supplementary Tables 11 and 12). The non-recombining mitochondrial DNA was similar to that of the 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].

We used two tests to investigate whether the island populations were likely to represent independent replicates, both developed in an earlier study[17]. The first tests whether signals of introgression are correlated among identical genome windows for each of the islands. We find correlations of f_d between 0 and 0.37 for the same genome windows among Italian sparrow population pairs (Supplementary Fig. 5a and Supplementary Table 13). These low correlations are consistent with independent hybridization events or long periods of independent evolutions with no gene flow. In comparison, a previous study[17] found no pairwise correlations <0.617 and a conspecific pairwise correlation >0.8 in Lake Victoria and Lake Kivu cichlids, and concluded that they represent a single hybridization event. Our second test detected significant differences in ancestry tract lengths among island populations ($F_{3,124894}=545.4$, $P=2.2\times10^{-16}$, Supplementary Fig. 5b), further supporting independent genome evolution in the island populations.

To address whether differentiation within a hybrid species is likely to result from positive selection or whether differentiated areas mainly arise as a by-product of long-term linked background selection in low-recombination areas in the parent species[18], we first investigated whether differentiation among Italian populations correlates with that between the parent species. If strong differentiation between Italian sparrow populations mainly occurs at positions that are highly differentiated between parent species, this could indicate that the island populations are different as a by-product of the existence of differentiation between parent species. If there is no correlation or if the standardized directly comparable slope of the regression between F_{ST} of the different population pairs (β) is lower between Italian sparrow population pairs, selection for different parental alleles could play a more important role. We then tested whether variation in differentiation reflects variation in recombination rate, as has previously been shown in flycatchers[18]. Differentiation among the Italian sparrow populations was not strongly correlated with differentiation between the parents (Supplementary Fig. 6). β (which is not influenced by overall levels of differentiation) for the relationship between within-Italian sparrow differentiation and recombination rate was much shallower than that between the parent species differentiation and recombination rate (Supplementary Fig. 6 and Supplementary Tables 14 and 15); therefore, we find no strong evidence that the house sparrow recombination rate accounts for the genomic heterogeneity observed within this hybrid species. Sorting of parental variants, potentially due to Hill–Robertson effects[19], may have contributed to this reduction in differentiation in low-recombination regions compared with that of the parent species. Regions with high differentiation may instead reflect divergent selection, or, alternatively, the recombination landscape may be altered following hybridization and spur differentiation in different areas of the genome in the hybrid than in its parents.

To identify outlier loci that are highly differentiated among all populations of the hybrid Italian sparrow and may have been subjected to, for example, divergent selection, we extracted the genomic windows in which the island populations were most differentiated (measured as relative similarity to the parent species in terms of F_{ST}). Among the genes found in the 1% 100-kb windows that are most differentiated between the island populations, 89 different gene ontology (GO) categories were over-represented relative to the rest of the genome (Supplementary Table 15). To test whether the significant gene ontologies found within the outlier regions are likely to be observed by chance, 30 GO permutations on randomly selected sets of genes were also run. For all GO analyses, we considered only GO terms that were not present in any of the 30 permutations as significant, as these significant GO groups are unlikely to be identified randomly. As differentiation in outlier genes may be hard to interpret without follow-up studies of their specific effects, we have focused on the GO terms and candidate genes for which there is such information from relevant study systems for all outlier analyses. Among the significantly over-represented GO categories were genes related to neuron function, including nervous system development, transmission of nerve impulse and synaptic transmission (Supplementary Table 16), suggesting that these categories of genes have been under divergent selection. The variation in Tajima's D was much higher for the outlier loci than for a size-matched set consisting of randomly selected loci, consistent with a role for selection in causing outlier differentiation (Supplementary Table 10). The high variation in Tajimas's D values implies that there are both high values consistent with balancing selection or population contraction, and low values consistent with directional selection among these outliers. Interestingly, genes related to neuron function have also been targets of recent positive selection in great tits[20]. The outliers also included *FGF10*, a gene explaining beak shape divergence between Darwin's finches[21]. Sicily and Crete were strongly differentiated at this beak shape candidate gene and the Italian sparrow has previously been shown to exhibit adaptive beak shape divergence[22] (Fig. 2a–c), although we have not been able to investigate the effect of the gene within a population with the same genomic background. A gene involved in feather development[23] and melanogenesis[24], *WNT7A*, was also a highly differentiated outlier among the otherwise genetically very similar (mean $F_{ST}=0.016$) Sicilian and Maltese populations[16] (Fig. 2d–f and Supplementary Table 16) that show considerable differences in plumage. These two examples underscore that repeated hybridization between the same parental species can generate locally adapted populations through reshuffling of parental alleles at biologically important loci.

To identify areas of unique Italian sparrow evolution, we targeted regions in which the Italian sparrow populations are differentiated from both parents by extracting the 1% of windows exhibiting the largest differences in F_{ST} between each Italian/parent comparison, keeping only windows overlapping between hybrid/parent in both comparisons, regions where the Italian sparrow populations are differentiated from both parent species. These regions have highly positive values of Tajima's D compared with the genome-wide average (Fig. 2g), consistent with strong selection in the Italian sparrow. High Tajima's D values result from an excess of medium-frequency alleles, consistent with balancing selection or a population bottleneck[25]. Low values suggest an excess of low-frequency polymorphisms (for example, from a selective sweep or population expansion)[25]. The high Tajima's D values for these outliers suggest that the Italian sparrow segregates for alleles that have undergone selection in the parents (Tajima's D is on average -1.66 ± 0.82 s.d. for house and -1.26 ± 0.85 s.d. for Spanish sparrows for these outliers) and subsequently have been subject to balancing selection

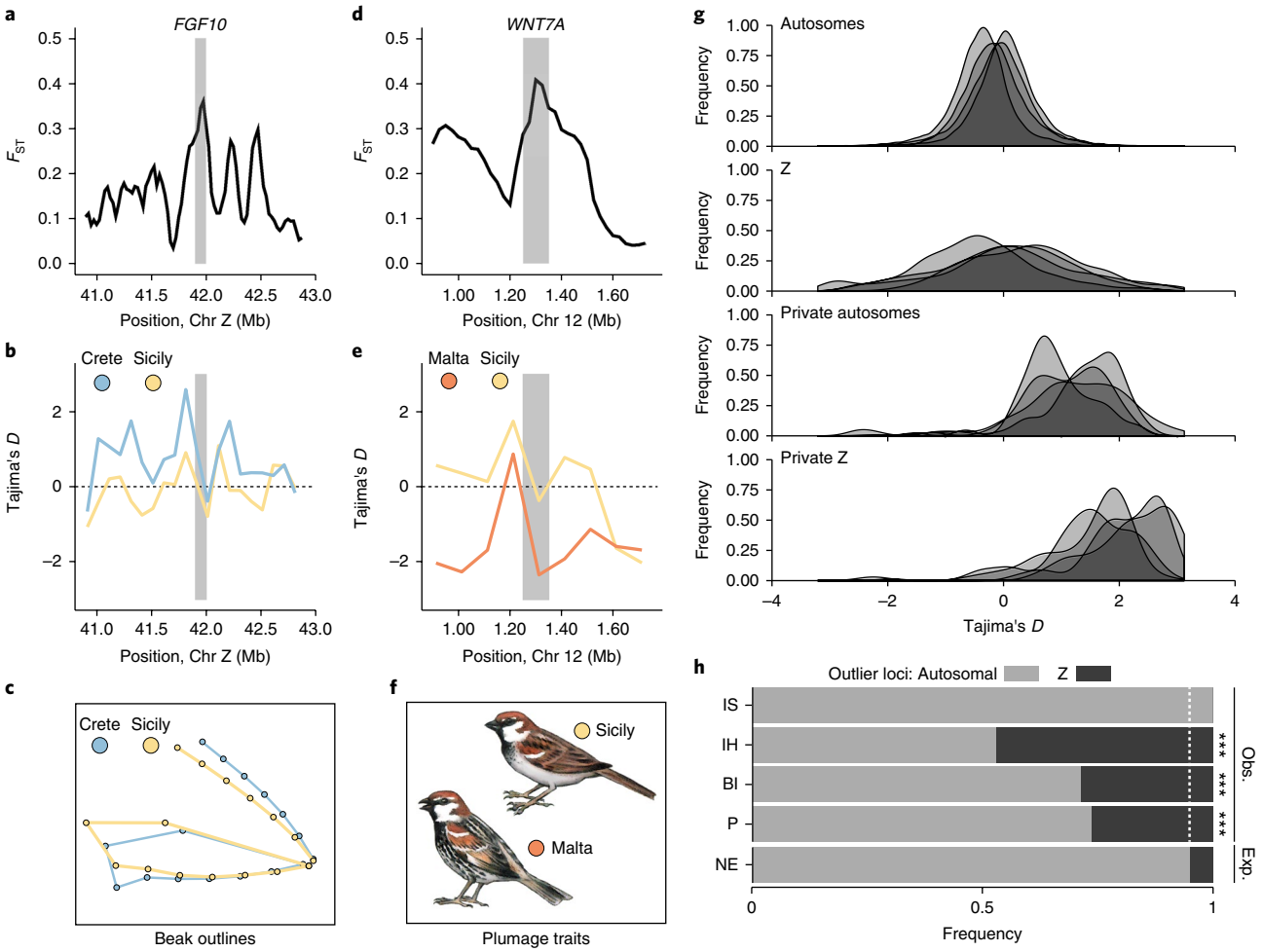

Fig. 2 | Local adaptation, private variation and strong selection on the Z chromosome. a, The beak shape candidate gene *FGF10* is differentiated between Italian sparrows on Crete and Sicily (genome-wide mean F_{ST} 0.20 ± 0.075 s.d.). **b**, Tajima's *D* for the *FGF10* region. **c**, Beak shape differences between Cretan and Sicilian sparrows. **d**, The plumage candidate gene *WNT7A* is differentiated between genetically similar sparrow populations on Sicily and Malta (genome-wide mean F_{ST} 0.13 ± 0.05 s.d.). **e**, Tajima's *D* for the *WNT7A* region. **f**, Schematic illustration of plumage differences between the populations. **g**, Strength of selection on the autosomes, the Z chromosome and on private outlier loci where the Italian sparrow is differentiated from both parent species. **h**, Distribution of outlier genes between the Z chromosome and the autosomes. IS denotes invariably Spanish, IH invariably house, BI between islands, P private and NE neutral expectations based on the whole genome. The lower N_e of the Z chromosome may contribute to the high representation of loci located on the Z chromosome among outliers.

within the hybrid lineage. We found five over-represented GO categories among the genes differentiated between Italian sparrow populations and both parent species (Supplementary Table 17). These GO categories include circadian rhythm and entrainment of the circadian clock, and the associated genes showed strong signals of stabilizing selection (Fig. 2g) and elevated linkage disequilibrium (Supplementary Table 18). These results illustrate how population-specific selection in concert with the parental mosaic is able to form unique features in the genomes of hybrid populations.

The level of differentiation between the Italian sparrow and the parent species is elevated on the Z chromosome compared with autosomes, even when accounting for the elevated parental differentiation on the Z chromosome (mixed model using data downsampled to every 2,500 kb to conservatively exclude all windows in linkage disequilibrium[9], with parental differentiation as a random factor and F_{ST} $F_{3,3141} = 7704.3$; $P < 2.2 \times 10^{-16}$ and whether the F_{ST} is against the house or Spanish sparrow as fixed effects; $F_{2,3141} = 4131.5$; $P < 2.2 \times 10^{-16}$; Supplementary Figs. 7 and 8). This is expected on the basis of the lower effective population size of this sex chromosome and hence elevated rates of genetic drift[26]. However, increased differentiation on the Z chromosome is also expected from the faster X(Z) effect of elevated rates of adaptive evolution on the macro sex chromosome due to hemizygous exposure[27]. Outlier loci are strongly over-represented on the Z chromosome (Fig. 2h; all $P < 0.001$; Supplementary Table 19), except for the outliers in the category where Italian populations invariably had inherited Spanish sparrow alleles (see Methods), none of which resided on the Z. Interestingly, Tajima's *D* estimates for the Z chromosome have significantly higher variance than those for the autosomes (Fig. 2c; repeated-measures ANOVA $F_{1,3} = 56.94$; $P = 0.005$; Supplementary Figs. 9 and 10) although this may partly be explained by the lower effective population size (N_e) of the Z chromosome. Moreover, dn/ds was higher on the Z chromosome compared with autosomes (goodness of fit $P < 0.001$ for fixed differences against both house and Spanish sparrows; Supplementary Table 20). The smaller N_e of the Z chromosome and the reduced recombination rate will increase the probability of loss of chromosomes without deleterious mutations by drift (Muller's ratchet), such that fixation of deleterious mutations in coding positions occurs more rapidly, increasing the dn/ds ratio[28]. However, taken

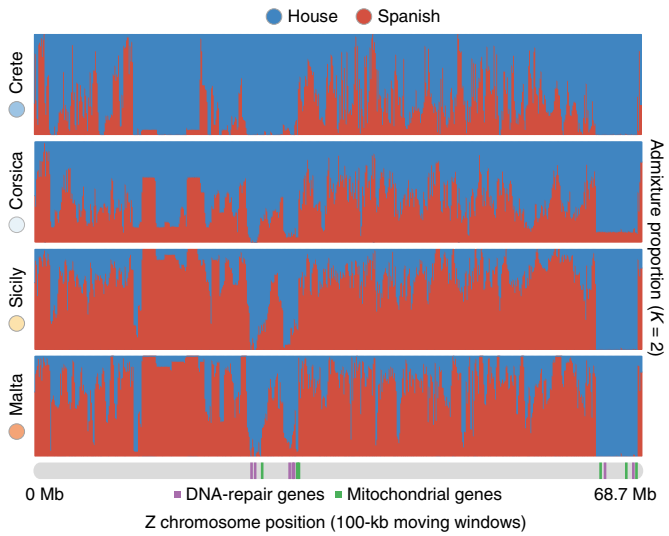

Fig. 3 | Parental similarity across the Z chromosome. Sliding-window ADMIXTURE analysis of the probability of house sparrow (blue) or Spanish sparrow (red) inheritance over the Z chromosome for the four Italian sparrow populations. Areas highly constrained to house sparrow inheritance harbour a significantly higher proportion of DNA-repair genes (purple lines) and many mitonuclear (green lines) genes.

together, these findings are consistent with a role for selection in the strong Z chromosome differentiation.

Across taxa with heteromorphic sex chromosomes, introgression on sex chromosomes is reduced[6,10]. To detect loci potentially important in causing such reduction in introgression on the sex chromosomes, we identified regions invariably inherited from a specific parent across all populations. We summed the F_{ST} against the house sparrow across island populations and subtracted the summed F_{ST} against the Spanish sparrow before extracting the extremes at both ends of the distribution (the 2% of the 100-kb windows with values most different from 0; Supplementary Tables 21 and 22). Genes invariably inherited from the Spanish sparrow are much fewer than these invariably inherited from house sparrow, but include a candidate pigmentation gene *WNT4* (ref. [29]) and a gene involved in vision, *OLFML2B*[30] (Supplementary Table 22). We found strong evidence for genes invariably inherited from the house sparrow, especially on the Z chromosome. There were ten significantly over-represented GO terms among the genes invariably inherited from house sparrows and these included both DNA repair and response to DNA-damage stimulus. DNA repair was significantly over-represented among these genes ($P_{\text{DNArepair}} = 0.026$; Supplementary Table 21). There were 11 mitonuclear loci and although these were not generally over-represented ($P_{\text{mitonuclear}} = 0.11$), 7 were found in the areas on the Z chromosome strongly constrained to house sparrow inheritance, which is significantly higher than the expectation ($P_{\text{mitonuclear Z}} = 0.00000002$). The mitonuclear genes on the Z chromosome included the previously identified candidate incompatibility gene *HSDL2* (ref. [9]; Supplementary Table 23 and Fig. 3). There were also six DNA-mismatch-repair genes on the Z chromosome, among them the candidate incompatibility gene *GTF2H2* (ref. [9]) that is involved in nucleotide excision repair (Fig. 3). This suggests that hybrid genome formation is restricted to uniparental inheritance for DNA-mismatch-repair genes, and is consistent with previous findings of strong selection on some mitonuclear loci[9].

Whereas mitonuclear genes are known as drivers of reproductive isolation[9] 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[31] 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[32]. 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[9].

Discussion

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[33] has been suggested to be important for shaping hybrid genomes. We find that house sparrow DNA-repair genes are necessary for 'escaping the mass of unfit recombinants'[34], most likely due to epistatic interactions. We also confirm the findings that many mitonuclear genes are under stronger selection to be inherited from house sparrows than the rest of the genome. Genes invariably inherited from the Spanish sparrow are fewer but include genes 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 suggest that hybridization is a potent force for creating novel variation, as many different combinations of the parental genomes can arise in hybrids and allow for adaptive differentiation between isolated populations of the hybrid species. Importantly, we show that the variation is limited for DNA repair and genes with mitochondrial function. 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). At least ten individuals from each population or species were sampled to adequately represent the variation within populations. 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), and stored in Qiagen Elution Buffer (Qiagen). Whole-genome re-sequencing was performed with Illumina sequencing technology. An Illumina TruSeq gDNA 180-base-pair (bp) library was created and sequenced on the Illumina HiSeq 2000 platform with 100-bp read length and 3 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 4 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 European Nucleotide Archive repository under (Accessions ERS1988350–ERS1988400).

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 (ref. [35]) with the mem, -M and -R options. A sorted BAM file was produced using SAMTOOLS version 1.0 (ref. [36]) using view with the -b, -U and -s options and a pipe to the sort command. Duplicates were identified and filtered out with 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 of these tools, which are components of GATK 3.3.0 (refs [37,38]). Final sequencing

coverage of these final BAM files (excluding duplicates) was 8× per individual (minimum 5.33×, maximum 15.8×; 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 the GATK GenotypeGVCFs tool was run using standard settings to achieve joint genotyping. Two different versions of the VCF file were created: one with only variable sites (49,237,560 single-nucleotide polymorphisms (SNPs)), and one where all sites were called as specified by the --includeNonVariantSites option (1,040,518,317 SNPs). We did not use a training set of SNPs as there are no prior expectations for our non-model organism.

For both VCF files, indels were first filtered out using VCFtools version 0.1.12b (ref. [39]), and hard filtering according to the Broad Institute's recommendations was performed with bcftools-1.2 (ref. [36]) to filter out sequencing artefacts. This included requiring a QualByDepth ≥ 2.0 and a FisherStrand phred-scaled *P* value of <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 ≥ 40.0 to ensure high mapping quality of the reads across all samples, a MappingQualityRankSumTest value of >−12.5 to exclude reads where the alternative allele has a lower mapping quality than reads with the reference allele, and finally a ReadPosRankSumTest value <−8.0 was required to ensure that reads with the alternative allele were not shorter than those with the reference allele, potentially indicating sequencing artefacts. In addition, we filtered out sites with a mean number of reads per individual <3, with a genotype quality <20 or with a mapping quality lower than 20 using VCFtools version 0.1.12b (ref. [39]). All allele-frequency-based methods were performed with ANGSD (ref. [40]), a tool that is particularly suited for low-coverage data that is applied directly to the BAM files, and hence the filtering options will not affect analyses that may yield spurious results at low coverage. To avoid paralogues, we also excluded sites with read depths above five times the variance of coverage. This left a total of 38,341,426 sites for further analysis.

Principal component analysis. Principal component analysis was performed using ANGSD (ref. [40]) version 0.911 and ngsTools version 1.0.1 (ref. [41]). This pipeline was chosen because it does not rely on genotype calls but instead takes allele frequency likelihoods and genotype probabilities into account[42]. We first estimated genotype probabilities from BAM files with ANGSD, including only bi-allelic sites and allowing a minimum mapping and site quality of 20 (Phred score) and a minimum coverage of 30× across all individuals. Principal component analysis 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. Eigenvalues for each principal component were then estimated from the covariance matrix produced by ngsTools. We used the broken stick criteria to assess which principal component axes were biologically informative from a simple scree plot and then extracted the covariance from these axes (Supplementary Table 3). The analysis was performed on three data sets: all sites, sites from the Z chromosome alone, and autosomal sites alone.

Population genomic analysis. Population genetic parameters were estimated for non-overlapping 100-kb windows along the genome, as this window size was larger than the distance of linkage disequilibrium decay[13]. Population genetic inference was based on genotype likelihoods whenever possible. ANGSD (ref. [40]) version 0.911 was used to estimate allele frequency likelihoods and to obtain a maximum-likelihood estimate of the unfolded site frequency spectrum for estimation of Tajima's *D*. Nucleotide diversity was estimated by dividing the pairwise Θ (population-scaled mutation rate) estimates by the number of variable sites per window. The ancestral sequence was reconstructed using genotypes from the outgroup tree sparrow. A Fasta file of the tree sparrow genome was obtained by using the -doFasta 2 command with the -GL 1 -doCounts 1, -setMinDepth 3 and -setMaxDepth 65 options in ANGSD (ref. [40]) version 0.911. Here, BAM files from 8 additional tree sparrows, sequenced to 8–12× depth and processed as described for the other samples above, were used. Genetic differentiation (F_{ST}) based on genotype likelihood was estimated on the basis of the two-dimensional site frequency spectrum using ngsTools (ref. [42]). All of these analyses were performed on the final BAM files. Sequence divergence (d_{xy}) was calculated from the VCF file with all positions called using the script developed in an earlier study[43] (https://github.com/johnomics/Martin_Davey_Jiggins_evaluating_introgression_statistics/blob/master/egglib_sliding_windows.py; version August 2014).

Admixture analysis. Genetic admixture was estimated using ADMIXTURE (ref. [44]) version 0.911. The VCF file was converted to plink's PED format using VCFtools version 0.1.12b (ref. [39]) and plink version 1.07[45]. Log likelihood values for *K*, the number of genetic clusters in the data sets, between $K=1$ and $K=8$ were estimated (Supplementary Table 4), and admixture analyses were run for the most appropriate value of *K*. Analyses were first run for a linkage-disequilibrium-pruned whole-genome data set (sites within a 50-SNP stepping window with a correlation coefficient higher than 0.1 were omitted; pruning of the BED file was performed with plink version 1.07 using the --indep-pairwise command; this left 438,443 sites for analysis). Sliding-window analyses with 100-kb windows were then carried out to investigate variability in the probability of parental inheritance across the genome. This analysis was performed individually for each island population together with the two parent species. The VCF file was then split into individual VCF files per 100 kb using the –L option in GATK 3.3.0 (refs [37,38]). Individual ADMIXTURE analyses for $K=2$ were then performed for each 100 kb BED file. A mean estimated cluster assignment probability for all individuals per population was computed for each analysis using a custom python script.

Phylogenetic analyses. To investigate whether phylogenetic relationships varied across the genome due to introgression or incomplete lineage sorting, a machine-learning approach implemented in SAGUARO version 0.1[46] was used to identify genomic regions characterized by distinct similarity matrices. To focus on breakpoints at which recombination led to changes in the topology of populations or species, rather than topological changes within populations, one representative high-coverage individual 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 1 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 Fig. 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 (ref. [47]). To this end, chromosome-length alignments were first phased using SHAPE-IT version 2 (ref. [48]). To improve phasing, this analysis was conducted with a subset consisting of the 6 individuals with the highest coverage per population (24 individuals in total), rather than just the 6 individuals used for SAGUARO analyses; however, the 6 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 generated with RAXML version 8.2.4 (ref. [49]) and the probability that the alignment is free of recombination determined with the Phi test[50]. We assumed that alignments with a low proportion of heterozygous sites are less likely to contain paralogous sequences, and that alignments with many parsimony-informative sites and high mean bootstrap support contain strong phylogenetic signal. Thus, alignments were selected according to the following 'relaxed' and 'strict' filters: a proportion of missing data below 0.2 (relaxed) or 0.1 (strict), at least 75 (relaxed) or 100 (strict) parsimony-informative sites, a proportion of heterozygous sites below 0.005 (relaxed) or 0.0025 (strict), a mean bootstrap support of at least 90 (both relaxed and strict), and a Phi test *P* value above 0.005 (relaxed) or 0.01 (strict). A total of 1,234 and 116 alignments were selected with these relaxed and strict filters, respectively. To include an outgroup for phylogenetic analyses with BEAST, consensus sequences of tree sparrow reads from the Naxos1 individual mapped to the house sparrow reference genome[13] were added to each selected alignment. To avoid bias towards the reference, missing data were not replaced by the reference alleles. The phylogeny of each alignment was then inferred with BEAST, using the GTR model of sequence evolution with estimated base frequencies, a Yule tree prior[51], and 50 million Markov chain Monte Carlo iterations. The ingroup, combining house sparrow, Spanish sparrow and the representatives of Italian sparrow populations, was constrained to be monophyletic. In the absence of a reliable absolute time line for sparrow divergences, the time of divergence of ingroup and outgroup was fixed at 1 time unit, so that all divergence ages within the ingroup are estimated relative to this initial split. Convergence of all Markov chain Monte Carlo chains was confirmed by effective sample sizes greater than 500 for all model parameters.

Introgression. The presence of introgression was estimated using Patterson's *D* (refs [52,53]) calculations, using the scripts provided in an earlier study[43]. ABBA–BABA estimates were calculated using a minimum coverage of 3, a 100-kb window size and 1,000 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 the tree sparrow as an outgroup.

Ancestry painting. To estimate heterozygosity levels and fixation of parental alleles, all sites for which parent species were fixed for different alleles were extracted. Specifically, a site was considered fixed if differentiation was complete ($F_{ST}=1.0$) between the parent species, and if alleles were observed in at least two individuals of both parents (missing data were allowed for other individuals). The genotypes of all Italian sparrow individuals were visualized for each of these fixed sites.

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 (refs [37,38]). Fitchi version 1.1.4 (ref. [54]) was used to reconstruct a haplotype genealogy based on Fitch distances.

Testing for independence. Following an earlier study[17], we used f_d correlations, based on the introgression measure f_d, calculated as the difference between ABBA and BABA patterns compared with the maximum possible difference[43] and ancestry block distribution to evaluate independence. f_d was estimated for 10-kb non-overlapping sliding windows along the genome using the scripts provided in a previous study[43]. Pearson's product-moment correlations of f_d between the Italian sparrow populations were estimated using the stats package in R. Sizes of blocks inherited from different parental populations, ancestry blocks, were estimated from the frequency of ABBA and BABA sites in non-overlapping windows of 3 kb following an earlier study[17]. Windows with ABBA/(ABBA + BABA) > 0.7 were considered candidate house sparrow ancestry windows, whereas those with a proportion < 0.3 were considered candidate Spanish sparrow ancestry windows. Ancestry block length was estimated from the number of consecutive windows (allowing for single windows with missing data) with the same ancestry (see ref. [17]).

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 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 only on autosomes.

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 (that is, shared differentiation amongst groups of populations) was used. The latter was calculated by performing principal component analysis on multiple pairwise comparisons of differentiation featuring the same focal species following an earlier study[55]. Common differentiation was estimated among: all Italian sparrow populations; between all Italian sparrow populations and the house sparrow; and between all Italian sparrow populations and the Spanish sparrow. In each case, all pairwise F_{ST} comparisons between these species were included, and the first principal component was 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 of population branch statistics[9]. Three categories of outliers were of interest. First, 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. Second, private outliers, windows in which Italian sparrows are diverged from both parent species, show where unique adaptation is putatively strongly selected for. Third, 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.

For between-island outliers, the 1% of 100-kb regions that 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 that are not dependent on these factors. Furthermore, F_{ST} was not strongly dependent on the recombination rate between Italian sparrow populations (Supplementary Fig. 4).

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, keeping only windows overlapping between both hybrid/parent comparisons. This was performed separately for each Italian population, and all outliers detected across the populations were then merged for gene ontology analyses.

For 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 the house sparrow, and by subtracting the sum of the F_{ST} values between these populations and the 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 extracted the 2% of the windows that deviated most strongly from 0 (Supplementary Fig. 10B), 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 analyses. One analysis was performed on all private outliers identified across populations, one on all outliers between populations, including all combinations of populations, one on outliers that resembled the house sparrow across all populations, and finally one on outliers that resembled the Spanish sparrow across all populations. As only 14 outliers resembling the Spanish sparrow across all populations were found, no significant GO terms were found for this analysis. Therefore, we do not provide a table of significant terms for this analysis. These analyses were performed using GO stat[56], with a human reference base. We implemented standard settings for GO analyses (that is, a value of 3 as the minimal length of the considered GO paths and no merging of GOs if the gene lists overlap). To test whether the significant GOs found within the outlier regions are likely to be observed by chance, 30 GO permutations on randomly selected sets of genes of the same size as the outlier sets were run. We considered only GO categories that were not present in any of the 30 permutations as significant. Mitonuclear genes were identified using MITOMINER 4.0 (ref. [57]) with a human reference database and standard settings. Over-representation of mitonuclear genes was subsequently tested using Chi-square tests. Corrections for multiple testing were performed with the Benjamini method[58].

dn/ds analyses. To test whether the Z chromosome was under stronger selection, synonymous and non-synonymous fixed differences within genes against each parent species for all autosomes and the Z chromosome were estimated for each Italian sparrow population. A goodness-of-fit test was applied to test whether the number of non-synonymous substitutions on the Z chromosome was higher than expected for each parent species. To this end, the R package PopGenome[59] was used. The splitting.data command in PopGenome was used to extract genes and fixed sites were extracted. Synonymous and non-synonymous sites were then identified using the options subsites = 'syn' and subsites = 'nonsyn', respectively.

Linkage disequilibrium decay. To address whether linkage disequilibrium was higher and decayed more slowly in outlier windows than in randomly selected windows, plink version 1.90b3b (ref. [45]) 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 to compare if linkage disequilibrium was higher and decayed more slowly in the outlier windows than in these randomly selected windows. A linear model was fitted for each outlier window, and intercept and slope were recorded and used in glm's to test whether linkage disequilibrium was higher and decayed more slowly in outlier windows than in randomly selected windows.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Code availability. All software versions and options needed to reproduce the results are specified in the Methods.

Data availability. The data generated and analysed during the current study are available in the European Nucleotide Archive repository http://www.ebi.ac.uk/ena under project PRJEB22939, accession numbers ERS1988350–ERS1988400. The raw data for the parental reference individuals have been deposited at the NCBI Sequence Read Archive under BioProject PRJNA255814 accession numbers SRR5369936–SRR5369966. The house sparrow reference assembly has been deposited at DDBJ/ENA/GenBank under accession MBAE00000000.

Received: 10 January 2017; Accepted: 4 December 2017;
Published online: 15 January 2018

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. Sankararaman, S. et al. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354–357 (2014).
7. Fontaine, M. C. et al. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* **347**, 1258524 (2015).

8. Baack, E. J. & Rieseberg, L. H. A genomic view of introgression and hybrid speciation. *Curr. Opin. Genet. Dev.* **17**, 513–518 (2007).
9. 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 (2014).
10. Martin, S. H. et al. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.* **23**, 1817–1828 (2013).
11. Qvarnström, A. & Bailey, R. I. Speciation through evolution of sex-linked genes. *Heredity* **102**, 4–15 (2008).
12. 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).
13. Elgvin, T. O. et al. The genomic mosaicism of hybrid speciation. *Sci. Adv.* **3**, 1–15 (2017).
14. Hermansen, J. S. et al. Hybrid speciation through sorting of parental incompatibilities in Italian sparrows. *Mol. Ecol.* **23**, 5831–5842 (2014).
15. Saetre, G. P. et al. Single origin of human commensalism in the house sparrow. *J. Evol. Biol.* **25**, 788–796 (2012).
16. Bache-Mathiesen, L. *The Evolutionary Potential of Male Plumage Color in a Hybrid Sparrow Species*. MSc thesis, University of Oslo (2015); https://www.duo.uio.no/handle/10852/45473
17. Meier, J. I. et al. Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nat. Commun.* **8**, 1–11 (2017).
18. Burri, R. et al. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome Res.* **25**, 1656–1665 (2015).
19. Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet. Res.* **8**, 269–294 (1966).
20. Laine, V. N. et al. Evolutionary signals of selection on cognition from the great tit genome and methylome. *Nat. Commun.* **7**, 1–9 (2016).
21. Lamichhaney, S. et al. Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* **518**, 371–375 (2015).
22. Eroukhmanoff, F., Hermansen, J. S., Bailey, R. I., Sæther, S. A. & Sætre, G.-P. S. Local adaptation within a hybrid species. *Heredity* **111**, 286–292 (2013).
23. Noramly, S., Freeman, A. & Morgan, B. A. Beta-catenin signaling can initiate feather bud development. *Development* **126**, 3509–3521 (1999).
24. Guo, H. et al Wnt/beta-catenin signaling pathway activates melanocyte stem cells in vitro and in vivo. *J. Dermatol. Sci.* **83**, 45–51 (2016).
25. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
26. Mank, J. E., Nam, K. & Ellegren, H. Faster-Z evolution is predominantly due to genetic drift. *Mol. Biol. Evol.* **27**, 661–670 (2010).
27. Charlesworth, B., Coyne, J. A. & Barton, N. H. The relative rates of evolution of sex chromosomes and autosomes. *Am. Nat.* **130**, 113–146 (2016).
28. Charlesworth, B. & Charlesworth, D. The degeneration of Y chromosomes. *Philos. Trans. R. Soc. Lond. B* **355**, 1563–1572 (2000).
29. Poelstra, J. W., Vijay, N., Hoeppner, M. P. & Wolf, J. B. W. Transcriptomics of colour patterning and coloration shifts in crows. *Mol. Ecol.* **24**, 4617–4628 (2015).
30. Tomarev, S. I. & Nakaya, N. Olfactomedin domain-containing proteins: possible mechanisms of action and functions in normal development and pathology. *Mol. Neurobiol.* **40**, 122–138 (2009).
31. David, W. M., Mitchell, D. L. & Walter, R. B. DNA repair in hybrid fish of the genus *Xiphophorus*. *Comp. Biochem. Physiol. C.* **138**, 301–309 (2004).
32. Greig, D., Travisano, M., Louis, E. J. & Borts, R. H. A role for the mismatch repair system during incipient speciation in *Saccharomyces*. *J. Evol. Biol.* **16**, 429–437 (2003).
33. Schumer, M. & Brandvain, Y. Determining epistatic selection in admixed populations. *Mol. Ecol.* **25**, 2577–2591 (2016).
34. Barton, N. H. The role of hybridization in evolution. *Mol. Ecol.* **10**, 551–568 (2001).
35. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
36. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
37. McKenna, A. et al The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
38. Van der Auwera, G. A. et al. From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **11**, 1–33 (2013).
39. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
40. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinforma.* **15**, 356 (2014).
41. 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).
42. Fumagalli, M. et al. Quantifying population genetic differentiation from next-generation sequencing data. *Genetics* **195**, 979–992 (2013).
43. 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).
44. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
45. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
46. Zamani, N. et al. Unsupervised genome-wide recognition of local relationship patterns. *BMC Genom.* **14**, 1–11 (2013).
47. Bouckaert, R. et al. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).
48. O'Connell, J. et al. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* **10**, e1004234 (2014).
49. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
50. Bruen, T. C., Philippe, H. & Bryant, D. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**, 2665–2681 (2006).
51. Yule, U. G. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Phil. Trans. R. Soc. Lond. B* **213**, 21–87 (1925).
52. Green, R. E. et al. A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
53. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
54. Matschiner, M. Fitchi: haplotype genealogy graphs based on the Fitch algorithm. *Bioinformatics* **32**, 1250–1252 (2016).
55. Burri, R. et al. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome Res.* **25**, 1656–1665 (2015).
56. Beissbarth, T. & Speed, T. P. GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics* **20**, 1464–1465 (2004).
57. Smith, A. C., Blackshaw, J. A. & Robinson, A. J. MitoMiner: a data warehouse for mitochondrial proteomics data. *Nucleic Acids Res.* **40**, D1160–D1167 (2011).
58. Benjamini, Y. Simultaneous and selective inference: current successes and future challenges. *Biom. J.* **52**, 708–721 (2010).
59. Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E. & Lercher, M. J. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* **31**, 1929–1936 (2014).

Acknowledgements

We thank M. Tesaker and BirdLife Malta for help with field work, L. Piñeiro and L. Bache-Mathiesen for providing morphological data, and A. Nilsson for comments on the manuscript. This work was funded by a Swedish Research Council post doctoral grant and a Wenner-Gren Fellowship to A.R. and a Norwegian Science Foundation grant to G-P.S. and A.R.

Author contributions

A.R. conceived the study, carried out field work and laboratory work, designed analyses, analysed data and wrote the manuscript. C.N.T. helped design analyses, and provided scripts, F.E. carried out field work and the gene ontology analyses, J.S.H. carried out field work and the final touches in figure preparation, M.M. performed the BEAST and Saguaro analyses and M.R. performed the recombination rate analyses and principal component analysis. T.O.E. provided the house sparrow reference genome, and G.P.S. identified the study system, designed the sampling strategy and carried out field work. All co-authors commented on the manuscript.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-017-0437-7.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.R.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.