Genomic islands of speciation harbor genes underlying coloration differences in a pair of Neotropical seedeaters

Tram N. Nguyen[1], Márcio Repenning[2], Carla Suertegaray Fontana[3,4], Leonardo Campagna[1,5,*]

1. Department of Ecology and Evolutionary Biology, Cornell University, 215 Tower Road, Ithaca, NY 14853, USA.
2. Universidade Federal do Rio Grande (FURG), Laboratório de Aves Aquáticas e Tartarugas Marinhas (LAATM). Av. Itália km 8, Campus Carreiros, 96203-900, Rio Grande, Rio Grande do Sul, Brazil.
3. Universidade Federal do Rio Grande do Sul (UFRGS). Laboratório de Ecologia de Comunidades e Populações, Instituto de Biociências, Avenida Bento Gonçalves, 9500, 91501-970, Porto Alegre, Rio Grande do Sul, Brazil.
4. Programa de Pós-graduação em Biodiversidade Animal, Universidade Federal de Santa Maria.
5. Fuller Evolutionary Biology Program, Cornell Lab of Ornithology, 159 Sapsucker Woods Road, Ithaca, NY, 14850, USA.
*Correspondence: lc736@cornell.edu

Short title: Islands of speciation in two Neotropical seedeaters

Abstract

Incomplete speciation can be leveraged to associate phenotypes with genotypes, thus providing insights into the traits relevant to the reproductive isolation of diverging taxa. We investigate the genetic underpinnings of the phenotypic differences between *Sporophila plumbea* and *S. beltoni*. *S. beltoni* has only recently been described based, most notably, on differences in bill coloration (yellow vs. black in *S. plumbea*). Both species are indistinguishable through mtDNA or reduced-representation genomic data, and even whole-genome sequencing revealed low genetic differentiation. Demographic reconstructions attribute this genetic homogeneity to gene flow, despite divergence in the order of millions of generations. We found a narrow hybrid zone in southern Brazil where genetically, yet not phenotypically, admixed individuals appear to be prevalent. Despite the overall low genetic differentiation, we identified three narrow peaks along the genome with highly differentiated SNPs. These regions harbor six genes, one of which is involved in pigmentation (*EDN3*) and is a candidate for controlling bill color. Within the outlier peaks we found signatures of resistance to gene flow, as expected for islands of speciation. Our study shows how genes related to coloration traits are likely involved in generating prezygotic isolation and establishing species boundaries early in speciation.

Key words

Ancestral Recombination Graph, Genome scan, Hybridization, Islands of speciation, *Sporophila*.

Introduction

Increasing knowledge about the genetic, ecological, and biogeographic contexts in which new species are formed is central to understanding the processes which lead to speciation. General patterns can be found by accumulating examples of how speciation has taken place throughout the tree of life [1-4]. The genetic architecture of reproductive barriers can be studied by leveraging non-model systems in the early stages of speciation and/or by focusing on hybridizing species where these barriers break down.

For recently diverged taxa, genetic differentiation tends to be localized around areas relevant to speciation, but with time, genomes will continue to diverge as other differences start to follow [5]. In such cases, natural hybridization or experimental crossing can help break up genetic linkage through recombination, and phenotypic traits can still be associated statistically to genetic changes [6]. Whereas studies following these types of designs have revealed how genomic landscapes are clearly heterogeneously differentiated [5], the processes driving such differentiation remain debated. Areas of high differentiation (or outlier regions) can be shaped by processes such as demographic range expansions [7], background selection combined with low levels of recombination [8], selective sweeps [9] or resistance to gene flow [10,11]. These processes can leave different signatures in the outlier genomic regions which researchers can examine with various traditional summary statistics, such as F_{ST}, D_{XY}, Tajima's D, and Pi, to uncover why genomes are heterogeneously differentiated [11]. However, clarity on the processes which have shaped many of these genomic regions is still generally lacking [8]. With the increasing feasibility of whole-genome resequencing, the toolkit for inferring evolutionary processes from genomic patterns continues to grow, offering researchers greater power to pinpoint regions of the genome under selection that mediate phenotypes of interest [12].

In birds, genome scans among recently diverged taxa have repeatedly uncovered coloration genes in areas of high differentiation in species that otherwise show shallow overall levels of genomic differentiation [10,13-17]. Collectively, these findings point to the importance of coloration differences in the early stages of speciation, and their likely effectiveness as mechanisms of prezygotic isolation [18-20]. Notably, *de novo* mutations in coloration genes, or the reassembly of different variants from standing variation across permeable species limits [21], may promote speciation. However, many taxa in the early stages of speciation may be ephemeral [22,23], with their persistence dependent on the accumulation of additional mutations over time, perhaps facilitated by the biogeographic context in which they were formed. Nevertheless, recently diverged species, whether they persist or not, will be informative of the types and location of genetic changes which can lead to initiating speciation.

The Neotropical genus *Sporophila* contains over 40 species and has the highest speciation rates of the species-rich Tanager family, to which it belongs [24,25]. This high speciation rate is in part the consequence of several small groups of rapidly radiating (recent and closely related) taxa in the early

stages of speciation, making the genus a valuable study system to understand the origin of reproductive isolation and the process of speciation. Many of these groups have been studied from a genetic perspective, and consistently show differentiation in male sexual traits (including coloration patterns) despite little genome-wide genetic differentiation [14,26,27]. The largest of such groups is known as the capuchino seedeaters, which contains 12 species differing most notably in the plumage patterns of adult males and in their vocalizations (a primarily cultural trait) [28,29]. Despite their phenotypic diversity, capuchino seedeaters show extremely low genetic differentiation, except for a few areas of the genome which are enriched for pigmentation genes. These genes likely mediate the different coloration phenotypes observed across males from different species [14,20,30], which in turn, impact species recognition [20]. Similarly, the Variable Seedeater superspecies complex (*S. corvina*, *S. intermedia*, *S. murallae*, and *S. americana*) and Morelet's Seedeater (*S. morelletti*) represent additional cases of plumage diversity, yet shallow genetic differentiation within *Sporophila* [26,27,31,32]. Overall, these examples suggest that localized genetic differences may mediate divergence in coloration traits, contributing to reproductive isolation and speciation [20].

Another *Sporophila* species that is not part of the previously mentioned groups, the Plumbeous Seedeater (*S. plumbea*), was traditionally thought to contain black and yellow-billed individuals. However, this presumed intraspecific variation was not a common polymorphism and did not correlate with seasonality or age (Figure 1A and 1B, [33]). In 2013, the yellow-billed individuals were shown to constitute a new species, the Tropeiro seedeater (*S. beltoni*), that bred in the grassland highlands of southern Brazil [34,35]. *S. beltoni* is differentiated from *S. plumbea* primarily through bill coloration, but also through male size and bill shape, vocalizations, migration phenology and subtle aspects of male plumage coloration. Furthermore, both species also differ in their breeding ranges, habitat, and phenology [34,35]. *S. beltoni* is sexually dimorphic and shares a narrow contact zone (~50-100 km) in the north of its distribution with the southernmost breeding population of *S. plumbea* (Figure 1A). In this contact zone, individuals are mostly segregated by habitat and elevation, though they can breed in close proximity where their respective habitats are available [34]. Although hybrids have not been definitively identified in this region, a few individuals with irregularly colored and streaky bills were observed, raising the possibility that hybridization occurs in this contact zone. However, a preliminary study using reduced-

representation genomic data failed to distinguish these two species [35], suggesting they were of very recent origin and/or continued to experience high levels of gene flow. Together, the phenotypic difference and low genetic differentiation between this species pair provides an opportunity to further investigate the process of speciation.

Here we assessed the degree of genetic differentiation between *S. beltoni* and *S. plumbea* using whole-genome resequencing, and conducted the first study to search for genomic regions associated with their phenotypic differences. Using the genomes of individuals from each species across both their allopatric ranges (i.e., outside of the contact zone) and within the contact zone, we conducted a demographic reconstruction to understand the demographic patterns in these different geographic regions. We then performed a genome scan which highlighted three narrow regions containing genes that likely mediate differences in beak coloration. We subsequently used genealogy-based statistics derived from topologies extracted from the ancestral recombination graph (ARG) to infer the processes that have shaped genomic outlier regions, and found signatures of reduced gene flow, as is expected for islands of speciation. Our study contributes to the emerging body of literature showing differentiation in genes related to bird coloration, which may mediate prezygotic isolation, before speciation is complete.

Results

Low genomic differentiation between *S. beltoni* and *S. plumbea*

While *S. beltoni* and *S. plumbea* individuals sampled outside of the contact zone (n=11 for each species) clustered separately in a PCA derived from ~15 million genome-wide SNPs, those sampled within the contact zone resembled each other, overlapping in the PCA (n=7 for each species; Figure 1C, Figure S1). Moreover, individuals could not be assigned to species using mtDNA (cytochrome c oxidase subunit 1 sequences), irrespective of sampling location (Figure 1D). Although the overall level of genome-wide differentiation was consistently low across groups, it was higher when comparing between allopatric individuals versus between sympatric individuals (F_{ST} = 0.0031 for allopatric, ~0 for sympatric, and on average 0.0015 for all samples combined). Using these genomic data, we then reconstructed the history

of divergence between the two species. We found similarly large current effective population sizes for both taxa (in the order of millions of individuals), which were much larger than the size of the inferred ancestral population (~5 fold; Figure 2A). The divergence time between *S. beltoni* and *S. plumbea* was estimated to be between ~3.2 and 3.6 million generations (Figure 2B), and the genetic similarity between the species could be partly explained by high levels of gene flow since their split (between 2 and 4 migrants per generation; Figure 2C). There was some variation in the parameter values obtained from models using birds sampled from different parts of the range (allopatric, sympatric, or a combination of both; Figure 2). However, the most notable difference was the direction of gene flow (Figure 2C), which we interpret with caution as it can be particularly challenging to infer under certain scenarios [36]. While migration was inferred from *S. beltoni* into *S. plumbea* within the contact zone, the opposite was true for allopatric individuals (or when all samples were combined). It is possible that the patterns of hybridization may be different within and outside of the contact zone, with the former being representative of more recent gene flow (see Discussion).

Differentiated SNPs are concentrated in three outlier peaks

When comparing all samples, a small number of SNPs showed differentiation above F_{ST}=0.20 (1.1%) and only 40 SNPs (out of ~15 million) had an F_{ST} value above 0.7 (Figure 3A). Of the 40 variants showing the highest F_{ST} values, 39 were clustered in three divergence peaks: on contigs 404 (24 elevated SNPs in ~100kb on the Z chromosome), 33 (9 elevated SNPs in ~66 kb on chromosome 20), and 382 (6 elevated SNPs in ~2 kb on chromosome 11) (Figure 3A). Taken together, the variants from these peaks could be combined to cluster individuals by species in a PCA (Figure S1), and the same was true when distinguishing allopatric individuals using the variants from each peak separately (Figure 3B). However, with the exception of contig 33, sympatric individuals were genetically similar to each other and could not always be assigned to species based on their genotypes in these genomic regions (Figure 3B, Figure S2). The 39 variants showing the highest F_{ST} values in these peaks were in high linkage-disequilibrium, both within but also between peaks, suggesting they are co-inherited despite being on different chromosomes (Figure S3). Moreover, each species showed a common haplotype on each peak, with less

intraspecific variation compared to interspecific differences (Figures S4-S6). Only a few sympatric individuals were heterozygotes or sometimes homozygotes for haplotypes from the other species. We did not find individuals that showed admixture at all three peaks simultaneously (Figures S4-S6).

EDN3 and _PRLR_ are candidate genes implicated in the phenotypic differences between species

The divergence peaks we identified involved a total of six genes (Figure 3B, Table 1). Two of these genes, *EDN3* which encodes the protein Endothelin-3, and *PRLR* which encodes the Prolactin receptor, may mediate the difference in beak coloration between *S. beltoni* and *S. plumbea* (Table 1; see Discussion). The remaining genes could be related to differences between these taxa that are not immediately obvious to us (Table 1). The variants showing the highest F_{ST} values in these peak regions were mostly non-coding (74% fell in intergenic regions and 13% in introns; Figure 3B), however two were responsible for non-synonymous mutations in the *HSDL1* gene on contig 382. Three additional mutations fell within *EDN3* but our annotation of that region was not of sufficient quality to distinguish if they fell within introns or in exons.

Outlier genomic regions are speciation islands

We explored the divergence peaks to search for signatures that could reveal the processes involved in shaping these regions. We first looked at Tajima's D and nucleotide diversity, which tend to be reduced after selective sweeps. Although Tajima's D was on average negative and nucleotide diversity was reduced in the three peaks, more extreme values could be found outside of the peaks within the same contigs (Figure S7). Therefore, based on these statistics alone, it is not conclusive whether selective sweeps have occurred within these outlier peaks. We next looked at H12 and H2/H1 which can be used to distinguish hard and soft selective sweeps, but again did not find conclusive patterns (Figure S7). We also compared F_{ST} and D_{XY} values and observed slightly elevated D_{XY} in some windows within the F_{ST} divergence peaks (compared to both windows outside of the F_{ST} peaks or in contigs without F_{ST} peaks), suggesting these regions may have accumulated some absolute sequence differences between

the species (Figure S8). This prompted us to search for signatures of selection in more depth using three statistics derived from topologies extracted from the ancestral recombination graph (ARG) (Figure 4A). The species enrichment score measures the probability of observing clades containing individuals from a certain species [9]. The relative TMRCA half-life, RTH' [37], is a normalized version of the time to most recent common ancestry (TMRCA). Finally, the inter-species cross-coalescence measures the average age of the most recent coalescence events between tips in the tree which belong to the different species. We followed the reasoning outlined by Hejase et al. [9], where areas under selection are expected to show species enrichment, but these could be the product of either species-specific and recent selective sweeps, or they could be older islands of speciation that resist gene flow. These two extreme scenarios can be distinguished by using RTH', which we expect to be low when selective sweeps produce shallow clades, and the inter-species cross-coalescence, which we expect to show larger values when gene flow is selected against (i.e., delayed cross-coalescence [9]). Following this logic, we obtained values for the three statistics in the divergence peaks and for a control set of contigs which represented ~10% of the genome, and we used the latter to establish species-specific thresholds of significance for each statistic. We also conducted this analysis for all samples together, and for sympatric and allopatric samples separately. The analysis shows genealogies within the peak regions which are significantly enriched for each species (Figure 4B, Table 2, Table S2, Figures S9-S12). These clades show low RTH' values, sometimes for one species but in some cases for both, and for the peaks on contigs 404 and 382 we also observed delayed cross-coalescence (Table 2, Table S2). These patterns were accentuated when conducting the analysis with samples from outside of the contact zone and generally weaker or lost when using sympatric birds from within the contact zone. Taken together, these regions show evidence of having undergone selection and experiencing reduced migration compared to other areas in the genome, especially when analyzing allopatric samples which may be more representative of historical processes (see Discussion).

Discussion

Whole-genome sequencing of *Sporophila beltoni* and *Sporophila plumbea* revealed little genetic differentiation between this species pair despite their marked phenotypic differences, similar to what has been found in other *Sporophila* species [14,26,32]. The highest levels of differentiation were concentrated in three narrow peaks, containing only six genes, which are candidates for mediating the phenotypic differences between our focal taxa. However, unlike the other *Sporophila* species which have shown this pattern of genetic homogeneity [9,14], we inferred *S. beltoni* and *S. plumbea* to be comparatively old, in the order of 3 million generations (~1 Myr). For comparison, the 10 species of southern capuchino seedeaters (those found south of the Amazon River [38]) began to diverge within a narrow window over the last million years [39]. Our demographic modelling suggests high levels of gene flow, and not recent origin, as a possible explanation for the genetic homogeneity between *S. beltoni* and *S. plumbea*. This scenario of relatively deep divergence, high overall gene flow, and a small number of highly differentiated regions suggests that these divergence peaks could be islands of speciation [40], i.e., regions of the genome that resist homogenization through gene flow. In contrast, in capuchino seedeaters most divergence peaks, which are also enriched for coloration genes, were found to be shaped through recent species-specific selective sweeps [9]. Therefore, divergence in regions of the genome containing pigmentation genes has likely evolved through different processes in *S. beltoni*/*plumbea* and capuchino seedeaters [9,14].

The two outlier peaks with the highest levels of differentiation and encompassing most of the highly differentiated SNPs (on contig 404 and contig 33), contained only three genes, two of which have functions that may mediate the phenotypic differences between our focal taxa. *EDN3* functions in melanoblast proliferation and differentiation, and mutations (including duplications) in either the *EDN3* gene itself or its receptors (*EDNR*s) can lead to hypo or hypermelanization in quails and chickens [41], and to several piebalding phenotypes in domestic pigeons [42]. These differences in pigmentation have been shown to affect not only feathers, but also other body parts, like the chicken comb [43] or even internal organs [41]. The Prolactin receptor (*PRLR*) is implicated in keratinization and feather formation [44] and Prolactin itself is known to stimulate molt [45] and is involved in coloration phenotypes in other taxa [46]. It is therefore likely that these genes are involved in mediating the coloration differences between *S. beltoni* and *S. plumbea*, the most notable of which is in the beak. These differences in beak

color may contribute to promoting prezygotic reproductive isolation, as has been found between subspecies of Long-tailed Finches (*Poephila acuticauda*) [47,48]. Within the genus *Sporophila,* species have either gray, black, yellow or orange bills [24]. The clade containing *S. plumbea* and *S. beltoni* has a total of six species (also including *S. albogularis*, *S. falcirostris*, *S. collaris* and *S. schistacea*) [49], and four of these six species have orange bills, suggesting that the black melanized bill is possibly a derived trait in *S. plumbea*. More work is needed to understand if *EDN3* and *PRLR* mediate beak coloration across multiple *Sporophila* species, as is the case with the *BCO2* gene and carotenoid-based coloration in *Setophaga* warblers [50].

It was previously unclear whether *S. beltoni* and *S. plumbea* hybridized in their contact zone in Southern Brazil (Figure 1A, [34]), mainly due to the lack of individuals with a clear mixed phenotype in the wild. However, our study shows extensive admixture in this region, likely due to ongoing hybridization. Therefore, this narrow area of contact likely constitutes a hybrid zone. This hybrid zone coincides with a region where the breeding habitat of both species is degraded due to human activities, an aspect which may have contributed to increasing the overlap between the breeding ranges of both species and resulted in high levels of recent gene flow [51]. It remains to be determined whether this habitat alteration, and the possibly associated increase in gene flow, remains restricted to this narrow contact zone or if it may have consequences for the completion of speciation. Within the divergence peaks, sympatric individuals were more similar to each other than allopatric individuals, and in some analyses (e.g., PCA or admixture plots), some were indistinguishable (Figure 3B, Figure S1-S2). The exception was the peak on contig 33 which contained the *EDN3* gene that may be involved in controlling beak coloration, allowing the correct species identification in the contact zone (Figure 3B). Heterogeneous genomic landscapes tend to show a larger number of peaks outside of contact zones, where gene flow is reduced [52-54]. This increased level of divergence could be due to selection on traits in allopatry, unrelated to reproductive isolation and where gene flow does not occur [52]. If this is true for our focal taxa, the lack of admixture in *EDN3* in both allopatry and sympatry constitutes further evidence that this gene may mediate phenotypic differences relevant to reproductive isolation. However, we note that the genome-wide level of differentiation between *S. beltoni* and *S. plumbea* is very low in allopatry (F_{ST} = 0.0031), and that we infer high levels of genome-wide gene flow when analyzing sympatric or allopatric samples (Figure 2C). It is

therefore possible that the three peaks of differentiation that we have identified have resisted gene-flow during speciation. This prompted us to search for signatures of different evolutionary processes which could have shaped our candidate genomic regions.

We did not find clear patterns using traditional summary statistics (e.g., Tajima's D, nucleotide diversity, D_{XY}) within the peak regions that would allow us to draw robust conclusions about the processes that have shaped them. We therefore leveraged the ARG and genealogy-based statistics that would potentially give us greater resolution and allow us to distinguish between two different extreme scenarios which could elevate F_{ST}: recent and species-specific selective sweeps and resistance to gene flow [9,17]. Regions which have undergone recent species-specific selective sweeps are expected to show clades containing most individuals from the relevant species that are significantly shallower (i.e., showing reduced diversity) than other areas of the genome. Islands of speciation, on the other hand, should show signatures of reduced gene flow (i.e., delayed cross-coalescence) when compared to other regions of the genome. We note that these two extreme models are not mutually exclusive, and a locus could undergo a selective-sweep after which gene flow between species would be reduced [9]. This latter scenario is supported by our data in the peak regions, where we tend to see comparatively shallow clades for one or both species, as expected for loci that have undergone selective sweeps. Moreover, we also see evidence for delayed cross-coalescence (primarily outside of the hybrid zone, but also more generally when the p-value thresholds are not as stringent; Table 1 vs. Table S2). Taken together, our findings support the islands of speciation model [11,40], where a few loci that may be relevant to generating coloration differences and prezygotic isolation, have undergone selection and resisted gene flow.

Our study represents the first investigation into the phenotypic differences between *S. beltoni* and *S. plumbea* and shows how coloration genes can diverge early in speciation and likely mediate prezygotic isolation. We also exemplify how the ARG can be leveraged to distinguish between competing processes which shape genome evolution, finding evidence of reduced gene flow in speciation islands compared to the rest of the genome. Taken together, our genome scan, demographic reconstructions, and ARG-based analyses reveal the complicated interplay between selection and gene flow in shaping species boundaries.

Materials and methods

Sampling and dataset

We sampled a total of 36 individuals for whole genome resequencing, 18 *S. beltoni* and 18 *S. plumbea*. Samples originated from allopatric portions of the ranges of both species (n=11 for each species), as well as from the contact zone (n=7 for each species) (Figure 1; Table S1). For this study we only used adult males so that identification would be unambiguous, and so that our depth of coverage for autosomes and sex chromosomes would be equivalent. Birds were captured during breeding seasons (austral spring/summer) between 2008 and 2014 using mist nest, and subsequently banded and bled before being released. To conduct fieldwork in Brazil and collect samples, we secured licenses from the National Center for Research on the Conservation of Wild Birds (CEMAVE, licenses 3711 and 3778) and the Instituto Chico Mendes of Conservation of Brazilian Biodiversity (ICMBio), through the Brazilian System and Information on Biodiversity (SISBIO, license numbers: 13310, 36881, 35434, and 52714).

Sequencing and variant discovery

We used the DNEasy blood and tissue kit (Qiagen, CA, USA) to obtain DNA from blood and prepared libraries following the protocol recommended for the TruSeq Nano DNA library preparation kit protocol (550 bp inset size). Libraries were pooled into two groups of 18 samples, using concentrations of adapter-ligated DNA (determined through digital polymerase chain reaction), and each group was sequenced on its own Illumina NextSeq 500 lane at the Cornell Institute for Biotechnology core facility. This produced a total of ~1.6 billion 151 bp paired-end reads, and consequently based on the number of raw reads obtained for each individual, we expected an average depth of coverage of 5.6 +/- 1.1. We assessed sequence quality and performed filtering and adapter removal as described previously [17]. We first trimmed low-quality bases from individual reads and collapsed overlapping paired reads using AdapterRemoval version 2.1.1 (options: --trimns --trimqualities --minquality 10) [55]. We then aligned the filtered sequences from each individual to the *Sporophila hypoxantha* reference genome [14] using

Bowtie2 version 2.4.3 [56] with the very sensitive local option. Despite using a reference genome from a different, yet closely related species, the average alignment rate was 96% (Table S1). The average depth of coverage across all samples and sites, calculated using Qualimap v2.2.1 [57], was 5.4 +/- 1.1x (Table S1). We followed the genotyping pipeline described in detail in [17] and filtered the resulting SNP dataset with VCFtools version 0.1.16 [58]. Briefly, we first marked PCR duplicates and realigned around indels using “MarkDuplicates” and “IndelRealigner”, respectively. We subsequently used the “Haplotypecaller” and “GenotypeGVCFs” modules from GATK version 3.8.1 [59] to identify variants and conduct joint genotyping across samples, resulting in a single variant file for the entire dataset. We then selected SNPs with the “SelectVariants” module of GATK (option: --selectType SNP) and filtered out those that did not satisfy the following filters: QD < 2, FS > 60.0, MQ < 30.0, ReadPosRankSum < -8.0. Finally, we performed additional post-hoc filtering using VCFtools to retain 15,663,387 variants present in at least 85% of individuals, with mean depth of coverage between 2 and 50 and a minor allele count of at least six. We explored the sensitivity of our dataset to a range of filtering parameters in different combinations, including up to a minimum depth of coverage of 5, a minor allele count of at least 10 and less that 5% missing data. These different filters produced congruent patterns in our PCAs and Manhattan plots, yet reduced the number of total SNPs retained. We therefore decided to retain a larger number of SNPs and, unless otherwise stated, use the dataset described above for downstream analyses.

Analysis of mitochondrial DNA

Because we did not recover high-quality mitochondrial genomes for every individual from the whole genome data, possibly due to miss-assemblies related to nuclear sequences of mitochondrial origin, we decided to sequence the cytochrome c oxidase I (*COI*) gene, which is commonly used for species identification [60]. We used primers BirdF1 and COIBirdR2 and followed the procedures described by Kerr et al. [61]. We Sanger sequenced PCR-amplified DNA using both the forward and reverse primers at the Cornell Institute for Biotechnology core facility. Forward and reverse sequences were combined into a consensus sequence for each individual and aligned using Geneious version 10.2.6 [62]. In total we obtained 27 *COI* sequences, 8 allopatric and 6 sympatric *S. beltoni* and 7 allopatric and 6

sympatric *S. plumbea* (Table S1). We used these sequences to construct a *COI* haplotype network using the R version 4.0.2 [63] package pegas v1.1 [64].

Genetic differentiation, admixture and summary statistics

We first assessed genome-wide patterns of differentiation among species and sampling locations (i.e., allopatric or sympatric) by conducting a Principal Component Analysis (PCA) in the R package SNPRelate version 3.3 [65]. Two samples originated from individuals that were later found to be related (TNN26 and TNN27, relatedness of 0.538) and separated from the remaining samples in a PCA (Figure S1A). We therefore excluded these samples from the PCA in Figure 1C, which produced the same pattern as when the samples were retained, and PC2 vs. PC3 were plotted (Figure S1B). We used the same programs to produce PCAs from the SNPs found within the divergence peaks, after subsetting the dataset by genomic location in VCFtools. We also calculated F_{ST} values for individual SNPs using VCFtools and displayed these in a histogram produced in R or as a Manhattan plot using the R package qqman version 0.1.8 [66]. For the latter, we simplified the plot by only visualizing SNPs with $F_{ST} > 0.2$. We also used VCFtools to subset the dataset by species and contig to calculate statistics base on the frequency of the most common haplotypes (H1, H2, H12 and H2/H1) in SelectionHapStats [67]. We used a window size (non-overlapping) of 25 SNPs and merged haplotypes differing in one position (--distanceThreshold 1). We also used VCFtools to calculate R2 values and estimate linkage disequilibrium (LD) among all the different sites showing the highest differentiation ($F_{ST}>0.7$) in the divergence peaks. For these same variants (24 on contig 404, 9 on contig 33 and 6 on contig 382), within each divergence peak, we explored local haplotypes by first phasing and imputing missing data using default parameters in BEAGLE version 3.3.2 [68]. We plotted the two haplotypes for each peak region from every individual using the function phylo.heatmap from the R package phytools [69], and clustered them according to similarity by producing a distance matrix in the R package vegan [70]. Finally, we explored ancestry along the contigs of interest by running Admixture version 1.3.0 [71] in 50 kb (non-overlapping) sliding windows using a K value of two. As the peak on contig 382 was narrow (~2 kb), we used a sliding window of 10 kb in this case, as a compromise between not averaging out across too many SNPs while maintaining

sufficient resolution. We plotted these values separately for allopatric and sympatric samples using a smoothing line in ggplot2 [72]. To calculate Tajima's D and nucleotide diversity (π) in the contigs of interest, we used a vcf file that followed the same filtering steps described above except for that it was not filtered by minor allele frequency (the same file used in the demographic reconstruction), as this would bias both summary statistics that are sensitive to low frequency alleles. We calculated these statistics in non-overlapping, 5 kb windows. Finally, we calculated D_{XY} using the program pixy version 1.2.7 [73]. For this specific analysis we included invariant sites by running GATK's "GenotypeGVCFs" module with the option "–includeNonVariantSite". Following filtering recommendations [73], we then separated our data into variant and invariant sites and filtered them independently, as applying population genetic based filters such as minor-allele frequency (MAF) filtering would result in the loss of our invariant sites. We omitted the MAF filter to obtain invariant sites, but all other filtering steps were implemented as described previously (e.g., missing data or coverage filters).

Demographic reconstruction

We estimated current and ancestral effective population sizes for both species, the splitting time between the taxa and migration rates (total of six demographic parameters) using an isolation with migration model as implemented in G-PhoCS version 1.3 [74]. Because this analysis is computationally intensive, we subsampled individuals from the dataset. We conducted three different analyses using individuals from different parts of the ranges of both species: allopatric (eight from each species), sympatric (seven from each species) or a combination of both types of samples (4 allopatric and 4 sympatric from each species). To avoid biasing our analyses by excluding low frequency alleles, we used a vcf file that was not filtered for minor allele frequency (~61M SNPs), but otherwise had the same filters applied as described previously. We subsequently discarded contigs that were smaller than 1 Mb, Z-linked, or had outlier SNPs (contig 404, 33, 382 and 910). We used the "FastaAlternateReferenceMaker" module in GATK to generate sequence files for each individual for the remaining contigs, and sampled and aligned 2,500 loci, each 1 kb in length and at least 100 kb apart. We ran G-PhoCS for 1 million generations (for the sympatric and allopatric analyses) or for 300,000 iterations (for the combined

analysis) and discarded the initial 50,000 as burn-in. We used the coda package in R [75] to check for convergence and subsequently converted median and 95% Bayesian credible intervals to generations or individuals (from mutation scale) by using an approximate mutation rate estimate of 10^{-9} per bp per generation [76]. We have focused our interpretations primarily on relative comparisons between the species which are independent of our assumption of mutation rate, as is the number of migrants per generation [74].

Identification of genes in outlier genomic regions

We defined F_{ST} outlier regions as the genomic coordinates encompassed by the SNPs showing $F_{ST} > 0.7$. By using this approach, we aim to focus on the regions showing the highest levels of differentiation between species, however this does not imply that other regions, showing more subtle patterns, are not relevant to shaping phenotypic differences. We had previously mapped every contig in the reference genome to its corresponding chromosome in the Zebra Finch [14]. Most of these outlier SNPs were clustered together and located in three peaks (contig 404, contig 33 and contig 382), yet there was a single additional elevated SNP on scaffold 910 (near the gene LRRC15), corresponding to chromosome 9. We did not consider this a divergence peak as it included a single position, and don't discuss it further. We searched for the genes within each divergence peak (+/- 5 kb) by inspecting our reference genome annotation using Geneious version 10.2.6 [62]. We used the NCBI database (http://www.ncbi.nlm.nih.gov/) to find information on the function of these different genes in other organisms. Finally, the outlier genomic regions were similar to the rest of the genome in average missing data (0.05 vs. 0.05), variant quality (QUAL; 1683.19 vs 1678.24) and depth of coverage (4.97 vs 5.00).

Statistics derived from the Ancestral Recombination Graph (ARG)

We estimated local trees by inferring ARGs using the arg-sample module in ARGweaver version 1 [37]. We ran ARGweaver on unphased data in 5-7 mb intervals which included the areas of high differentiation in the three contigs with outlier peaks, plus a control set consisting of 10 contigs totaling

~110 megabases. For the control set we broke large contigs into 5 mb fragments and ran the software following the settings described previously [17]. Briefly, we set the mutation and the recombination rates to 10−9/bp/gen, the effective population size to 500,000 individuals and the remaining parameters as follows: -c 5 -ntimes 20 -maxtime 1e7 -delta 0.005 -resample-window-iters 1 -resample-window 10000 -n 1000. We retained the last of 1,000 MCMC iterations, discarded the first and last 50 kb of each ARG block, and extracted trees every 500 bp. We used these trees to calculate three statistics as described in detail in [9] (Figure 4A). The species enrichment score measures the probability of observing subtrees of different sizes containing individuals from a certain species, and is calculated for both species and every subtree. For each tree we retained the maximum enrichment score obtained for a given species across all subtrees. A high enrichment score is obtained when a local tree has a clade with most or all individuals from that species. We note that when only two species are present in a tree like in this study, when one species enrichment score is high it means that individuals from the second species may also group together on the tree, and therefore both species enrichment scores will tend to be elevated. The relative TMRCA half-life [9,37], RTH', is a normalized version of the time to most recent common ancestry (TMRCA). To calculate this statistic, we divided the time to the most recent common ancestor of half of the haploid samples for a species by the age of the youngest subtree containing at least half of all the haploid samples (irrespective of species). This normalization controls for the variation is coalescence times in different trees across the genome [37]. Finally, the inter-species cross coalescence measures the average age of the most recent coalescence events between tips in the tree which belong to the different species. We added the ages of the ten most recent cross-coalescence events in each tree and normalized them in the same way we did to calculate RTH' [9]. We obtained these statistics for each tree (sampled every 500 bp) and then averaged those values across 5 kb non-overlapping windows. We generated empirical distributions for these statistics from the ~22,000 windows (~110 mb) obtained from the control contigs, and used these distributions to assess statistical significance as described previously [9,17]. The analysis was conducted separately for allopatric and sympatric samples, and a third time for all samples combined. We exported trees which show extreme enrichment scores for each species for illustration purposes. The code to calculate these three statistics and conduct these statistical tests is deposited in GitHub (https://github.com/CshlSiepelLab/bird_capuchino_analysis; see [9]).

Acknowledgements

We thank Bronwyn Butcher for help with library preparation and sequencing. This project was funded through the Athena grant from the Cornell Lab of Ornithology (to T.N.N.) and NSF DEB-2232929 (to L.C.). The Brazilian Federal Agencies (CNPq and CAPES) provided scholarships to M.R. The Neotropical Grassland Conservancy (NGC) supported field work. The Brazilian National Research Council (CNPq grants (303318/2013, 309438/2016) supported C.S.F. We would like to acknowledge the property owners who kindly authorized our fieldwork. We thank Renata Grieco for permission to reproduce the bird illustrations in Figure 1. The sequence data generated for this project is archived on GenBank (Bio Project PRJNA382416).

Author contributions

Designed project: T.N.N., M.R. and L.C. Performed fieldwork: M.R. and C.S.F. Obtained genomic data: T.N.N. Analyzed genomic data: T.N.N. and L.C. Supervised research: L.C. Wrote manuscript: T.N.N. and L.C., with edits from all authors.

Declaration of interests

The authors declare no competing interests.

Figure titles and legends

Figure 1: Genome-wide differentiation between *S. beltoni* and *S. plumbea*. (A) Range map for both species showing sampling localities in Brazil, and the area of contact (encompassed in the rectangle on the map and also in the topographic map in the inset) following [34]. Samples are color-coded by species and geographic origin (as shown in **C**). (**B**) Examples of adult male *S. beltoni* and *S. plumbea* individuals; photos by Márcio Repenning. (**C**) PCA derived from ~15 million genome-wide SNPs. (**D**) Haplotype network derived from sequences of the mitochondrial gene *COI*. Different haplotypes are connected

either by solid (based on the minimum spanning tree) or by dashed lines (indicating alternative paths). The number of mutational steps between haplotypes is shown as solid bars.

Figure 2: Demographic reconstruction. (**A**) Current and ancestral population sizes, (**B**) divergence time and (**C**) migrants per generation. Each set of parameters was estimated three times, once with a random subset of all samples combined and also after subsampling either exclusively allopatric or sympatric individuals. Bars indicate median values and 95% credible intervals.

Figure 3: Outlier peaks between *S. beltoni* and *S. plumbea*. (**A**) Manhattan plot showing F_{ST} values calculated using all individuals from both species. Note that for simplicity the plot displays the 1.1% of SNPs with $F_{ST} > 0.2$. The inset shows a histogram of the distribution of F_{ST} values from all genome-wide SNPs, with those showing values above 0.7 in red. The plots in (**B**) represent a zoom into each peak shown in (**A**) and display the genes that are annotated in that region. The insets show PCAs derived from the variants found in each peak. Samples are color-coded by species and the geographic region to which they belong (as in Figure 1).

Figure 4: ARG-based statistics. (**A**) Graphical representation of how species enrichments, RTH' and cross-coalescence values are calculated (modified from [17]). The species enrichment score is calculated for each subtree (example shaded in gray) in a topology and represents the probability of observing that number of samples of a particular species (the one represented in red in the example) in that subtree under a hypergeometric distribution. The score for a given tree and species is the maximum score associated with the full tree. RTH' is the ratio between the TMRCA of half of the samples from a species and the age of the youngest subtree that contains at least half of all samples (irrespective of species). The cross-coalescence time is calculated by adding the time of the ten most recent cross-coalescence events in the tree and dividing it by the same value used to normalize TMRCA to obtain RTH'. In the example the color of the leaves indicates two hypothetical species. (**B**) Plots showing these statistics for contig 404 in 5 kb windows, with the peak region shown by dashed vertical blue lines. These particular plots show statistics derived from individuals sampled outside of the hybrid zone. Species enrichment

521 values for *S. beltoni* and *S. plumbea* are higher than the genome-wide threshold of statistical significance,
522 while RTH' values are significantly lower. Therefore, subtrees from topologies in this region contain most
523 individuals from each species and are comparatively younger than in the rest of the genome. The cross-
524 coalescence times in the peak region are also significantly delayed, meaning that there is less recent
525 gene flow than across other areas of the genome.

526 **Tables**

527

528 **Table 1. Regions of high divergences between species.** Details for each outlier peak, including the genes they contain and a summary of their

529 functions.

530

531

Contig	Peak size (kb)	Highest SNP F_{ST} in window	Number of SNPs with F_{ST} > 0.7	Number of genes	Genes	Gene function
404	102	0.794	24	2	*PRLR*, *AGXT2*	*PRLR* encodes for the prolactin hormone receptor affecting breeding behavior and parental care in birds. *PRLR* also plays a role in keratinization. *AGXT2* encodes for a protein which catalyzes the conversion of glyoxylate to glycine.
33	66	0.901	9	1	*EDN3*	Encodes a protein affecting the development of melanocytes, resulting in both hypo and hyperpigmentation in chickens and is also related to egg production.

382	2	0.795	6	3	*DNAAF1*, *HSDL1*, *MBTPS1*	*DNAAF1* encodes a protein required for cilia stability. *HSDL1* enables oxidoreductase activity. *MBTPS1* encodes a member of the subtilisin-like proprotein convertase family.

532

Table 2: Results from statistical tests conducted on ARG-based statistics. Tests were conducted separately for allopatric, sympatric, or all samples combined. Species enrichment and RTH' scores were compared to the distribution of these values across a set of control contigs without outlier peaks. Cross-coalescence (CC) values were compared between the peaks and their adjacent regions, and for control regions, between focal and adjacent areas. The distribution of the difference in CC values between focal and adjacent areas was assessed across all control regions and used to determine the statistical significance of the difference in CC values observed between the peak regions and those areas immediately adjacent to the peaks. Statistically significant results are shown by asterisks, which also denote the level of significance (*0.005; **0.001). Note that we ran species enrichment and RTH' tests for each species, while the cross-coalescence test is run for the species pair and therefore only presented once in the top of the table.

		peak on contig 404			peak on contig 33			peak on contig 382		
		Species enrichment	RTH'	CC	Species enrichment	RTH'	CC	Species enrichment	RTH'	CC
S. beltoni	all samples	8.95**	0.246**	0.15	7.4**	0.367*	-0.05	4.15**	0.845	0.3*
	allopatric	9.16**	0.239*	0.32*	6.39**	0.346*	0.17	4.39**	0.772	0.5**
	sympatric	1.83	0.545	0.07	2.44**	0.49*	0.06	1.25	0.72	0.06
S. plumbea	all samples	9.28**	0.235**		4.76**	0.709		5.66**	0.753	
	allopatric	10.34**	0.193*		4.65**	0.742		6.24**	0.469*	
	sympatric	1.76	0.44*		1.7	0.693		1.33	0.828	

542 **References**

543

544 1. Schluter D. The ecology of adaptive radiation. Oxford University Press; 2000.

545 2. Price T. Speciation in birds. Roberts and Co.; 2008.

546 3. Coyne JA, Orr A. Speciation. Oxford University Press; 2004.

547 4. Grant PR, Grant BR. How and why species multiply: the radiation of Darwin's finches. Princeton
548 University Press; 2007.

549 5. Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, et al. Genomics and
550 the origin of species. Nature Reviews Genetics. 2014;15: 176–192. doi:10.1038/nrg3644

551 6. Barton NH, Gale KS, Harrison R. Genetic analysis of hybrid zones. Hybrid zones and the
552 evolutionary process. Oxford University Press; 1993: 13–45.

553 7. Klopfstein S, Currat M, Excoffier L. The fate of mutations surfing on the wave of a range expansion.
554 Molecular Biology and Evolution. 2006;23: 482–490. doi:10.1093/molbev/msj057

555 8. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due to
556 reduced diversity, not reduced gene flow. Molecular ecology. 2014;23: 3133–3157.

557 9. Hejase HA, Salman-Minkov A, Campagna L, Hubisz MJ, Lovette IJ, Gronau I, et al. Genomic islands
558 of differentiation in a rapid avian radiation have been driven by recent selective sweeps.
559 Proceedings of the National Academy of Sciences of the United States of America. 2020;48: 30554-
560 30565 doi:10.1101/2020.03.07.977694

561 10. Toews DPL, Taylor SA, Vallender R, Brelsford A, Butcher BG, Messer PW, et al. Plumage genes and
562 little else distinguish the genomes of hybridizing warblers. Current Biology. 2016;26: 2313–2318.
563 doi:10.1016/j.cub.2016.06.034

564 11. Irwin DE, Alcaide M, Delmore KE, Irwin JH, Owens GL. Recurrent selection explains parallel evolution
565 of genomic regions of high relative but low absolute differentiation in a ring species. Molecular
566 Ecology. 2016;25(18): 4488-4507.
567
568 12. Lewanski AL, Grundler MC, Bradburd GS. The era of the ARG: an empiricist's guide to ancestral
569 recombination graphs. ArXiv [Preprint]. 2023 Oct 18:arXiv:2310.12070v1. PMID: 37904740; PMCID:
570 PMC10614969.

571 13. Poelstra JW, Vijay N, Bossu CM, Lantz H, Ryll B, Muller I, et al. The genomic landscape underlying
572 phenotypic integrity in the face of gene flow in crows. Science. 2014;344: 1410–1414.
573 doi:10.1126/science.1253226

574 14. Campagna L, Repenning M, Silveira LF, Fontana CS, Tubaro PL, Lovette IJ. Repeated divergent
575 selection on pigmentation genes in a rapid finch radiation. Science Advances. 2017;3: e1602404.
576 doi:10.1126/sciadv.1602404

577 15. Stryjewski KF, Sorenson MD. Mosaic genome evolution in a recent and rapid avian radiation. Nature
578 Ecology and Evolution. 2017;1: 1912–1922. doi:10.1038/s41559-017-0364-7

16. Wang S, Rohwer S, Zwaan DR, Toews DPL, Lovette IJ, Mackenzie J, et al. Selection on a small genomic region underpins differentiation in multiple color traits between two warbler species. Evolution Letters. 2020;4: 502–515. doi:10.1002/evl3.198

17. Campagna L, Mo Z, Siepel A, Uy JAC. Selective sweeps on different pigmentation genes mediate convergent evolution of island melanism in two incipient bird species. Bierne N, editor. PLoS Genetics 2022;18: e1010474. doi:10.1371/journal.pgen.1010474

18. Uy JAC, Irwin DE, Webster MS. Behavioral isolation and incipient speciation in birds. Annual Review of Ecology, Evolution, and Systematics. 2018;49: 1–24. doi:10.1146/annurev-ecolsys-110617-062646

19. Campagna L, Toews DPL. The genomics of adaptation in birds. Current Biology. 2022;32: R1173–R1186. doi:10.1016/j.cub.2022.07.076

20. Turbek SP, Browne M, Di Giacomo AS, Kopuchian C, Hochachka WM, Estalles C, et al. Rapid speciation via the evolution of pre-mating isolation in the Iberá Seedeater. Science. 2021;371: eabc0256. doi:10.1126/science.abc0256

21. Marques DA, Meier JI, Seehausen O. A combinatorial view on speciation and adaptive radiation. Trends in Ecology & Evolution. 2019; S0169534719300552. doi:10.1016/j.tree.2019.02.008

22. Rosenblum EB, Sarver BAJ, Brown JW, Des Roches S, Hardwick KM, Hether TD, et al. Goldilocks meets Santa Rosalia: An ephemeral speciation model explains patterns of diversification across time scales. Evolutionary Biology 2012;39: 255–261. doi:10.1007/s11692-012-9171-x

23. Kautt AF, Kratochwil CF, Nater A, Machado-Schiaffino G, Olave M, Henning F, et al. Contrasting signatures of genomic divergence during sympatric speciation. Nature. 2020;588: 106–111. doi:10.1038/s41586-020-2845-0

24. Winkler DW, Billerman SM, and Lovette IJ. Tanagers and Allies (Thraupidae), version 1.0. In Birds of the World (SM Billerman, BK Keeney, PG Rodewald, and TS Schulenberg, Editors). Cornell Lab of Ornithology, Ithaca, NY, USA. 2020. https://doi.org/10.2173/bow.thraup2.01.

25. Burns KJ, Shultz AJ, Title PO, Mason NA, Barker FK, Klicka J, et al. Phylogenetics and diversification of tanagers (Passeriformes: Thraupidae), the largest radiation of Neotropical songbirds. Molecular Phylogenetics and Evolution. 2014;75: 41–77. doi:10.1016/j.ympev.2014.02.006

26. Mason NA, Olvera-Vital A, Lovette IJ, Navarro-Sigüenza AG. Hidden endemism, deep polyphyly, and repeated dispersal across the Isthmus of Tehuantepec: Diversification of the White-collared seedeater complex (Thraupidae: *Sporophila torqueola*). Ecology and Evolution 2018;8: 1867–1881. doi:10.1002/ece3.3799

27. Ocampo D, Winker K, Miller MJ, Sandoval L, Uy JAC. Rapid diversification of the Variable Seedeater superspecies complex despite widespread gene flow. Molecular Phylogenetics and Evolution. 2022;173: 107510. doi:10.1016/j.ympev.2022.107510

28. Campagna L, Benites P, Lougheed SC, Lijtmaer DA, Di Giacomo AS, Eaton MD, et al. Rapid phenotypic evolution during incipient speciation in a continental avian radiation. Proceedings of the Royal Society of London. Series B: Biological Sciences. 2012;279: 1847–1856. doi:10.1098/rspb.2011.2170

29. Repenning M, Fontana CS. Distinguishing females of capuchino seedeaters: call repertoires provide evidence for species-level diagnosis. Revista Brasileira de Ornitologia. 2019;27: 70–78. doi:10.1007/BF03544451

30. Estalles C, Turbek SP, José Rodríguez-Cajarville M, Silveira LF, Wakamatsu K, Ito S, et al. Concerted variation in melanogenesis genes underlies emergent patterning of plumage in capuchino seedeaters. Proceedings of the Royal Society of London. Series B: Biological Sciences. 2022;289: 20212277. doi:10.1098/rspb.2021.2277

31. Stiles FG. When black plus white equals gray: The nature of variation in the variable seedeater complex (Emberizinae: *Sporophila*). Ornitologia Neotropical. 1996;7: 75–107.

32. Ocampo D, Winker K, Miller MJ, Sandoval L, Uy JAC. Replicate contact zones suggest a limited role of plumage in reproductive isolation among subspecies of the variable seedeater (*Sporophila corvina*). Molecular Ecology. 2023; mec.16938. doi:10.1111/mec.16938

33. Hellmayr CE. Catalogue of birds of the Americas and the adjacent islands in Field Museum of Natural History. Chicago; 1918. doi:10.5962/bhl.title.5570

34. Repenning M, Fontana CS. A new species of gray seedeater (Emberizidae: *Sporophila*) from upland grasslands of southern Brazil. The Auk. 2013;130: 791–803. doi:10.1525/auk.2013.12167

35. Repenning M. Variacao geografica em *Sporophila* (Aves: Thraupidae) com base em evidencias fenotipicas, ecologicas e geneticas. PhD Thesis, Pontifícia Universidade Católica do Rio Grande do Sul. 2017. Available: http://tede2.pucrs.br/tede2/handle/tede/7598

36. Thawornwattana Y, Huang J, Flouri T š, Mallet J, Yang Z. Inferring the direction of introgression using genomic sequence data. Molecular Biology and Evolution. 2023; msad178. doi:10.1093/molbev/msad178

37. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. Genome-Wide inference of ancestral recombination graphs. PLoS Genetics. 2014;10: e1004342. doi:10.1371/journal.pgen.1004342

38. Lijtmaer DA, Sharpe NM, Tubaro PL, Lougheed SC. Molecular phylogenetics and diversification of the genus *Sporophila* (Aves: Passeriformes). Molecular Phylogenetics and Evolution. 2004;33(3): 562-579. doi: 10.1016/j.ympev.2004.07.011

39. Campagna L, Silveira LF, Tubaro PL, Lougheed SC. Identifying the sister species to the rapid capuchino seedeater radiation (Passeriformes: *Sporophila*). The Auk. 2013;130: 645–655. doi:10.1525/auk.2013.13064

40. Wolf JBW, Ellegren H. Making sense of genomic islands of differentiation in light of speciation. Nature Reviews Genetics. 2017;18: 87–100. doi:10.1038/nrg.2016.133

41. Akiyama T, Kinoshita K. Body color expression in birds. Pigments, Pigment Cells and Pigment Patterns. 2021; 91–126.

42. Maclary ET, Wauer R, Phillips B, Brown A, Boer EF, Samani AM, et al. An allelic series at the *EDNRB2* locus controls diverse piebalding patterns in the domestic pigeon. PLoS genetics. 2023; 19(10), p.e1010880. https://doi.org/10.1371/journal.pgen.1010880

43. Dong X, Li J, Zhang Y, Han D, Hua G, Wang J, et al. Genomic analysis reveals pleiotropic alleles at *EDN3* and *BMP7* involved in chicken comb color and egg production. Frontiers in Genetics. 2019;10: 612. doi:10.3389/fgene.2019.00612

44. Luo C, Shen X, Rao Y, Xu H, Tang J, Sun L, et al. Differences of Z chromosome and genomic expression between early- and late-feathering chickens. Molecular Biology Reports. 2012;39: 6283–6288. doi:10.1007/s11033-012-1449-7

661 45. Kuenzel W. Neurobiology of molt in avian species. Poultry Science. 2003;82: 981–991.
662 doi:10.1093/ps/82.6.981

663 46. Skold H, Amundsen T, Svensson P, Mayer I, Bjelvenmark J, Forsgren E. Hormonal regulation of
664 female nuptial coloration in a fish. Hormones and Behavior. 2008;54: 549–556.
665 doi:10.1016/j.yhbeh.2008.05.018

666 47. Hooper DM, Griffith SC, Price TD. Sex chromosome inversions enforce reproductive isolation across
667 an avian hybrid zone. Molecular ecology. 2019; 28(6): 1246-1262.
668 https://doi.org/10.1111/mec.14874

669 48. McDiarmid CS, Finch F, Peso M, van Rooij E, Hooper DM, Rowe M, Griffith SC. Experimentally
670 testing mate preference in an avian system with unidirectional bill color introgression. Ecology and
671 Evolution. 2023;13(2): e9812. https://doi.org/10.1002/ece3.9812

672 49. Mason NA, Burns KJ. Molecular phylogenetics of the Neotropical seedeaters and seed-finches
673 (*Sporophila*, *Oryzoborus*, *Dolospingus*). Ornitologia Neotropical. 2013;24: 139–155.

674 50. Baiz MD, Wood AW, Brelsford A, Lovette IJ, Toews DPL. Pigmentation genes show evidence of
675 repeated divergence and multiple bouts of introgression in *Setophaga* warblers. Current Biology.
676 2021;31: 643-649.e3. doi:10.1016/j.cub.2020.10.094

677 51. Grabenstein KC, Taylor SA. Breaking barriers: Causes, consequences, and experimental utility of
678 human-mediated hybridization. Trends in Ecology & Evolution. 2018;33: 198–212.
679 doi:10.1016/j.tree.2017.12.008

680 52. Vijay N, Bossu CM, Poelstra JW, Weissensteiner MH, Suh A, Kryukov AP, et al. Evolution of
681 heterogeneous genome differentiation across multiple contact zones in a crow species complex.
682 Nature Communications. 2016;7: 13195. doi:10.1038/ncomms13195

683 53. Semenov GA, Linck E, Enbody ED, Harris RB, Khaydarov DR, Alström P, et al. Asymmetric
684 introgression reveals the genetic architecture of a plumage trait. Nature Communications. 2021;12:
685 1019. doi:10.1038/s41467-021-21340-y

686 54. Walsh J, Kovach AI, Olsen BJ, Shriver WG, Lovette IJ. Bidirectional adaptive introgression between
687 two ecologically divergent sparrow species. Evolution. 2018;72: 2076–2089. doi:10.1111/evo.13581

688 55. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and
689 read merging. BMC Research Notes. 2016; 9: 88. https://doi.org/10.1186/s13104-016-1900-2

690 56. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9:
691 357–359. doi:10.1038/nmeth.1923

692 57. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, et al. Qualimap:
693 evaluating next-generation sequencing alignment data. Bioinformatics. 2012;28: 2678–2679.
694 doi:10.1093/bioinformatics/bts503

695 58. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format
696 and VCFtools. Bioinformatics. 2011;27: 2156–2158. doi:10.1093/bioinformatics/btr330

697 59. Van der Auwera GA, O'Connor BD. Genomics in the cloud: Using docker, GATK, and WDL in Terra.
698 O'Reilly Media, Incorporated; 2020.

699 60. Hebert PDN, Ratnasingham S, de Waard JR. Barcoding animal life: cytochrome *c* oxidase subunit 1
700 divergences among closely related species. Proceedings of the Royal Society of London. Series B:
701 Biological Sciences. 2003;270. doi:10.1098/rsbl.2003.0025

702 61. Kerr KCR, Lijtmaer DA, Barreira AS, Hebert PDN, Tubaro PL. Probing evolutionary patterns in
703 Neotropical birds through DNA barcodes. DeSalle R, editor. PLoS ONE. 2009;4: e4379.
704 doi:10.1371/journal.pone.0004379

705 62. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: An
706 integrated and extendable desktop software platform for the organization and analysis of sequence
707 data. Bioinformatics. 2012;28: 1647–1649. doi:10.1093/bioinformatics/bts199

708 63. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R
709 Foundation for Statistical Computing; 2021. Available: https://www.R-project.org/

710 64. Paradis E. pegas: an R package for population genetics with an integrated–modular approach.
711 Bioinformatics. 2010;26: 419–420. doi:10.1093/bioinformatics/btp696

712 65. Zheng X, Levine D, Shen J, Gogarten S, Laurie C, Weir B. AhHigh-performance computing toolset for
713 relatedness and principal component analysis of SNP data. Bioinformatics. 2012;28: 3326–3328.
714 doi:10.1093/bioinformatics/bts606

715 66. Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots.
716 Journal of Open Source Software. 2018;3: 731–732.

717 67. Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent selective sweeps in North American
718 *Drosophila melanogaster* show signatures of soft sweeps. PLoS Genetics. 2015;11: e1005004.

719 68. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for
720 whole-genome association studies by use of localized haplotype clustering. The American Journal
721 of Human Genetics. 2007/09/21 ed. 2007;81: 1084–1097. doi:10.1086/521987

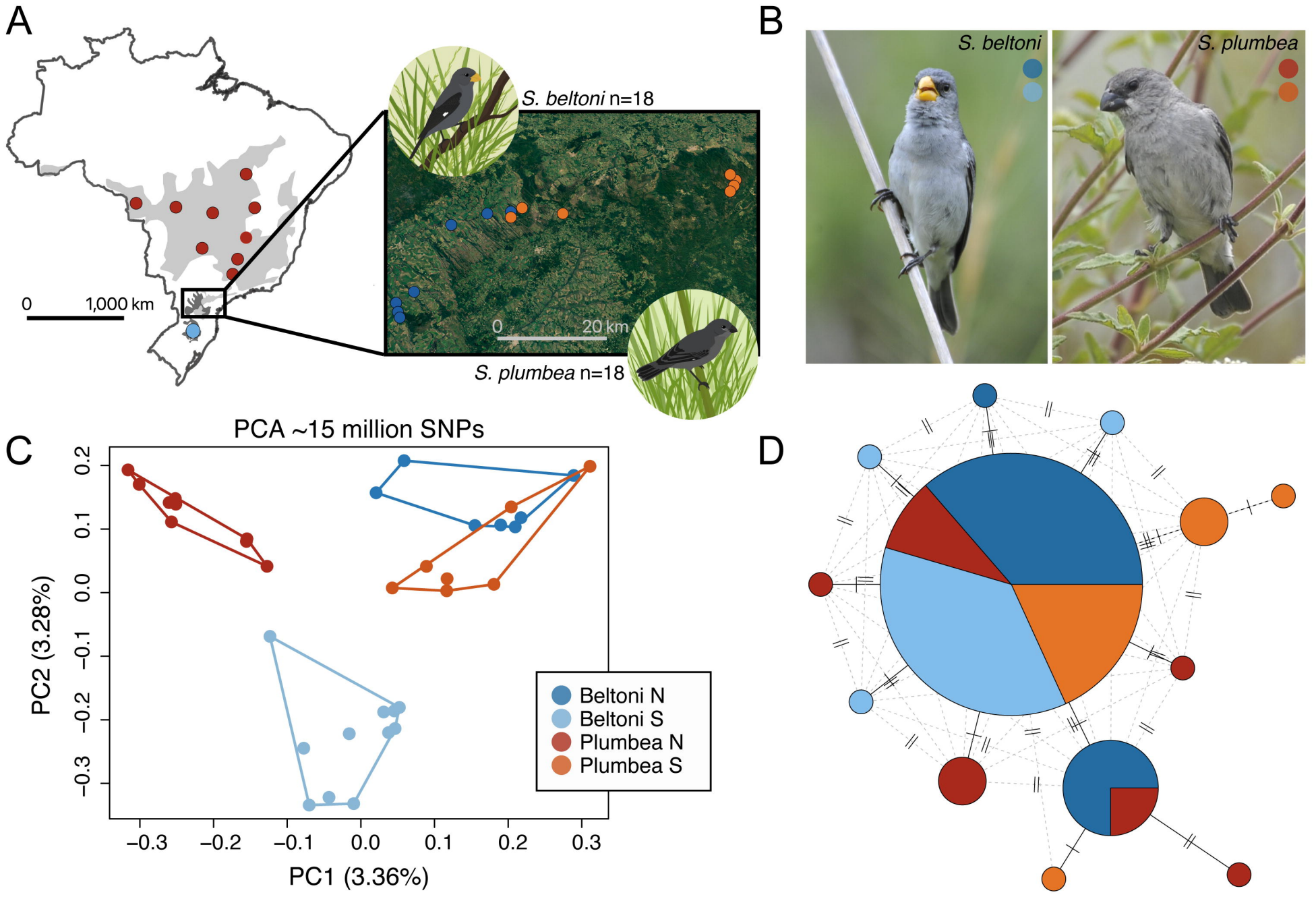
722 69. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things): phytools:
723 R package. Methods in Ecology and Evolution. 2012;3: 217–223. doi:10.1111/j.2041-
724 210X.2011.00169.x

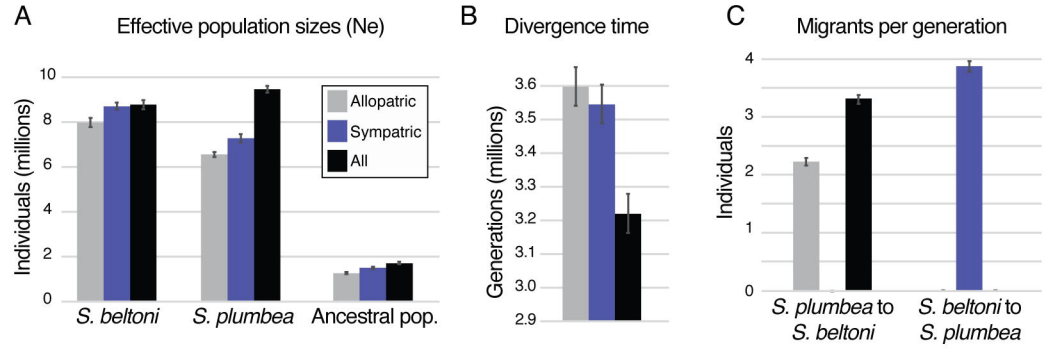
725 70. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community
726 ecology package. R package version 2.5-6. https://CRAN.R-project.org/package=vegan; 2019.
727 Available: https://CRAN.R-project.org/package=vegan
728
729 71. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated
730 individuals. Genome Research. 2009;19: 1655–1664.
731
732 72. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016. Available:
733 https://ggplot2.tidyverse.org
734
735 73. Korunes KL, Samuk K. pixy: Unbiased estimation of nucleotide diversity and divergence in the
736 presence of missing data. Molecular Ecology Resources. 2021;21: 1359–1368.

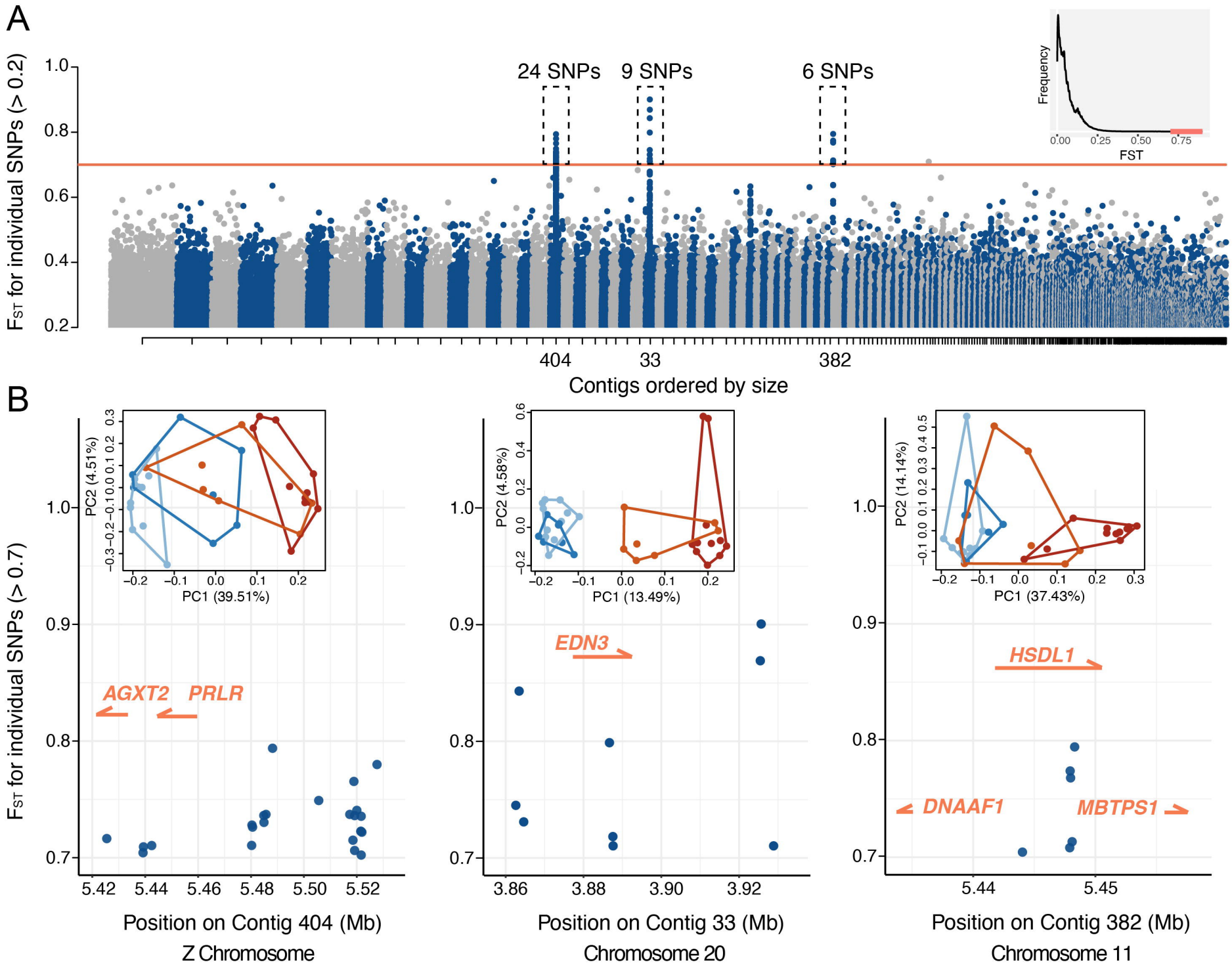
737 74. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. Bayesian inference of ancient human
738 demography from individual genome sequences. Nature Genetics. 2011;43: 1031–1034.
739 doi:10.1038/ng.937

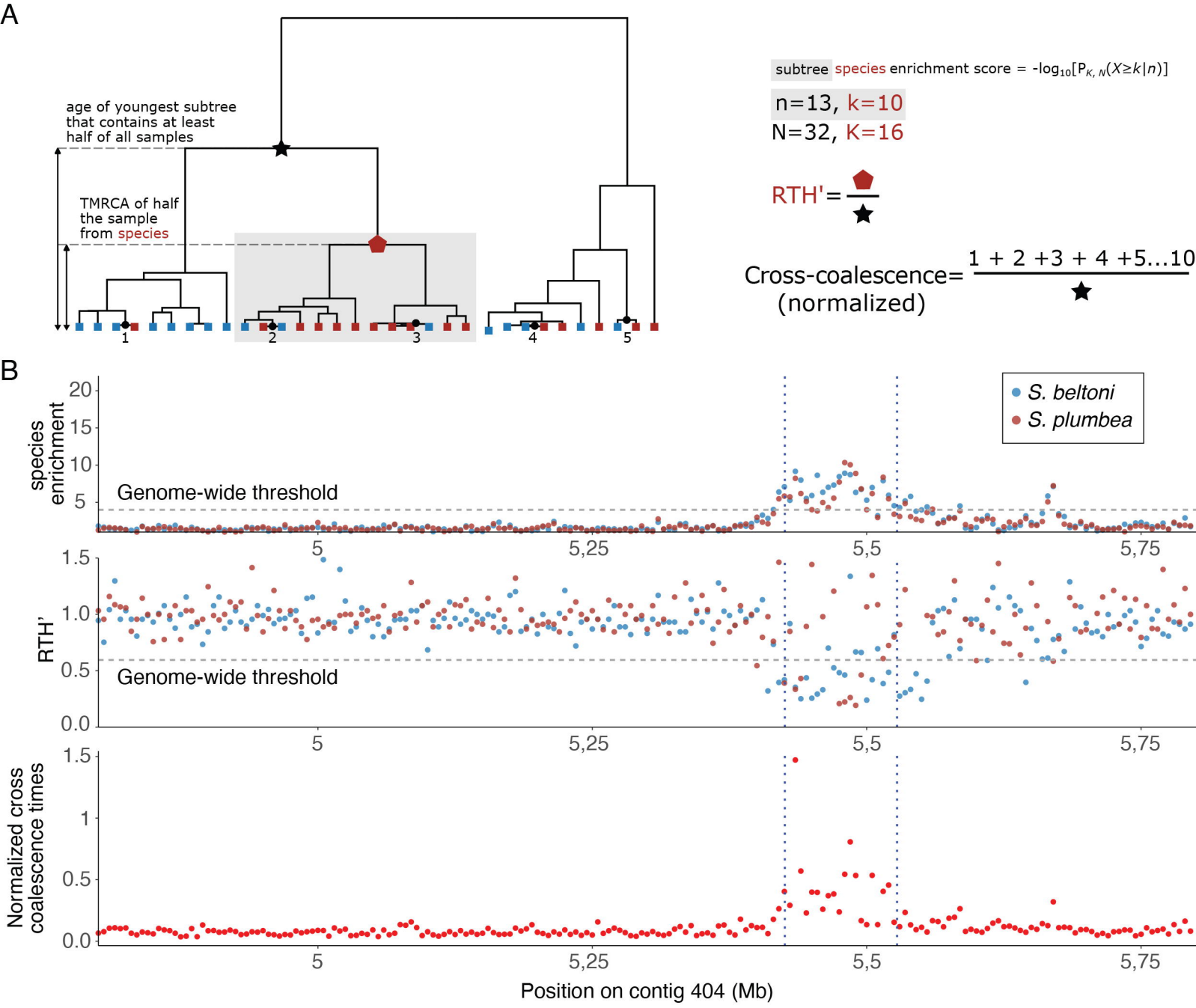
740 75. Plummer M, Best N, Cowles K, Vines K. CODA: Convergence diagnosis and output analysis for
741 MCMC. R News. 2006;6: 7–11.

742 76. Smeds L, Qvarnström A, Ellegren H. Direct estimate of the rate of germline mutation in a bird.
743 Genome Research. 2016;26: 1211–1218. doi:10.1101/gr.204669.116


A
S. beltoni n=18
S. plumbea n=18
0
1,000 km
0
20 km
B
S. beltoni
S. plumbea
C
PCA ~15 million SNPs
PC2 (3.28%)
PC1 (3.36%)
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
Beltoni N
Beltoni S
Plumbea N
Plumbea S
D


A
Effective population sizes (Ne)
Individuals (millions)
10
8
6
4
2
0
Allopatric
Sympatric
All
S. beltoni
S. plumbea
Ancestral pop.
B
Divergence time
Generations (millions)
3.6
3.5
3.4
3.3
3.2
3.1
3.0
2.9
C
Migrants per generation
Individuals
4
3
2
1
0
S. plumbea to S. beltoni
S. beltoni to S. plumbea


A
24 SNPs
9 SNPs
6 SNPs
Frequency
FST
FST for individual SNPs (> 0.2)
404
33
382
Contigs ordered by size
B
PC2 (4.51%)
PC1 (39.51%)
PC2 (4.58%)
PC1 (13.49%)
PC2 (14.14%)
PC1 (37.43%)
FST for individual SNPs (> 0.7)
AGXT2
PRLR
EDN3
HSDL1
DNAAF1
MBTPS1
Position on Contig 404 (Mb)
Z Chromosome
Position on Contig 33 (Mb)
Chromosome 20
Position on Contig 382 (Mb)
Chromosome 11


A
age of youngest subtree that contains at least half of all samples
TMRCA of half the sample from species
1
2
3
4
5
subtree species enrichment score = -log10[P_{K,N}(X≥k|n)]
n=13, k=10
N=32, K=16
RTH'=
Cross-coalescence= (normalized)
1 + 2 +3 + 4 +5...10
B
species enrichment
Genome-wide threshold
S. beltoni
S. plumbea
RTH'
Genome-wide threshold
Normalized cross coalescence times
Position on contig 404 (Mb)