Hybridization and extinction in a recent Passer sparrow zone

Vitalii Lichman

Master of Science Thesis

Department of Biosciences

Faculty of Mathematics and Natural Sciences

University of Oslo

01.11.2018

2018

Hybridization and extinction in a recent Passer sparrow zone

Vitalii Lichman

http://www.duo.uio.no/

Trykk: Reprosentralen, Universitetet i Oslo

Abstract

The avifauna of Cape Verde archipelago is represented by three species within the *Passer* genus. Due to its distant localization from the continent and its variety of landscapes, this group of islands serves as objects of interest for studies in the sphere of evolutionary biology. From the beginning of the age of naturalistic explorations in the middle of 19th century, only few detailed ornithological expeditions were conducted until recently. In this connection, nowadays we have at our disposal only superficial information concerning the disposition of population structure and interspecific interactions within bird species, particularly sparrows. Technical progress and development of technologies in the field of molecular biology giving us an opportunity to investigate these processes more closely. This study clarifies phylogenetic relationships between 3 Passer species: 2 invasive (*P. domesticus* and *P. hispaniolensis*) and 1 endemic (*P. Iagoensis*). I also revealed a pronounced presence of *P. hispaniolensis* ancestry in *P. domesticus* genome that indicates existence of recent hybridization in the range of their contact, and supporting the notion that these species are prone to interspecific breeding elsewhere. I also found that *P. Iagoensis* has relatively high genome divergence - wide fixation index, that suggests absence of interbreeding between endemic and any of the invasive species. This is the first study of sparrows on the Cape Verde based on genetics and bioinformatics that presents explicit results on population structure

Acknowledgements

First and foremost, I would like to thank my supervisors Glenn-Peter Sætre and Mark Ravinet for all of their guidance. Melissah Rowe along with PhD students Angelica Cuevas and Camilla Lo Cascio Sætre provided training with capturing birds and familiarized with technique of taking blood from avian species. Special thanks to Martin Melo and Juan Carlos Illeria for collecting blood samples from Portugal. Thank you again to Mark, Glenn and Camilla for collection and providing blood samples. Thank you to Fabrice Eroukhmanoff and Mark for arranging of field work and Åsta Dale for assistance during it. Thank you to Mariia Kuzmenko for inspiration and motivation during project. Thank you all who was giving help at all stages of project implementation and to all members of CEES for wonderful experience and opportunity to become a part of the scientific community.

Contents

Appendix 3: Filtering SNPs

Appendix 4: Conversion from BAM to FASTQ

Appendix 5: Conversion from FASTQ to psmcfa

Appendix 6: Conversion from psmcfa to psmc

Appendix 7: Plink linkage pruning

Appendix 8: Plink PCA

Appendix 9: Plotting PCA

Appendix 10: Substitution of chromosome names

Appendix 11: Running admixture

Appendix 12: Plotting admixture

Appendix 13: Estimation of pairwise Fst

Appendix 14: Producing data.frame

Appendix 15: Plotting Fst

Introduction

Hybridization is the phenomenon determined as “crossing of genetically distinguishable groups or taxa, leading to the production of viable hybrids” (Mallet 2005, Futuyma 2013). Even though the existence of hybridization has been observed since the time of Linnaeus (Mallet 2005), evolutionary biologists have been debating for a long time in an attempt to determine precisely its role in speciation. More explicit denotation of this biological event derives from studies in selection, genetics and notably the concept of gene flow. Basically, gene flow is characterized as mating among individuals from different populations (Futuyma 2013) and successful transfer of alleles from one population to another (Ellstrand, 2016).

In respect of historic background, the most dominant view on speciation throughout last century was a biological species concept, and hybridization was one of the keystones . Followers of this idea viewed species as reproductively isolated populations, and therefore considered hybridization as “breakdown of isolating mechanisms”. According to their arguments, due to low fertility, caused by structural differences between the chromosomes and inferior viability of hybrid offspring by reason of disharmonious interactions between the different genes, any assumed backcrossed genotypes are of low quality, and thus result in deleterious effects (Mallet 2005, Futuyma 2013).

Although recent studies sustain the idea of multiple and significant impacts of hybridization, modern perspectives are less conservative regarding its negative influence on the speciation process. On the one hand, present comprehension of evolutionary studies does not reject the evidence that gene flow may inhibit adaptation and divergence. On the other hand, referring to conventional definition of hybridization, current understanding implies its occurrence in all known processes of speciation except instantaneous speciation and complete allopatry (Abbott et al. 2013).

Contemporary researches reveal and ascertained the presence of hybrid speciation, that occurs through formation of new hybrid taxa and thereafter promotion of adaptive divergence due to adaptive introgression - invasion of foreign genetic material into a genome across a partial barrier (Abbott et al. 2013, Mallet 2005). In support of that given

mechanism is regular and ongoing, it is estimated that 10% of animal and 25% of plant species are hybridized with at least one species (Mallet 2007).

Even though introgression was thoroughly observed and described in plants since 1940-s (Anderson 1949), its patterns in animals have not been well-documented until recently. Modern studies indicate the occurrence of introgression between *Homo sapiens* and archaic humans, such as Neanderthal (Currat 2011) and Denisovan (Dannemann 2016); significant effect was also observed in divergent forms of stickleback (Yamada 2001, Ravinet et al. 2018), cyprinid (Turner et al. 2004, Aboim et al. 2010) and cichlid fishes (Salzburger et al. 2002, Meier et al. 2017). Studies show that introgressive hybridization can be more extensive than previously thought and is likely to be a source for rapid adaptive radiation. Recent explorations acted as breakthrough in understanding of the processes of speciation as a whole, but continuous debates whether hybridization is evolutionary noise or engine of biodiversity (Soltis 2013) are still ongoing.

Modern studies revealed also a strong signal for hybridization (Ait Belkacem et al. 2016, Hermansen et al. 2011) in sparrow species. For instance, Italian sparrow is known as a hybrid taxon, that was produced in a secondary contact zone between Spanish (*P. hispaniolensis*) and house (*P. domesticus)* sparrows and subsequently backcrossed to its ancestors (Trier et al. 2014 , Hermansen et al. 2014, Elgvin). Genus *Passer* is about 30 species, which are widespread around Eurasia and Africa. Due to both deliberate and accidental introductions by human, house sparrow, strongly associated as human-commensal species, enhanced the ranges of distributions and nowadays can be found on all continents except Antarctica (Ravinet et al. 2018). Since *P. domesticus* is widespread, it may come in to contact with other *Passer* species and consequently increases the opportunity for likely introgression.

The Cape Verde archipelago is located approximately 570 km away from Africa's west coast and consists of 10 islands (*Santo Antão, Santiago, São Nicolau, São Vicente, Fogo, Sal, Maio, Brava, Santa Luiza, Boa Vista*) and 4 islets (Pinheiro et al. 2013), see figure 1. The *Passer* genus is represented on the archipelago by 3 species – *P. hispaniolensis, P. domesticus,* and *P. iagoensis,* see figure 2*. P. iagoensis* is an endemic species whereas the two others are invasive and were introduced in the early 19th century and between 1922-1924 years respectively (Summers-Smith, 1988). The first inspection of bird fauna was conducted by Darwin in 1832, and even though ornithological observation has been managed and

documented until present, only one published study was focused on *Passer* species (Summers-Smith, 1988).

The first aim of the project was to define phylogenetic relationships among populations of the three *Passer* species on the Cape Verde islands. The second purpose was to determine when the two non-endemic species reached the archipelago. The last objective was to clarify whether there is genomic evidence of introgression between the Spanish sparrow (*P. hispaniolensis)* and house sparrow (*P. domesticus)* populations since they came in to contact on the archipelago. To reach first goal, I examined relatedness between *P. iagoensis* and 14 sparrow species, based on SNPs in whole mitochondrial genome and constructed a phylogenetic tree. To disclose second objective, I reconstructed the fluctuations in effective population size over time. To achieve last goal, I used statistical methods to estimate proportions of admixture and reveal signals for gene flow between populations.

Iago sparrow appears to be a reliable model organism for the study of speciation in the wild: it is abundant on all the islands and occurs in all habitats; moreover *P. iagoensis* is being considered as a recent species, that presents extensive phenotypic variation that has not been properly studied yet (Pinheiro de Melo et al. 2013). Additionally, due to specifics of aridity gradient and landscape features on archipelago, along with wide range of distances among groups of islands and islets, these species might serve as the trustworthy object to study divergence with different levels of gene flow and the effect of isolation by distance in the process of diversification (Pinheiro de Melo et al. 2013). Finally, it was found that *P.iagoensis* is co-occuring with *P.hispaniolensis* on five islands.

Single nucleotide polymorphism (SNP) is the form of variation in DNA that occurs due to substitution of one single nucleotide to another (Shastry 2009). SNPs also considered to be as the basic form of genetic variation and thus serve as the principal indicator in processes of molecular evolution (Oeveren 2009). They are widely spread through genome, thus they can act as a reliable source for analyses in population genetics, such as historical demography and speciation (Brumfield 2003).

Figure 1. Map of Cape Verde archipelago

Figure 2. Species of Passer genus distributed on Cape Verde archipelago: (from left to right) P.domesticus, P. hispaniolensis, P. Iagoensis

2 Materials and methods

2.1 Samples collection

A total set of 80 blood samples was constructed in the following way: 5 blood samples of Iago sparrow (*P. iagoensis)*, 12 of Spanish sparrow (*P. hispaniolensis)* and 8 of house sparrow (*P.domesticus*) from Cape Verde were collected by collaborators from Portugal; 15 and 10 sequences of house sparrow (*P. domesticus)* from Northern Norway and Sales, France respectively in addition to 20 samples of *P. hispaniolensis* from Tenerife, Spain were obtained from studies by Ravinet et al. 2018. To increase consistency of data set we also added 10 samples of Spanish sparrow from Lesina, Italy and 10 blood samples of Italian sparrows (*P. italiae*) (Elgvin et al. 2017).

Capturing of birds was conducted with the usage of mist nets, after which birds were placed in a cloth bag. Prior to sampling, the inner part of the wing was cleaned with a tissue wetted in 50% of ethanol, in order to prevent material from bacterial contamination. Blood samples have been taken from brachial vein using medical needle and capillary and subsequently were either added to tubes with Queen's lysis buffer or 70% ethanol for preservation. Samples were immediately put in a fridge after return from the field and stored at a constant temperature of 4°C. Captured individuals were released rapidly to reduce stress and prevent harm. Fieldwork was done under correct permits and permissions from appropriate local authorities.

Table 1. List of individuals included in a data set

ID	Species	Location	Additional information
8934547	*P.domesticus*	Northern Norway	*Aldra*
8L19766	*P.domesticus*	Northern Norway	*Aldra*
8L19786	*P.domesticus*	Northern Norway	*Alta*
8L52141	*P.domesticus*	Northern Norway	*Aldra*
8L52830	*P.domesticus*	Northern Norway	*Aldra*
8L64869	*P.domesticus*	Northern Norway	*Leka*
8L89915	*P.domesticus*	Northern Norway	*Træna*
8M31651	*P.domesticus*	Northern Norway	*Leka*
8M71932	*P.domesticus*	Northern Norway	*Løkta*
8M72455	*P.domesticus*	Northern Norway	*Kvål*
8N05240	*P.domesticus*	Northern Norway	*Aldra*
8N05890	*P.domesticus*	Northern Norway	*Leka*
8N06612	*P.domesticus*	Northern Norway	*Linesøya*
8N73248	*P.domesticus*	Northern Norway	*Lauvøya*
8N73604	*P.domesticus*	Northern Norway	*Lauvøya*
FR041	*P.domesticus*	France	*Sales*
FR044	*P.domesticus*	France	*Sales*
FR046	*P.domesticus*	France	*Sales*
FR048	*P.domesticus*	France	*Sales*
FR049	*P.domesticus*	France	*Sales*
FR050	*P.domesticus*	France	*Sales*
CVH02	*P.domesticus*	Cape Verde	*São Vicente*
CVH03	*P.domesticus*	Cape Verde	*São Vicente*
CVH05	*P.domesticus*	Cape Verde	*São Vicente*
CVH07	*P.domesticus*	Cape Verde	*São Vicente*
CVH13	*P.domesticus*	Cape Verde	*São Vicente*
CVH15	*P.domesticus*	Cape Verde	*São Vicente*
CVH16	*P.domesticus*	Cape Verde	*São Vicente*
CVH17	*P.domesticus*	Cape Verde	*São Vicente*
Guglionesi336	*P.italiae*	Italy	*Guglionesi*
Guglionesi426	*P.italiae*	Italy	*Guglionesi*
Guglionesi427	*P.italiae*	Italy	*Guglionesi*
Guglionesi428	*P.italiae*	Italy	*Guglionesi*
Guglionesi429	*P.italiae*	Italy	*Guglionesi*
Guglionesi431	*P.italiae*	Italy	*Guglionesi*
Guglionesi432	*P.italiae*	Italy	*Guglionesi*

Guglionesi433	*P.italiae*	Italy	*Guglionesi*
Guglionesi434	*P.italiae*	Italy	*Guglionesi*
Guglionesi435	*P.italiae*	Italy	*Guglionesi*
Lesina_280	*P.hispaniolensis*	Italy	*Lesina*
Lesina_281	*P.hispaniolensis*	Italy	*Lesina*
Lesina_282	*P.hispaniolensis*	Italy	*Lesina*
Lesina_285	*P.hispaniolensis*	Italy	*Lesina*
Lesina_286	*P.hispaniolensis*	Italy	*Lesina*
Lesina_287	*P.hispaniolensis*	Italy	*Lesina*
Lesina_288	*P.hispaniolensis*	Italy	*Lesina*
Lesina_289	*P.hispaniolensis*	Italy	*Lesina*
Lesina_292	*P.hispaniolensis*	Italy	*Lesina*
Lesina_295	*P.hispaniolensis*	Italy	*Lesina*
CSP03_161116	*P.hispaniolensis*	Spain	*Tenerife*
CSP17_151116	*P.hispaniolensis*	Spain	*Tenerife*
CSP18_151116	*P.hispaniolensis*	Spain	*Tenerife*
CSP22_151116	*P.hispaniolensis*	Spain	*Tenerife*
CSP31_151116	*P.hispaniolensis*	Spain	*Tenerife*
CSP36_151116	*P.hispaniolensis*	Spain	*Tenerife*
CSP4_161116	*P.hispaniolensis*	Spain	*Tenerife*
CSP6_151116	*P.hispaniolensis*	Spain	*Tenerife*
CSP7_151116	*P.hispaniolensis*	Spain	*Tenerife*
CSP9_151116	*P.hispaniolensis*	Spain	*Tenerife*
CV-BOA16	*P.hispaniolensis*	Cape Verde	*Boa Vista*
CV-BOA19	*P.hispaniolensis*	Cape Verde	*Boa Vista*
CV-SNI03	*P.hispaniolensis*	Cape Verde	*São Nicolau*
CV-SNI04	*P.hispaniolensis*	Cape Verde	*São Nicolau*
CV-SNI09	*P.hispaniolensis*	Cape Verde	*São Nicolau*
CV-STG03	*P.hispaniolensis*	Cape Verde	*Santiago*
CV-STG05	*P.hispaniolensis*	Cape Verde	*Santiago*
CV-STG11	*P.hispaniolensis*	Cape Verde	*Santiago*
CV-STG16	*P.hispaniolensis*	Cape Verde	*Santiago*
CV-STG18	*P.hispaniolensis*	Cape Verde	*Santiago*
IAG01	*P.iagoensis*	Cape Verde	*Santiago*
IAG02	*P.iagoensis*	Cape Verde	*Santiago*
PI-SA-2	*P.iagoensis*	Cape Verde	*Santo Antão*
PI-SA-3	*P.iagoensis*	Cape Verde	*Santo Antão*
PI-SA-5	*P.iagoensis*	Cape Verde	*Santo Antão*
CSV01	*P.hispaniolensis*	Cape Verde	*Santiago*
CVS02	*P.hispaniolensis*	Cape Verde	*Santiago*

FR051	*P.domesticus*	France	*Sales*
FR061	*P.domesticus*	France	*Sales*
FR063	*P.domesticus*	France	*Sales*
FR064	*P.domesticus*	France	*Sales*

2.2 Laboratory protocol

2.2.1 DNA extraction

Taking into consideration the type of sampling material, we applied Qiagen DNEasy Spin-Column protocol for purification of total DNA from animal blood or cells (DNeasy blood & tissue handbook, 2006). First, we removed a small amount of ethanol from original coagulated samples and carefully transferred blood from test tube to microcentrifuge tube by micro spoon and then dissolved clots by adding 100µL Queen's lysis buffer (QLB). Initially anticoagulated samples, that were stored in QLB, 125 µL each, were added to the mixture consisting of µL proteinase K and 75 µL phosphate-buffered saline (PBS). After an overnight (~12 hours) incubation at 56°C, we added 4 µL of RNase A in each tube and repeated incubation at room temperature. Thereafter lysis buffer (AL) was added to samples, 200 µL each, were mixed by vortexing to yield a homogenous solution and incubated at 56°C for 10 minutes. Further we added 200µL of ethanol, transferred to mini spin columns and centrifuged at 8000 rpm for 1 minute. Then we added 500 µL of wash buffer (AW1), repeating centrifuge step with same parameters, whereupon duplicated washing step using buffer (AW2) and centrifuged at 14000rpm for 3 minutes. To increase the final DNA concentration in the eluate, samples were split and pipetted with 100µL elution buffer (EB), incubated for 1 min at room temperature and centrifuged at 8000 rpm for 1 min.
To confirm successful DNA isolation, we measured the amount of DNA yield using Nanodrop and Qubit© 2.0 Fluorometer applying Qubit© dsDNA Broad-Range Assay (Thermo Fischer Scientific Inc., Waltham, MA, USA). We used Illumina TruSeq gDNA 180kb kit (Illumina, CA, USA) to prepare a library of extracted genomic DNA (concentration range 0,3-1,6 µg/mL).

2.2.2 Sequencing

Sequencing was performed on apparatus Illumina Hi-Seq 2000 and Illumina Hi-Seq X at either Norwegian Sequencing Center (NSC) or Genome Quebec, McGill University, Canada.

2.2.3 Data processing

Since raw Next Generation Sequencing (NGS) data may contain adapter sequences and errors, that decrease its overall quality, we used Trimmomatic v.036 (Bolger et al. 2014) to trim and filter for Illumina adapters. We removed base calls with a Phred quality score less than 5 at the start and the end of reads, retaining only those with accuracy more than 70%, regarding the growth of erroneous with increasing read length. Additionally, we kept solely reads with Phred score more than 10, removing base calls with accuracy less than 90% throughout 5 base pairs (bp) step windows to assure that data is of high degree of authenticity.

Raw sequencing data is stored in a FASTQ format. As a rule, it consists of millions of sequences, whereas every sequence is represented by an individual identifier, the DNA base calls and the value of quality /confidence/ for each individual base call (Bonfield et al. 2013). Commonly FASTQ files are of large size, thus entire indexation (appendix) is required for rapid and effective search throughout whole reference genome.

Further alignment of filtered reads to the sparrow reference genome (Elgvin et al. 2017) was processed using Burrows-Wheeler alignment tool (bwa) v.0.7.10 (Li&Durbin 2009). We mapped paired and unpaired reads independently and generated sequence alignment map.

The output of aligned data is stored in Sequence Alignment/Map (SAM) format. It is called by developers as “generic alignment format for storing read alignments against reference sequences” (Li et al. 2009). Binary Alignment/Map (BAM) is a binary representation of SAM that holds all information from SAM-files and allows to compress the records. BAM is an essential format for experiment, since it is used in calling SNPs, and the size of data is significant. SAMs were converted to BAMs using script and subsequently sorted. Finally, we realigned BAMs around insertion-deletion polymorphisms to avoid incorrect positive variant detection.

DNA polymorphism data, obtained after performing previous steps, is kept in Variant Call Format (VCF). It has been commonly used as a standardized format (Danecek et al. 2011) for multiple purposes. We used samtools mpileup for generating VCF file. To perform this step, we used pipes, executing samtools v1.3.1 to pileup BAMs and bcftools v1.1 for calling variants respectively. We treated all reads as one sample in one BAM and set coefficient for downgrading mapping quality as 50.

Final VCF was filtered using VCFtools v0.1.13; all insertions and deletions were excluded, all genotypes with quality, e.g. Phred score below 20 were excluded (thus, ensured accuracy is 99%), allowed 20% of missing data; we remained sites with depth values in the range within 10 and 40, preserving sites with mean depth in an interval between 5 and 60.
For evaluation of effective population size in terms of demographic history, we applied a specialization of the sequentially Markovian coalescent model referred to as pairwise sequentially Markovian coalescent (PSMC) approach (Li&Durbin 2011). Operating principle that underlie PSMC is evaluation of constant time to most recent common ancestor (TMRCA) blocks, that are divided due to recombination, through consistency of heterozygote sites over a single diploid genome (Ravinet et al. 2018). Following Nadachowska – Brzyska (), we investigated a data set of 29 resequenced genomes:1 individual of *P. domesticus* from Oslo, Norway and 10 individuals from Cape Verde, *P. montanus* from Giardini-Naxos, Italy, 12 *P. hispaniolensis* from Tenerife, 2 *P. hispaniolensis* from Cape Verde and 5 *P. Iagoensis* from Cape Verde. We ran program independently for every genome by reason of PSMC is unable to process several diploid individuals simultaneously. At the initial stage we converted BAM files to FASTQ format. For this purpose, we used samtools mpileup along with bcftools with implementation of pipeline. Sites with Phred quality score over 20 (i.e. 99% base call accuracy) were preserved; additionally, a minimal number of 10000 sites were kept and 100bp windows were applied along a scaffold. Further we converted FASTQ to psmcfa, so the output file could be readable and executable by Pairwise SMC Model v.0.6.5-r67 program.

Then we ran PSMC implementing minimum number of iterations N=30 along with a maximum coalescent time t=15 and adjusted 45 separate time intervals, whereas 4 first and 3 last intervals were incorporated due to incorrect representation of recent and ancient time

periods by method (Ravinet et al. 2018). To keep consensus sequences with high confidence (Mays et al. 2018) several filters were used (see appendix). We applied mutation rate μ=1.4e-9 and set generation time of 1 year relying on prior studies (Nadachowska – Brzyska et al, 2013) Finally, time constraints were delimited in a range between 10 000 (10kya) and 10 000 000 (10Mya) years ago. Visualization of output data was performed in GNUPLOT v.5.0.

2.2.4 Specification of phylogenetic relationships

For establishing phylogenetic relationships between species we used Maximum likelihood estimation based on variations in whole mitochondrial DNA genome. Total set of 20 genomes, including the following: 4 individuals of Spanish (*P. hispaniolensis*), house (*P. domesticus*), Dead Sea (*P. moabiticus),* tree (*P. montanus*), socotra (*P. insularis*), saxaul (*P. ammodendri*), southern grey-headed (*P. diffusus*), Sind (*P. pyrrhonotus*), Italian (*P. italiae*), Sudan golden (*P. luteus*), northern grey-headed (*P. griseus*), great (*P. motitensis*), Swahili (*P. suahelicus*), 2 individuals of Iago (*P. Iagoensis*) and bactrianus (*P. domesticus bactrianus*) sparrows respectively was aligned using MUSCLE v.3.8.31.

Additionally, the genome of zebra finch, *Taeniopygia guttata* was added to set as outgroup. I used jModelTest v.2.1.10 to examine which of sequence evolution models is the most applicable, and subsequently ran RAxML v.8.0.26 to generate best phylogenetic tree. Visualization was done using Dendroscope v.3.5.9.

2.2.5 Analysis of population structure

Principal component analysis (PCA) is a widespread data processing and dimension reduction technique (Zou et al. 2004). The underlying principle of PCA is a mathematical algorithm that decreases the dimensionality of the data while keeping most of the variation in the data set. Reduction of dimensionality is obtained by identifying directions and maximal variation of data across them. Representation by few numbers of components instead of a large number of variables with an opportunity for further plotting and, thus visualization, whether samples

are homogenous or can be grouped, provides significant simplification of data analysis (Ringnér 2008).

To examine the structure of populations, we first used PLINK v1.9 for linkage disequilibrium (LD) pruning on the minor allele frequency (MAF) 5%, or 0.05 data set, based on the fact that PCA method refers to assumption of statistical independence of components (Ringnér 2008). We applied a window size in variant of 50kb, used variant count to shift the window at the end of each step of 10 and set pairwise r^2 threshold over 0,1, as this value is basic for genome-wide LD. Finally, we performed PCA on allele frequencies for whole genome, autosomes and sex chromosomes. Results were plotted in R v.3.5.1.

Then, we ran ADMIXTURE v1.3.0 to assess ancestry. Files stored in Binary Plink (BED) format, obtained from previous step, were used as input. As ADMIXTURE process only digits as data input, we substituted all names of chromosomes with next following numbers. Then we tested cross-validation error assuming 5 likely ancestral populations and detected the lowest value of K=2 for all datasets.

2.2.6 Estimation of population genomic statistics

To clarify allocation of genetic variation among populations of *P. hispaniolensis*, *P. iagoensis* and *P. domesticus* on Cape Verde archipelago, we applied fixation index (F_{ST}) - the property of the distribution of allele frequencies (Holsinger & Weir 2009). F_{ST} is the ratio of within and between population of variation in allele frequencies, in a range from 0 (populations have equal allele frequencies) to 1 (populations are fixed for different alleles). For estimation of fixation index, I first segregated all sequences by species names and then split by geographical distribution. I produced a list of 8 distinct populations and ran VCFtoos v.0.1.13 to subsequently obtained 28 pairwise sequences.

3 Results

3.1 Phylogenetic relationships

I generated both maximum parsimony and best tree.

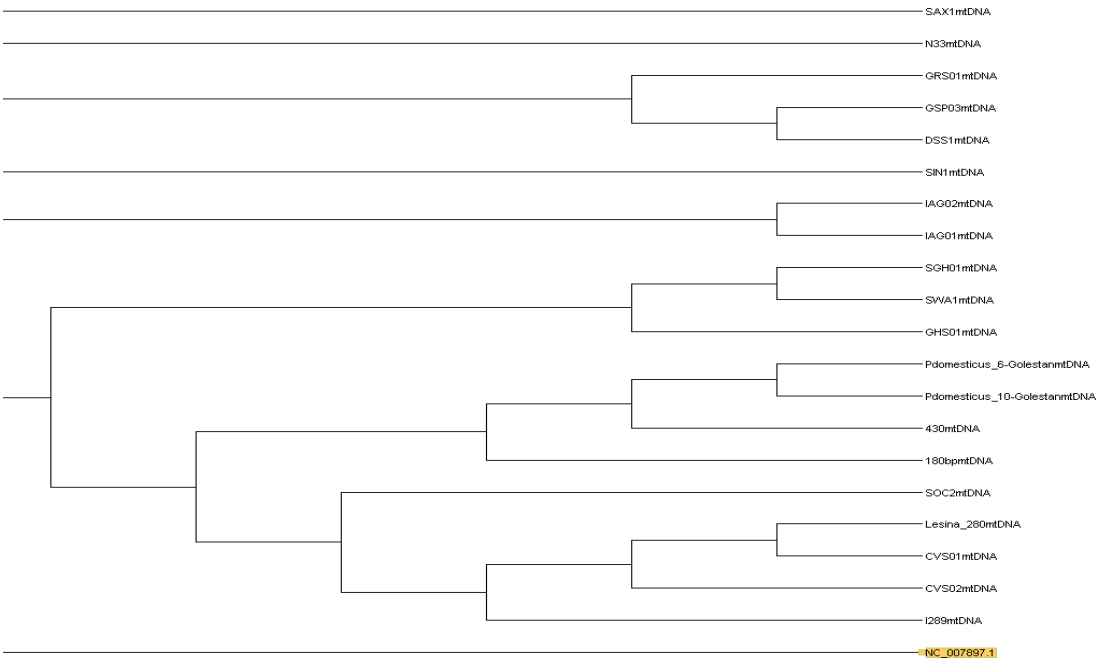

Figure 3. Phylogenetic tree based on maximum parsimony criterion

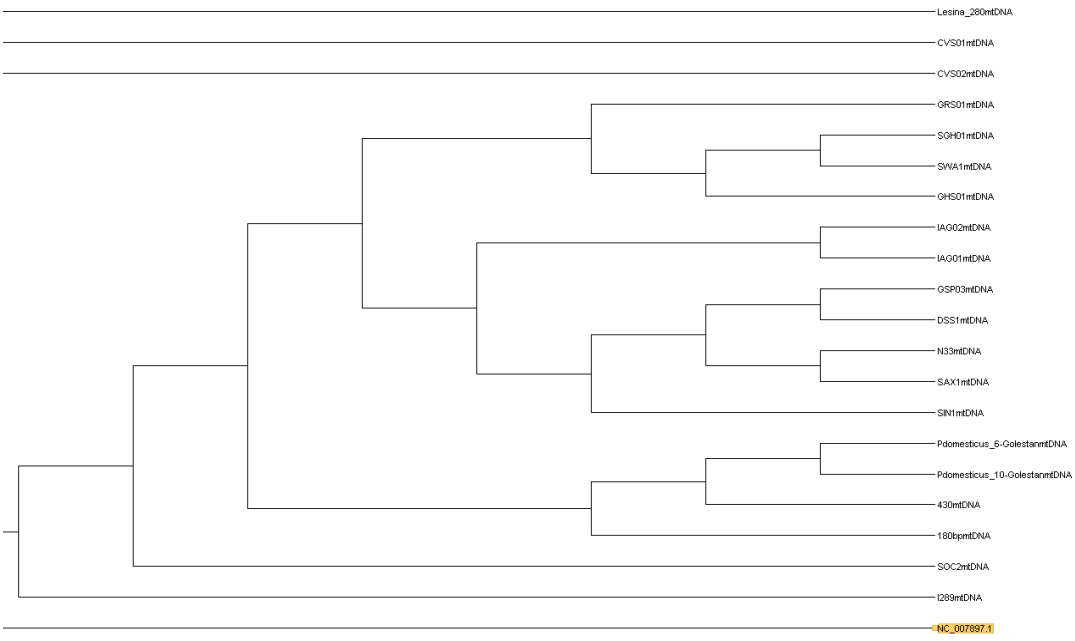

Figure 4. Best tree

I used GTRCAT model, which implies optimization of substitution rates and optimization of site-specific evolutionary rates along with GTR model itself. Both trees support monophyly for Bactrianus subspecies and Iago species. Best tree also demonstrates support for 2 more monophyletic groups: tree (N) with saxaul (SAX) sparrow species, Golden sparrow (GSP) with Dead sea (DSS) sparrow. Surprisingly, 2 individuals of Spanish sparrow (CVS) are depicted as totally different clades. Similar to best tree, maximum parsimony supporting monophyly for GSP and DSS, but is considering Great sparrow (GRS) as paraphyletic to it. Swahili (SWA) and Southern grey-headed (SGH) also form a monophyletic group.

3.2 Demographic history

For detailed investigation of demographic history, a final data set of 29 resequenced genomes was grouped into 3 categories: Iago sparrows only, inhabitants of Cape Verde archipelago and all individuals.

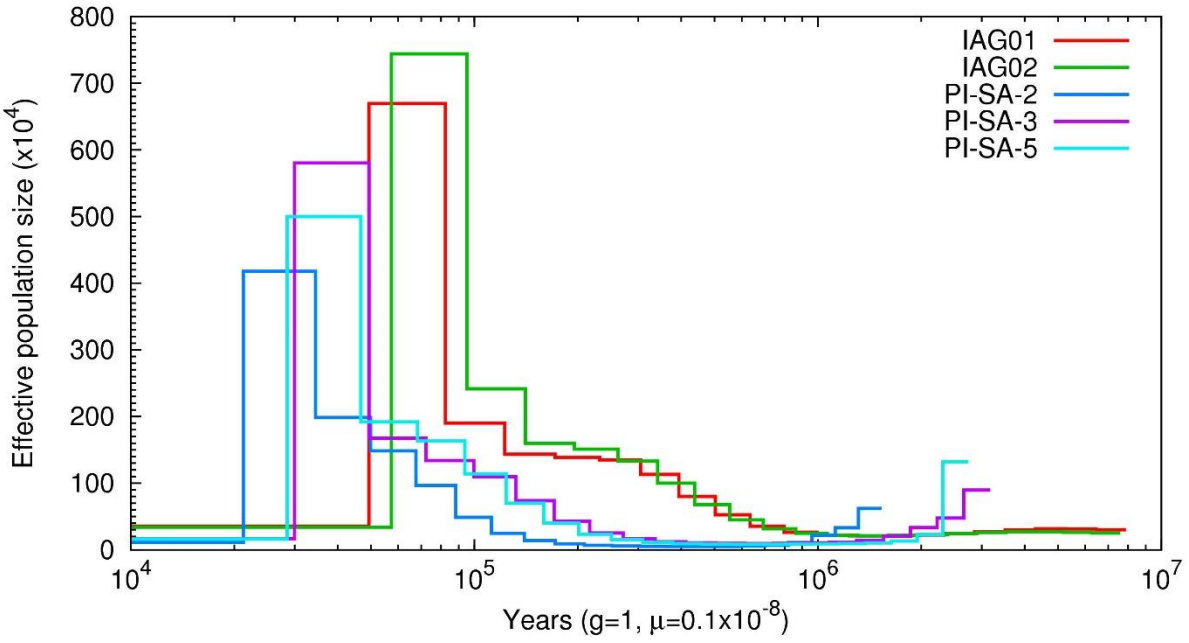

Figure 5. PSMC plot of 5 resequenced Iago sparrow genomes. Composition of curves indicates similarities in changes of N_e among individuals

Results of PSMC analysis processed for the first group, consisting of 5 individuals, demonstrate equivalent shape of curves in general, and thus similar patterns of fluctuations

in effective population size (N_e). After a moderate decline up to 1 Mya, all populations are showing steady growth, following rapid reduction around 20-60 kya and finally converge to relatively constant N_e after that time period. We can also observe a certain difference in N_e around 0,2-1 Mya between tightly convergent pairs of individuals (IAG01+IAG02 and PI-SA-3). Basically, the composition of plot depicts a sudden jump in population size which lasted 10-20 thousand years which then altered with a sudden slump. Dissimilar pattern is likely to point on population structure heterogeneity and might indicate diversification on archipelago.

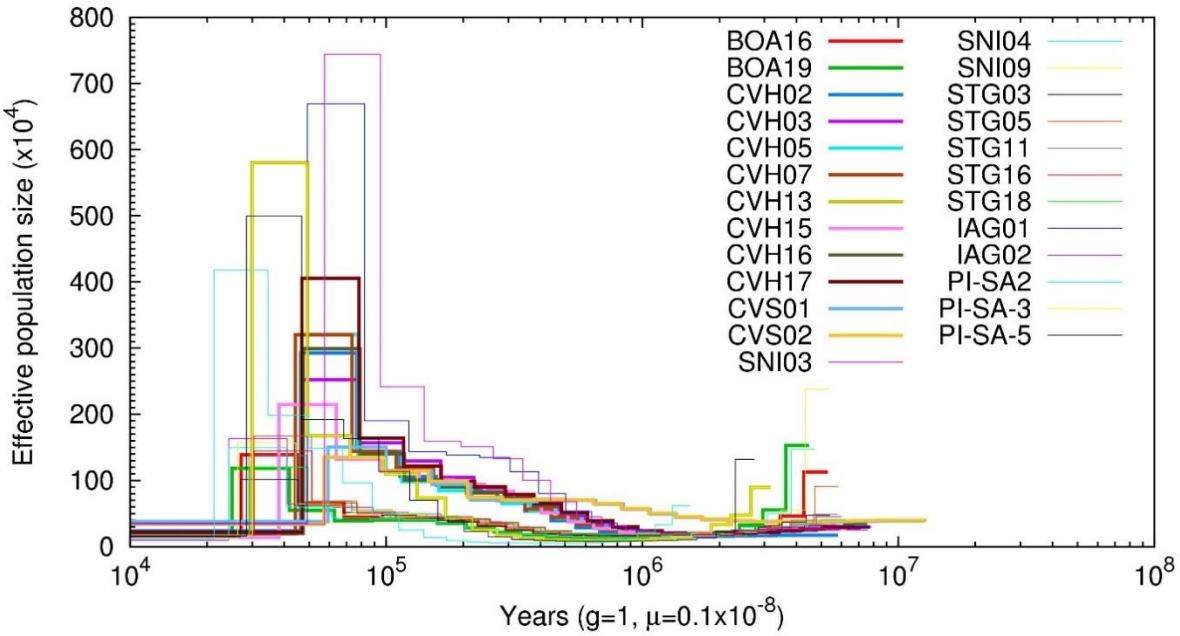

Figure 6. PSMC plot of 25 resequenced genomes of Cape Verde *Passer* inhabitants.

Here we examined genomes of *P. domesticus*, *P. hispaniolensis* and *P. iagoensis.* Common patterns of curves composition resemble the structure of Iago sparrows only on the whole. One of remarkable points is that curves representing individuals of *P. domesticus* (code CVH) are mostly superimposed, forming two highly convergent groups, which in its turn may designate a strong consistency in N_e fluctuations. Individuals of *P. hispaniolensis* populations (codes BOA, SNI and CVS) are also tightly conjugated. As opposite to above, curves corresponding to Iago and Spanish (STG) sparrows tend to be more dispersed and much differentiated in shape.

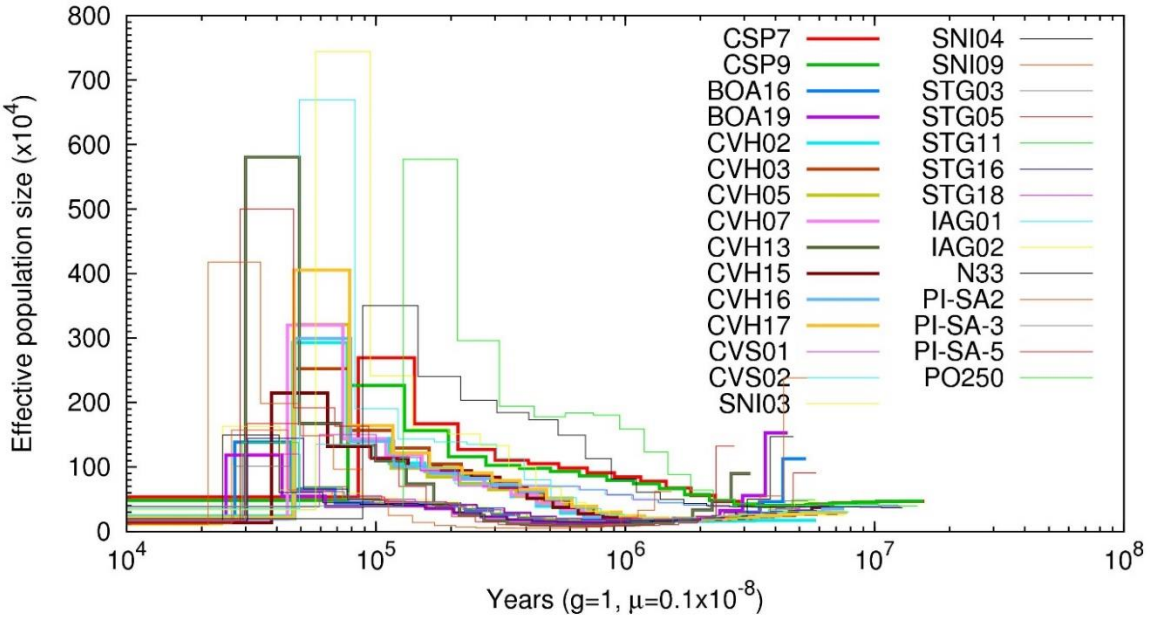

Figure 7. PSMC plot of 29 resequenced genomes of *Passer* individuals.

We added 4 individuals, including 2 genomes of Spanish sparrow from Tenerife population, 1 individual of tree sparrow from Naxos, Italy, and 1 of house sparrow, from Oslo to test descriptive comparison between Cape Verde populations and distinct individuals from different habitats. Common patterns remaining the same as in two previous categories, keeping one peak of abrupt growth preceding immediate decrease of effective population size. Notably we can highlight the dynamics of changes in tree sparrow, which is characterized by smooth rise, and the phase (more than 1 Mya) when house sparrow (Oslo) reached the highest population size.

3.3 Population structure

We constructed PCA plots for both genome-wide and autosomal data set.

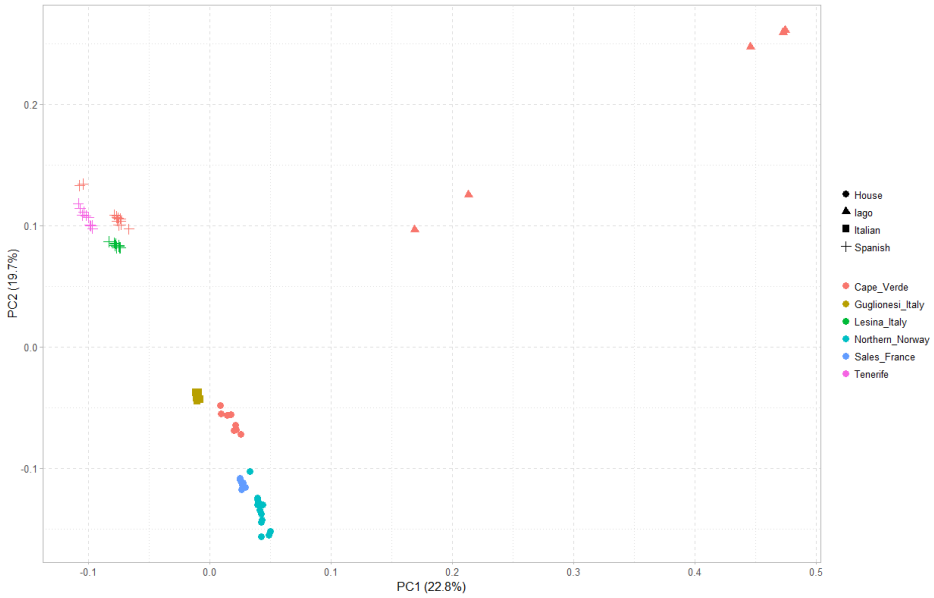

Figure 8. PCA plot of 80 *Passer* genome-wide data with highlighted geographic interconnection

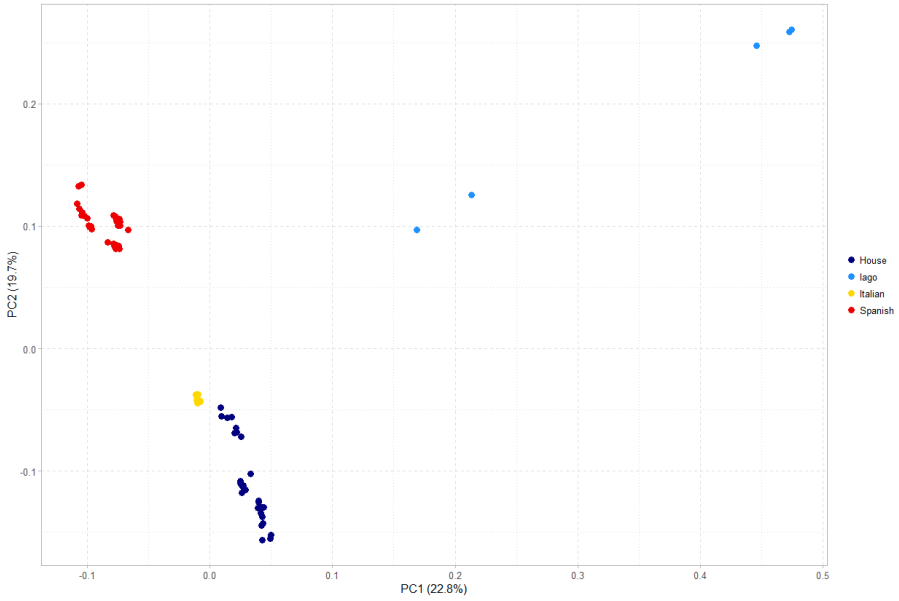

Figure 9. PCA plot of 80 *Passer* genome-wide data with highlighted population interconnection

House, Spanish and Italian sparrow species form a separate cluster, whereas Iago sparrow species are scattered along PC1 (22,8%) and PC2 (19,7%) axes. Since some of Iago sparrow individuals are displaced, a couple of them remains closer to main groups.

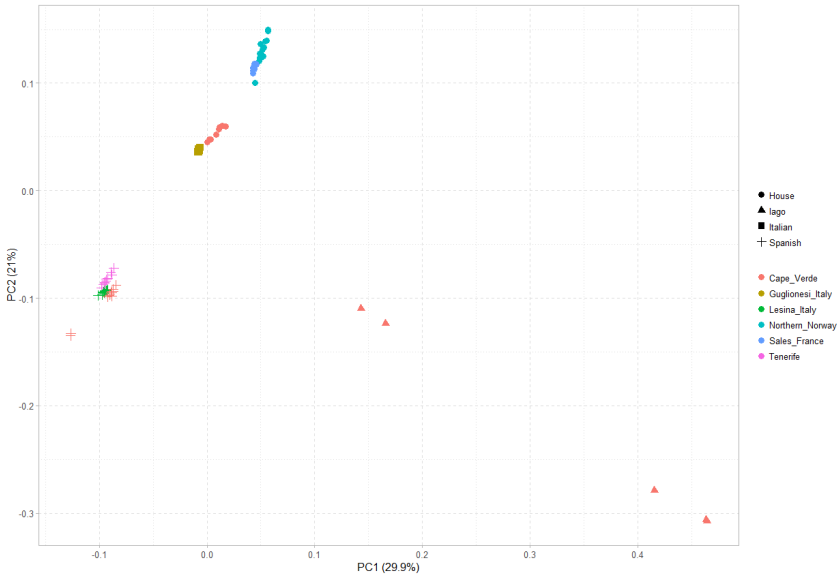

Figure 10. PCA plot of 80 *Passer* autosomal data with highlighted geographic interconnection

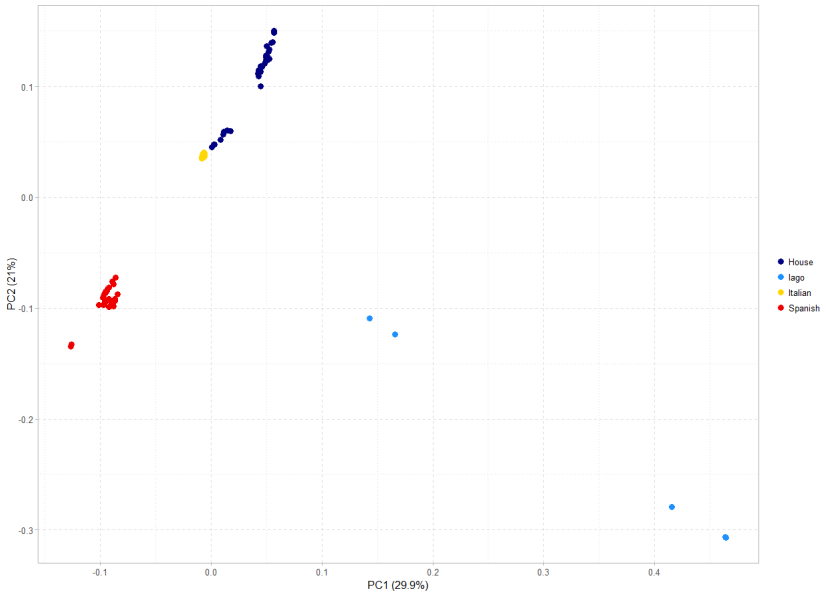

Figure 11. PCA plot of 80 *Passer* autosomal data with highlighted population interconnection

Population of house sparrow species from Cape Verde is extremely close to Guglionese population of Italian sparrow species in autosomal and genome-wide data sets. Spanish

species form more solid cluster, whereas house species are more structured along PC1 (29,9%). Iago sparrows keeping scattered pattern, remaining dispersed along both axes

Then, I performed a model based clustering analysis with ADMIXTURE. Primarily I plotted all values of cross-validation error to test whether the closest to optimal (K=2) parameters are also plausible.

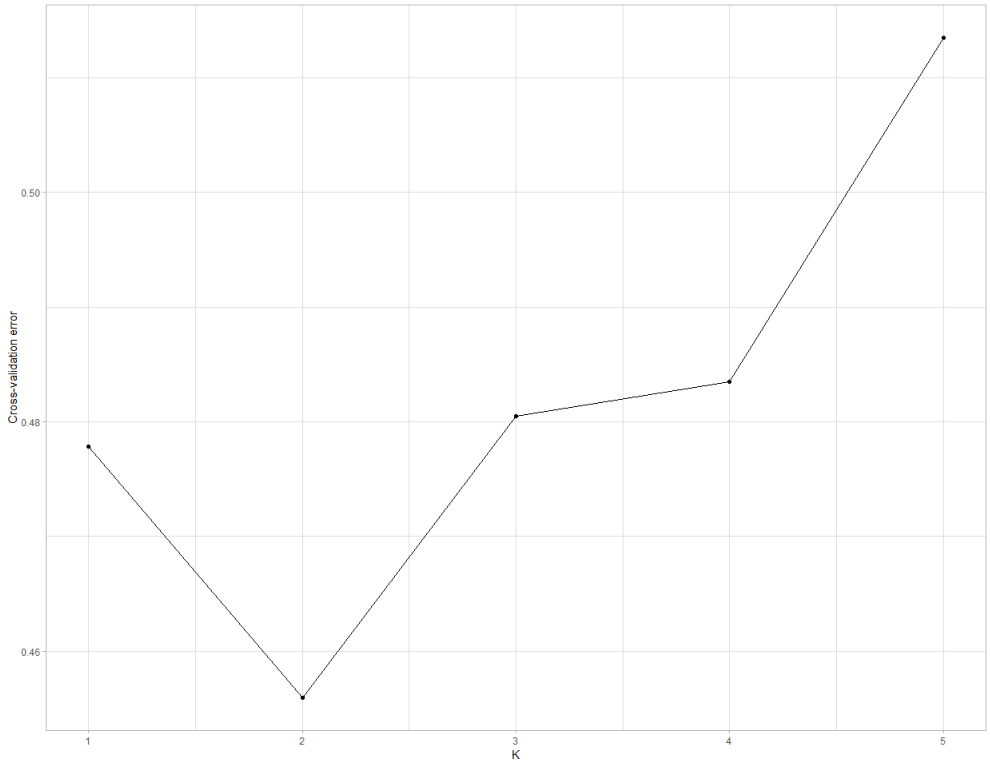

Figure 12. Plot of values for cross-validation error

Due to significant difference of values K=1 and K=3 from the favorable, the only reliable criterion is K=2, that minimizes the cross-validation error and is therefore the best value of K for ADMIXTURE run.

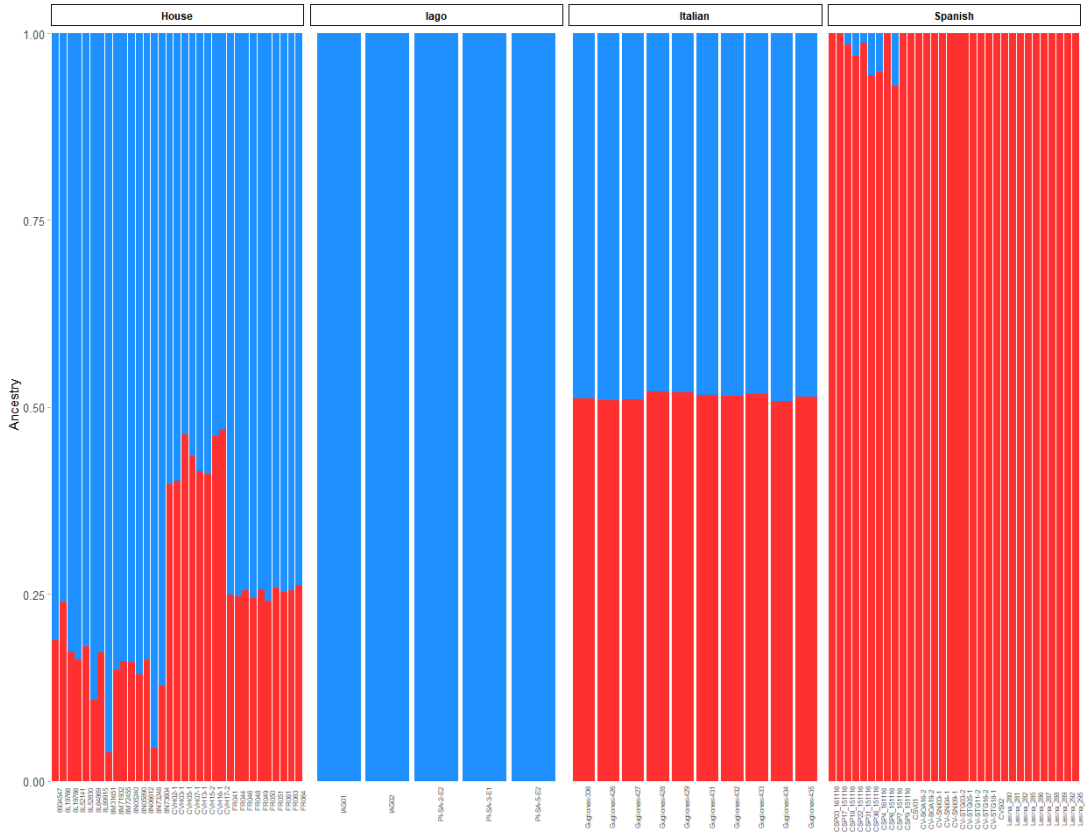

Figure 13. Admixture analysis for K=2

We observe 4 clusters, each corresponding to particular species. Remarkably, the only "pure" cluster contains Iago sparrow individuals, while Spanish have insignificant admixture, house species have a variation of admixture in a range from 46% to 6% and fractions of Italian sparrows cluster are roughly equal.

3.4 Population genomic statistics

Pairwise genome – wide fixation index was calculated for 28 pairs of populations. I used 100 kb non-overlapping windows with a 25 kb step to keep data credibility. Then generated a plot of pairwise weighted F_{ST} for populations from Cape Verde islands, including Italian sparrow species.

Graph is representing genome-wide fixation index for house and Italian (red), Iago and Spanish (blue), Iago and house (yellow), Spanish and house (black). Based on results of PCA analysis, F_{ST} for Italian and house sparrow was attached for additional comparative investigation.

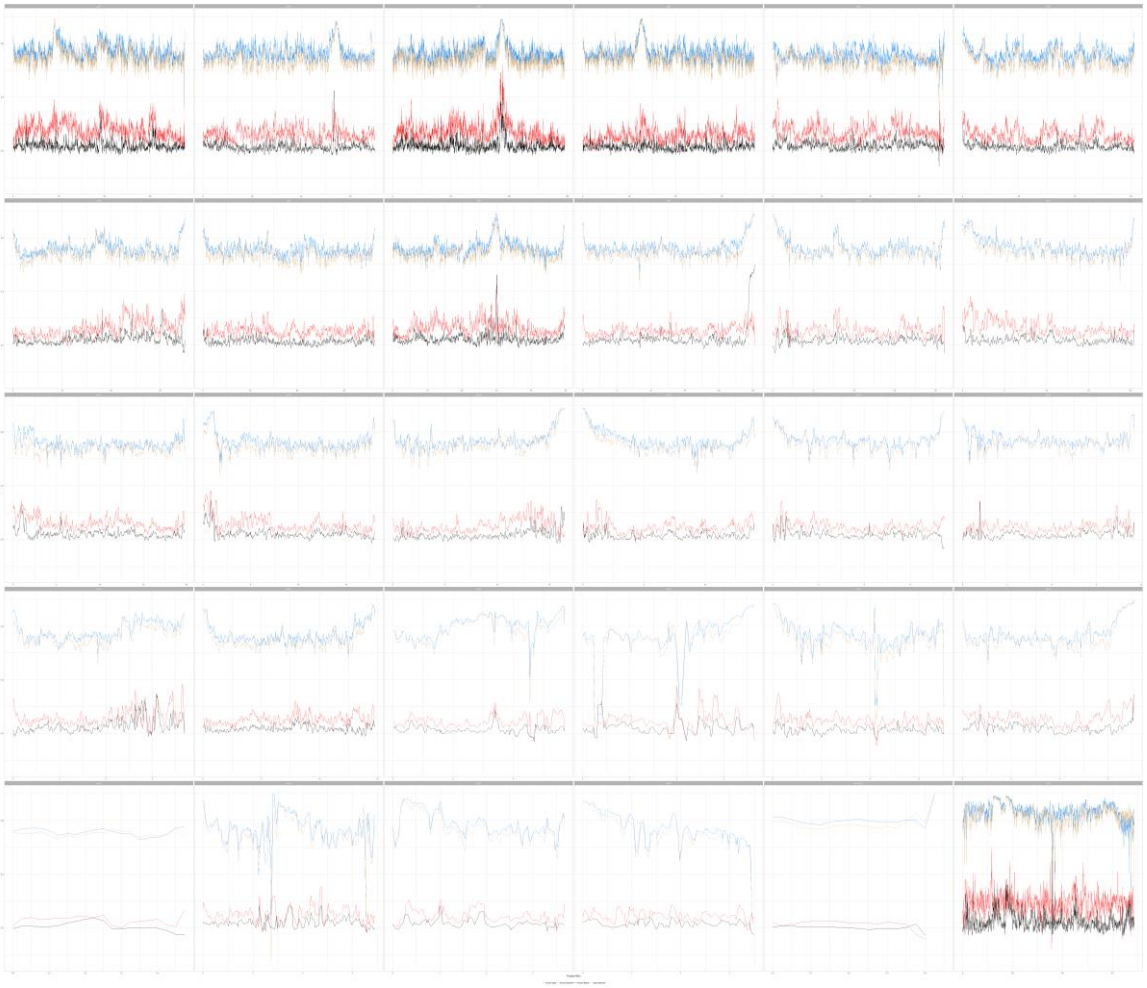

Figure 14. Plots of fixation index. F_{ST} for Iago x Spanish and Iago x house populations is high, whereas values for house x Italian and house x Spanish are low

Table 3. Pairwise fixation index across 8 populations of *Passer* species

Populations	Mean *Fst*	Weighted *Fst*
House CV - House FR	0.005573	0.031002
House CV - House NO	0.021937	0.046613
House CV - Iago	0.39292	0.69594
House CV - Italian	0.0089811	0.034396
House CV - Spanish CV	0.088326	0.13002
House CV - Spanish IT	0.064769	0.10969
House CV - Spanish TEN	0.067877	0.1057
House FR - Iago	0.26452	0.62357
House FR -Italian	0.010207	0.020452
House FR - Spanish CV	0.05508	0.11088
House FR - Spanish IT	0.051441	0.1031
House FR - Spanish TEN	0.052578	0.0976
House FR - House NO	0.015704	0.021407
House NO - Iago	0.29071	0.64667
House NO - Italian	0.02073	0.032435
House NO - Spanish CV	0.061947	0.11732
House NO - Spanish IT	0.05638	0.10897
House NO - Spanish Ten	0.059181	0.10748
Iago - Italian	0.2749	0.65356
Iago - Spanish CV	0.40866	0.73783
Iago - Spanish TEN	0.35467	0.69568
Iago - Spanish IT	0.34761	0.71714

Spanish IT - Italian	0.030335	0.055661
Spanish IT -Spanish TEN	0.036949	0.064463
Spanish IT - Spanish IT	0.018097	0.032101
Spanish CV - Italian	0.035008	0.067127
Spanish CV -Spanish TEN	0.047973	0.079505
Spanish TEN - Italian	0.037695	0.06767

We observe that Iago sparrow species (highlighted yellow and green) show considerably high F_{ST} in relation to either populations of *Passer* genus from both Cape – Verde islands and other geographical regions, in a range of between 62% and 74%. This is in a stark contrast to F_{ST} with both house and Spanish sparrows, thus we may state that *P. iagoensis* is much pure than the rest of species we investigated in this project. Low heterozygosity of Iago sparrow may be due to variety of factors, which we are not able to establish clearly, thus more detailed research with the scope of *P. iagoensis* ecology is required.

4 Discussion

4.1 Demographic history

Using PSMC I described the demographic history of *Passer* species on Cape Verde archipelago. Despite the fact this method was most commonly applicable across the vast majority of papers concerning determination of demographic history and thus regarded as highly efficient and relevant, it also includes a range of caveats. Recommended parameters for proper inference of demographic histories, should imply two chief factors – percentage of missing data (<25%) and mean genome coverage (optional >= 18x) (Nadachowska-Brzyska et al. 2016). I determined the values of essential parameters and imported results to table. Among 29 resequenced genomes, all data samples have a substantial mean mapping percentage (see Table 2). On the contrary to high quality of first parameter, only 8 (see Table 2, highlighted yellow) were of proposed scope of mean depth of coverage criterion (marked

as yellow), whereas the remaining sequence data was out of limits. Such disparity in values of input data may generate erroneous output, lead to bias and thus incorrect interpretation of results. Low depth thus may either underestimate heterozygosity entirely or cause unreliable SNP calls for heterozygous bases. For instance, we used different range of mean depth filters at the initial stage of generation PSMC file for sequences with low and high value, so such discrepancies could modify the final distribution of values. I ran a PSMC for sequences with high mean depth coverage only to verify the impact of given component on overall composition of plot (see Figure 15).

Table 2. Quality of sequencing data

Individual	Mean mapping percentage	Mean depth of coverage
IAG01	96.45	29.0312
IAG02	97.19	28.6674
PI-SA-2	95.06	4.57047
PI-SA-3	95.36	6.92751
PI-SA-5	95.04	6.61272
CV-BOA-16	93.87	6.41403
CV-BOA-19	93.73	5.67344
N33	94.83	28.5674
PO250	96.69	30.7537
CVH02	94.26	5.79189
CVH03	93.88	6.6438
CVH05	93.65	6.255
CVH07	93.58	6.48198
CVH13	94.03	6.45201
CVH15	93.92	6.51744
CVH16	94.29	7.02229
CVH17	94.22	6.91297
CV-SNI-03	93.80	6.78757
CV-SNI-04	93.80	5.98503
CV-SNI-09	93.64	6.46408

CV-STG-03	93.92	6.30794
CV-STG-05	93.35	6.8594
CV-STG-11	93.77	6.87092
CV-STG-16	93.31	6.8459
CV-STG-18	94.02	7.02515
CSP7	95.97	29.1039
CSP9	95.89	29.6527
CVS01	95.19	26.3097
CVS02	96.03	26.8284

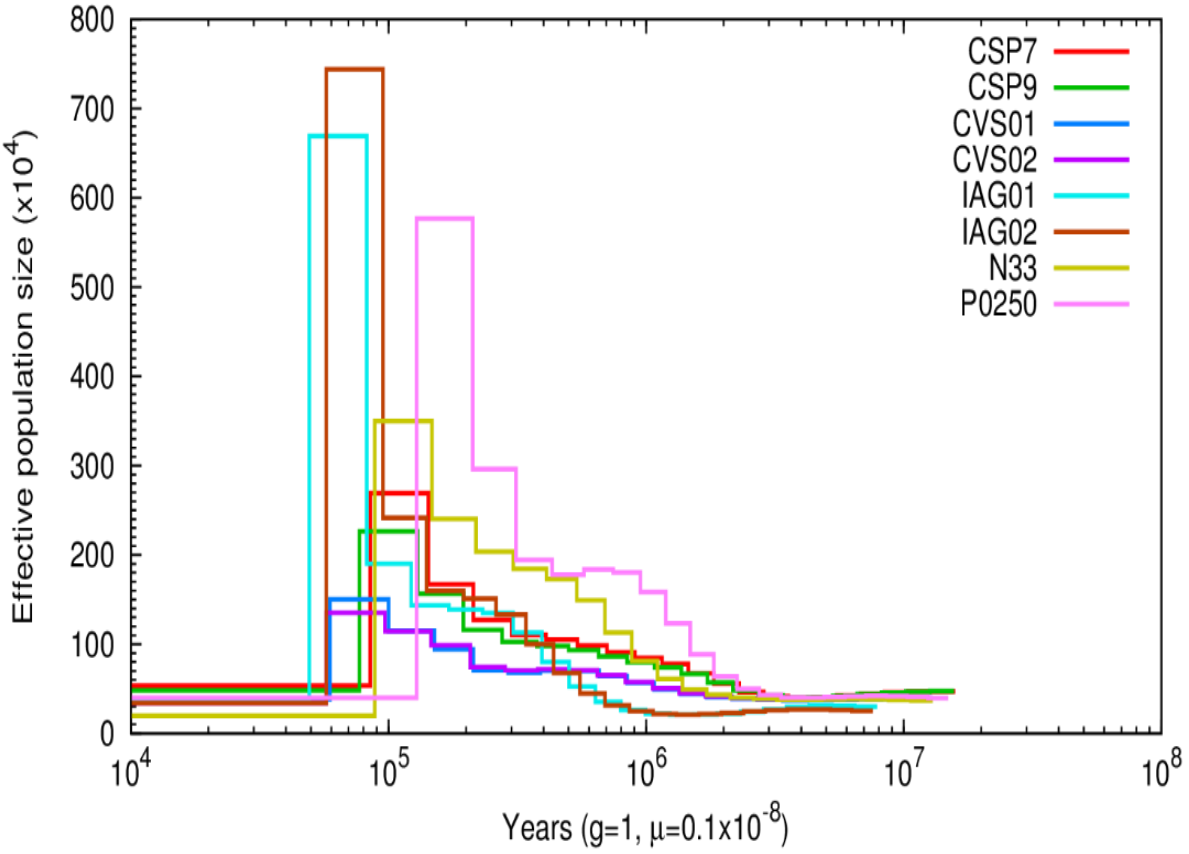

Figure 15. PSMC plot of samples with high mean depth coverage

Overall disposition of curves and the pattern of construction is slightly varied from those where low-coverage sequences included. Samples with mean depth >= 18x tend to be more structured and less scattered than samples with low coverage, although the difference is not substantial. We used genome of *P. domesticus* as the reference for generating PSMC of each individual, and analysis depicts that minor distinctions in patterns are caused due to

difference in mean depth value rather than mapping to the reference genome of different species.

4.2 Population structure specifics

ADMIXTURE and F_{ST} surprisingly revealed exceedingly high (up to 46%) proportions of Spanish sparrow ancestry in the genome of Cape Verde house sparrow population. Mean and weighted F_{ST} between those populations are 0.088326 and 0.13002 respectively. In addition, results of principal component analysis show explicitly relatedness between Italian sparrow and house sparrow (Cape Verde), and the evidence is reinforced by F_{ST} values (mean=0.008981, weighted=0.034396). These findings with a high probability specify several intriguing points. Firstly, considerable high proportion of Spanish sparrow ancestry in population of house sparrow, especially in individuals from *São Vicente* island (Cape Verde), denotes evidence of hybridization between *P. domesticus* and *P. hispaniolensis.* Since Spanish sparrow cluster remains almost unmixed, this indicates also the fact that hybridization occurs in one-way direction. This suggests that Cape Verde population of house sparrows might be an example of a recent hybrid zone between Spanish and house sparrows. Summers-Smith (1988) states that such event could definitely occur, as he documented the existence of one male hybrid individual between species that were mentioned above on *São Vicente* island. Nevertheless, our findings imply that house sparrow population is very admixed, even though they retain typical house sparrow morphological traits. Secondly, distribution of allele frequencies among these species denotes the presence of gene flow between them, and this process seems to be ongoing. It is well-documented (Trier 2014, Hermansen 2014), that *P. domesticus* is prone to breed with *P. hispaniolensis* in contact zones, however due to lack of historical observation and incompleteness of data we may not clarify how much admixture of Spanish ancestry this species had. However, relying on documentary conformation that house sparrow was introduced since 1924 (Summers – Smith 1988), hybridization on Cape Verde is occurred to secondary contact.

5 Concluding remarks and further work

The key finding of my project is the presence of considerable proportion of Spanish sparrow ancestry in the lineage of Cape Verde population house sparrow. Locally this exploration may provide the evidence for continuous hybridization events on restricted isolated area between these two distinct populations, and thus serve as example of recent hybrid zone. Broadly, due to extensive distribution of *P. domesticus*, they most likely may introduce new environments and come into contact with endemic species of *Passer* genus followed by interbreeding and backcrossing. However, results of my project clearly demonstrate the absence of house sparrow ancestry in Iago sparrow genome and maintenance of homozygosity in given population.

First challenging point regarding my project, was initial stage of extracting DNA. Due to the aspects that implied collecting blood samples from Cape Verde by collaborators, obtained samples were of different technique for storing blood, and thus a part of them was of low quality, that resulted in limitation of total number included in a final data set. Difference in depth coverage of genome might influenced the final result of PSMC, thus homogenous sequences of the same quality are required for getting results of high confidence. Our dataset included 5 Iago sparrow individuals from only one location, and such relatively restricted geographical range of samples is not providing explicitly information about real characteristics of population structure. Ideally, it would be better to have samples of Iago sparrows from the same place as Spanish and house sparrows.

References

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J., Bierne, N., ... & Butlin, R. K. (2013). Hybridization and speciation. *Journal of evolutionary biology*, *26*(2), 229-246

Aboim, M. A., Mavárez, J., Bernatchez, L., & Coelho, M. M. (2010). Introgressive hybridization between two Iberian endemic cyprinid fish: a comparison between two independent hybrid zones. *Journal of Evolutionary Biology*, *23*(4), 817-828.

Ait Belkacem, A., Gast, O., Stuckas, H., Canal, D., LoValvo, M., Giacalone, G., & Päckert, M. (2016). North African hybrid sparrows (Passer domesticus, P. hispaniolensis) back from oblivion–ecological segregation and asymmetric mitochondrial introgression between parental species. *Ecology and evolution*, *6*(15), 5190-5206.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120.

Bonfield, J. K., & Mahoney, M. V. (2013). Compression of FASTQ and SAM format sequencing data. *PloS one*, *8*(3), e59190.

Brumfield, R. T., Beerli, P., Nickerson, D. A., & Edwards, S. V. (2003). The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution*, *18*(5), 249-256.

Currat, M., & Excoffier, L. (2011). Strong reproductive isolation between humans and Neanderthals inferred from observed patterns of introgression. *Proceedings of the National Academy of Sciences*, 201107450.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... & McVean, G. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156-2158.

Dannemann, M., Andrés, A. M., & Kelso, J. (2016). Introgression of Neandertal-and Denisovan-like haplotypes contributes to adaptive variation in human Toll-like receptors. *The American Journal of Human Genetics*, *98*(1), 22-33.

Elgvin, T. O., Trier, C. N., Tørresen, O. K., Hagen, I. J., Lien, S., Nederbragt, A. J., ... & Sætre, G. P. (2017). The genomic mosaicism of hybrid speciation. *Science advances*, *3*(6), e1602996.

Ellstrand, Norman C., and Loren H. Rieseberg. "When gene flow really matters: gene flow in applied evolutionary biology." *Evolutionary applications* 9.7 (2016): 833-836.

Futuyma, D. J. (2013). Evolution. 3rd edn. Sunderland, MA. 464-465

Hermansen, J. S., Saether, S. A., Elgvin, T. O., Borge, T., Hjelle, E., & SÆTRE, G. P. (2011). Hybrid speciation in sparrows I: phenotypic intermediacy, genetic admixture and barriers to gene flow. *Molecular Ecology*, *20*(18), 3812-3822.

Hermansen, J. S., Haas, F., Trier, C. N., Bailey, R. I., Nederbragt, A. J., Marzal, A., & Sætre, G. P. (2014). Hybrid speciation through sorting of parental incompatibilities in I talian sparrows. *Molecular ecology*, *23*(23), 5831-5842.

Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F ST. *Nature Reviews Genetics*, *10*(9), 639.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*, *25*(14), 1754-1760.

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.

Mallet, James. "Hybridization as an invasion of the genome." *Trends in ecology & evolution* 20.5 (2005): 229-237.

Mallet, James. "Hybrid speciation." *Nature* 446.7133 (2007): 279.

Mays Jr, H. L., Hung, C. M., Shaner, P. J., Denvir, J., Justice, M., Yang, S. F., ... & Primerano, D. A. (2018). Genomic analysis of demographic history and ecological niche modeling in the endangered Sumatran rhinoceros Dicerorhinus sumatrensis. *Current Biology*, *28*(1), 70-76.

Meier, J. I., Marques, D. A., Mwaiko, S., Wagner, C. E., Excoffier, L., & Seehausen, O. (2017). Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature communications*, *8*, 14363.

Nadachowska-Brzyska, K., Burri, R., Olason, P. I., Kawakami, T., Smeds, L., & Ellegren, H. (2013). Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-genome re-sequencing data. *PLoS genetics*, *9*(11), e1003942.

van Oeveren, J., & Janssen, A. (2009). Mining SNPs from DNA sequence data; computational approaches to SNP discovery and analysis. In *Single Nucleotide Polymorphisms* (pp. 73-91). Humana Press, Totowa, NJ.

Pinheiro de Melo (2013). Darwin's Sparrows': a new model for the study of speciation in nature?

Ravinet, M., Yoshida, K., Shigenobu, S., Toyoda, A., Fujiyama, A., & Kitano, J. (2018). The genomic landscape at a late stage of stickleback speciation: High genomic divergence interspersed by small localized regions of introgression. *PLoS genetics*, *14*(5), e1007358.

Ravinet, M., Elgvin, T. O., Trier, C., Aliabadian, M., Gavrilov, A., & Sætre, G. P. (2018). Signatures of human-commensalism in the house sparrow genome. *Proc. R. Soc. B*, *285*(1884), 20181246.

Ringnér, M. (2008). What is principal component analysis?. *Nature biotechnology*, *26*(3), 303.

Salzburger, W., Meyer, A., Baric, S., Verheyen, E., & Sturmbauer, C. (2002). Phylogeny of the Lake Tanganyika cichlid species flock and its relationship to the Central and East African haplochromine cichlid fish faunas. *Systematic biology*, *51*(1), 113-135.

Shastry, B. S. (2009). SNPs: impact on gene function and phenotype. In *Single Nucleotide Polymorphisms* (pp. 3-22). Humana Press, Totowa, NJ.

Soltis, P. S. (2013). Hybridization, speciation and novelty. *Journal of Evolutionary Biology*, *26*(2), 291-293.

Summers-Smith, D. (1984). The Sparrows of the Cape Verde Islands. *Ostrich*, *55*(3), 141-146.

Trier, C. N., Hermansen, J. S., Sætre, G. P., & Bailey, R. I. (2014). Evidence for mito-nuclear and sex-linked reproductive barriers between the hybrid Italian sparrow and its parent species. *PLoS genetics*, *10*(1), e1004075.

Turner, T. F., Dowling, T. E., Broughton, R. E., & Gold, J. R. (2004). Variable microsatellite markers amplify across divergent lineages of cyprinid fishes (subfamily Leusicinae). *Conservation Genetics*, *5*(2), 279-281.

Yamada, M., Higuchi, M., & Goto, A. (2001). Extensive introgression of mitochondrial DNA found between two genetically divergent forms of threespine stickleback, Gasterosteus aculeatus, around Japan. *Environmental Biology of Fishes*, *61*(3), 269-284.

Zou, G. (2004). A modified poisson regression approach to prospective studies with binary data. *American journal of epidemiology*, *159*(7), 702-706.

Appendix

Conversion from SAM to BAM

```
cd~/alignment/sparrows_sam
module load samtools
samtools view ID.sam –b –o ID.bam
```

Sorting BAMs

```
samtools sort ID.bam -o ID_sort.bam
```

Generating VCF

```
cd~/sparrows_bam
module purge
module load
module load samtools/1.3
module load bcftools/1.3
cd ~/sparrows_bam
samtools mpileup -Rug -t DP,AD -C 50 -f $REF \
./sparrows_bam/*_sort.bam | bcftools call -f GQ,GP \
-vmO z -o ./sparrows_vitalii_final.vcf.gz
```

Filtering SNPs

```
vcftools --gzvcf sparrows_vitalii_final.vcf.gz --remove-indels --maf 0.05 --max-missing 0.8 --minDP 10 --maxDP 40 --min-meanDP 5 –min GQ 20 --max-meanDP 60--recode --stdout | gzip -c > sparrows_vitalii_final_filtered.vcf.gz
```

Conversion from BAM to FASTQ

```
module purge
module load python2/2.7.9
module load pyfasta/pyfasta-0.5.2
module load psmc/January14th_2016
cd ~/sparrows_bam

#set variables
ID=~/sparrows_bam/name_of individual #without filename extension

FASTQ=${ID}.fq.gz

PSMCFA=${ID}.psmcfa

PSMCFA_AUTOSOME=${ID}_auto.psmcfa

PSMCFA_SEX=${ID}_sex.psmcfa
```

Conversion from FASTQ to psmcfa

fq2psmcfa -q 20 -g 10000 -s 100 $FASTQ > $PSMCFA

creating a PSMC file removing mitochondrial DNA and sex chromosomes

pyfasta extract --fasta $PSMCFA --exclude chrZ mtD --header > $PSMCFA_AUTOSOME

generating PSMC with sex chromosomes only

pyfasta extract --fasta $PSMCFA --header chrZ > $PSMCFA_SEX

Conversion from psmcfa to psmc

module load psmc/January14th_2016

ID=~/sparrows_bam/name_of individual #without filename extension

psmc -N 30 -t 15 -r 5 -p 4+19*2+3 -o ${ID}.psmc ${ID}_auto.psmcfa

#plotting results

module load gnuplot/4.6.0

#labels are set in alphabet order

psmc_plot.pl -u 1.4e-9 –M CSP7,CSP9,CVS01,CVS02,IAG01,IAG02,N33,PO250 -g 1 -m 10 -L sparrows_final

Plink linkage pruning

module load plink2

cd ~/sparrows_bam

VCF=~/sparrow_vitalii_final_filtered.vcf.gz

runing linkage pruning

plink2 --vcf $VCF --double-id --allow-extra-chr --chr-set 30 --set-missing-var-ids @:# --indep-pairwise 50 10 0.1 --out sparrows_maf_0.05_ld

loci list preparation

grep -f targets sparrows_maf_0.05_ld.prune.in > hsi_maf_0.05_ld_genome.prune.in

loci list for sex chromosomes

grep "chrZ" sparrows_maf_0.05_ld_genome.prune.in > sparrows_maf_0.05_ld_sex.prune.in

loci list for autosomes

grep -v "chrZ" hsi_maf_0.05_ld_genome.prune.in > sparrows_maf_0.05_ld_autosome.prune.in

Plink PCA

running PCA for whole genome

plink --vcf $VCF --double-id --allow-extra-chr --chr-set 30 --set-missing-var-ids @:# --extract sparrows_maf_0.05_ld_genome.prune.in --maf 0.05 --geno 0.1 --pca --out sparrows_maf_0.05_genome_miss

running PCA for autosome only

plink --vcf $VCF --double-id --allow-extra-chr --chr-set 30 --set-missing-var-ids @:# --extract sparrows_maf_0.05_ld_autosome.prune.in --maf 0.05 --geno 0.2 --recodeA --out sparrows_maf_0.05_autosome_miss

running PCA for sex chromosomes only

plink --vcf $VCF --double-id --allow-extra-chr --chr-set 30 --set-missing-var-ids @:# --extract sparrows_maf_0.05_ld_sex.prune.in --maf 0.05 --geno 0.1 --pca --out sparrows_maf_0.05_sex_miss

Plotting PCA

rm(list = ls())

library(tidyverse)

read in main data

sparrowData <- tbl_df(read.csv("~/pca_data/sparrows_list.csv"))

read in pca_data

eigenvec_data <- "~/pca_data/sparrows_maf_0.05_genome_miss.eigenvec"

eigenval_data <- "~/pca_data/sparrows_maf_0.05_genome_miss.eigenval"

plot_path <- "~/pca_data/genome_pca_sparrows.pdf"

read in genome data

eigenvec <- tbl_df(read.table(eigenvec_data))[, -1]

join to main data

genome_PCA <- left_join(sparrowData, eigenvec, by = "ind")

also read in eigenvalues

```
genome_eigenval <- scan(eigenval_data)

genome_eigenval <- tbl_df(data.frame(pc = seq(1, ncol(eigenvec) - 1), eigenval
= genome_eigenval))

# mutate to produce PVE

genome_eigenval <- mutate(genome_eigenval, pve = eigenval/sum(eigenval))

# importance of eigenvectors

a <- ggplot(genome_eigenval, aes(pc, pve)) + geom_bar(stat = "identity")

a + theme_light()

# make plot with populations coloured

b <- ggplot(genome_PCA, aes(PC1, PC2, colour = pop, shape = spp2)) +
geom_point(size = 3)

b <- b + xlab(paste0("PC1 (", signif(genome_eigenval$pve[1], 3)*100, "%)"))

b <- b + ylab(paste0("PC2 (", signif(genome_eigenval$pve[2], 3)*100, "%)"))

b <- b + theme_light() + coord_equal()

b + theme(legend.position = "right",
        legend.title = element_blank(),
        legend.text = element_text(size = 10),
        panel.grid.minor.x = element_line(linetype = 3),
        panel.grid.minor.y = element_line(linetype = 3),
        panel.grid.major.x = element_line(linetype = 2),
        panel.grid.major.y = element_line(linetype = 2),
        axis.title = element_text(size = 12),
        axis.text = element_text(size = 10))

image_path <- "~/pca_data/"

dev.print(pdf, paste0(image_path, "PCA_sparrows.pdf"), width = 10, height = 8)

# make plot with spp coloured (main text)

b <- ggplot(genome_PCA, aes(PC1, PC2, colour = spp2)) + geom_point(size = 3)

b <- b + scale_colour_manual(values = c("navyblue", "dodgerblue1", "gold",
"red2"))
```

```
b <- b + xlab(paste0("PC1 (", signif(genome_eigenval$pve[1], 3)*100, "%)"))
b <- b + ylab(paste0("PC2 (", signif(genome_eigenval$pve[2], 3)*100, "%)"))
b <- b + theme_light() + theme_light()+ coord_equal()
b + theme(legend.position = "right",
        legend.title = element_blank(),
        legend.text = element_text(size = 10),
        panel.grid.minor.x = element_line(linetype = 3),
        panel.grid.minor.y = element_line(linetype = 3),
        panel.grid.major.x = element_line(linetype = 2),
        panel.grid.major.y = element_line(linetype = 2),
        axis.title = element_text(size = 12),
        axis.text = element_text(size = 10))

dev.print(pdf, paste0(poster_plot_path, "PCA_spp.pdf"), width = 10, height = 8)
dev.print(postscript, paste0(poster_plot_path, "PCA_spp.eps"), width = 10,
height = 8)

# make plot with spp coloured (main text)
b <- ggplot(genome_PCA, aes(PC1, PC2, colour = spp2)) + geom_point(size = 3)
b <- b + scale_colour_manual(values = c("navyblue", "dodgerblue1", "gold",
"red2"))
b <- b + xlab(paste0("PC1 (", signif(genome_eigenval$pve[1], 3)*100, "%)"))
b <- b + ylab(paste0("PC2 (", signif(genome_eigenval$pve[2], 3)*100, "%)"))
b <- b + theme_light() + coord_equal()
b + theme(legend.position = "none",
        legend.title = element_blank(),
        legend.text = element_text(size = 10),
        panel.grid.minor.x = element_blank(),
```

```
        panel.grid.minor.y = element_blank(),
        panel.grid.major.x = element_blank(),
        panel.grid.major.y = element_blank(),
        axis.title = element_text(size = 12),
        axis.text = element_text(size = 10))

dev.print(pdf, paste0(image_path, "PCA_spp_presentation.pdf"), width = 8, height = 8)
dev.print(postscript, paste0(image_path, "PCA_spp_presentation.eps"), width = 8, height = 8
```

Substitution of chromosome names

```
sed -i -e 's/chr1A/29/g' sparrows_maf_0.05_genome_miss.bed
sed -i -e 's/chrLGE22/30/g' sparrows_maf_0.05_genome_miss.bed
sed -i -e 's/chrZ/31/g' sparrows_maf_0.05_genome_miss.bed
```

Running admixture

```
module load admixture
cd~/plink_files
for K in 1 2 3 4 5
do admixture –cv sparrows_maf_0.05_ld_genome.bed $K | tee log {K}.out
done
```

Plotting admixture

```
rm(list = ls())
library(tidyverse)
library(gridExtra)
crossvalue_error_infile <- "~/admixture/cv_error.txt"
# read in individual data
ind <- as.character(read.table("~/admixture/sparrows_maf_0.05_genome_miss.fam")[, 1])
```

```
sparrowData <- tbl_df(read.csv("./admixture/sparrow_list.csv"))
group_by(sparrowData, spp2) %>% tally()
# first read in and examine cross-validation error
cv_error <- tbl_df(read.table(crossvalue_error_infile,
                              col.names = c("K", "error")))
# plot
a <- ggplot(cv_error, aes(K, error)) + geom_point() + geom_line()
a + ylab("Cross-validation error") + theme_light()

plot_path <- "~/admixture/"
dev.print(pdf, paste0(plot_path, "crossvalue_admixture.pdf"), height = 5, width
= 7)
## K2
# now plot actual data
K2 <- tbl_df(read.table("~/admixture/sparrows_maf_0.05_genome_miss.2.Q",
                        col.names = c("one", "two")))

# add individuals
K2 <- mutate(K2, ind = ind)
# join to ind data
K2 <- left_join(sparrowData, K2, by = "ind")
# gather to plot
K2 <- gather(K2, key = cluster, value = q, -ind, -spp, -spp2, -pop)
## K3
# now plot actual data
K3 <- tbl_df(read.table("~/admixture/sparrow_maf_0.05_genome_miss.3.Q",
                        col.names = c("one", "two", "three")))
# add individuals
```

```
K3 <- mutate(K3, ind = ind)

# join to ind data

K3 <- left_join(myData, K3, by = "ind")

# gather to plot

K3 <- gather(K3, key = cluster, value = q, -ind, -spp, -spp2, -pop)

# set colour palettes

k2_cols <- c("dodgerblue", "firebrick1")

k3_cols <- c("dodgerblue", "gold",  "firebrick1")

k3_cols <- c("firebrick1","dodgerblue", "gold")

# make plot

a <- ggplot(K2, aes(as.character(ind), q, fill = cluster)) + geom_bar(stat =
"identity")

a <- a + scale_fill_manual(values = k2_cols)

a <- a + xlab(NULL) + ylab("Ancestry") + scale_y_continuous(limits = c(0,1.01),
expand = c(0, 0))

a <- a + facet_wrap(~spp2, scales = "free_x", nrow = 1)

a <- a + theme_light() + theme(legend.position = "none",

                panel.border = element_blank(),

                panel.grid = element_blank(),

                axis.text.y = element_text(size = 10),

                axis.text.x = element_text(size = 6, angle = 90),

                axis.ticks.x = element_blank(),

                strip.text.x = element_text(colour = "black", face = "bold"),

                strip.background = element_rect(colour = "black", fill = "white"))

a

plot_path <- "~/admixture/"

dev.print(pdf, paste0(plot_path, "K2.pdf"), height = 5, width = 8)

# make plot
```

```
b <- ggplot(K3, aes(ind, q, fill = cluster)) + geom_bar(stat = "identity")

b <- b + scale_fill_manual(values = k3_cols)

b <- b + xlab(NULL) + ylab("Ancestry") + scale_y_continuous(limits = c(0,1.01),
expand = c(0, 0))

b <- b + facet_wrap(~spp2, scales = "free_x", nrow = 1)

b <- b + theme_light() + theme(legend.position = "none",
                    panel.border = element_blank(),
                    panel.grid = element_blank(),
                    axis.text.y = element_text(size = 10),
                    axis.text.x = element_text(size = 6, angle = 90),
                    axis.ticks.x = element_blank(),
                    strip.text.x = element_text(colour = "black", face = "bold"),
                    strip.background = element_rect(colour = "black", fill = "white"))

b

dev.print(pdf, paste0(plot_path, "K3.pdf"), height = 5, width = 8)


grid.arrange(a + theme(axis.text.x = element_blank(),
                    axis.ticks.x = element_blank()),
          b + theme(axis.text.x = element_blank(),
                    axis.ticks.x = element_blank()))


plot_path <- "~/admixture/"

dev.print(pdf, paste0(plot_path, "admixture.pdf"), height = 7, width = 14)

module load vcftools
```

Estimation of pairwise Fst

```
VCF=~/sparrow_vitalii_final_filtered.vcf.gz

WINDOW=100000

STEP=25000
```

```
OUT_PREF=~/sparrows

# run vcftools

echo "Running vcftools W&C fst for $1 and $2"

vcftools --gzvcf $VCF --fst-window-size $WINDOW --fst-window-step $STEP --
weir-fst-pop ${1} --weir-fst-pop ${2} –out ${OUT_PREF}_${1%.*}_${2%.*}
```

Producing data.frame

```
rm(list = ls())

library(tidyverse)

# import files

my_files <- list.files(path = "~/fst_sparrows", pattern = "*.fst", full.names = T)

# read in given data

fst_data <- lapply(my_files, read_delim, delim = "\t", skip = 1,

                   col_names = c("chr", "start", "stop", "n_var", "weighted_fst",
"mean_fst"))

# give names to the list

names <- sub("sparrows_", "", sub(".windowed.weir.fst", "",
basename(my_files)))


names <- sub("italian", "it", sub("spanish", "sp", sub("house", "ho", sub("iago",
"iag", names))))

names(fst_data) <- names

# rename names of columns

fst_data <- sapply(1:length(fst_data), function(x){ y <- names[x]
names(fst_data[[x]]) <- paste0(names(fst_data[[x]]), "_", y) fst_data[[x]]},
simplify = F)

# create an id for each dataset

fst_data<- lapply(fst_data, function(x){#x$id <- paste0(x[, 1], "_", start) x})

# add id column

fst_data <- lapply(fst_data, function(x){

  chr <- select(x, contains("chr")) %>% pull()
```

```
  start <- select(x, contains("start")) %>% pull()
  x$id <- paste0(chr, "_", start)  x})
# join all together
a <- left_join(fst_data[[1]], fst_data[[2]], by = "id")
b <- left_join(a, fst_data[[3]])
c <- left_join(b, fst_data[[4]])
# clear up
fst <- select(k, chr = chr_ho_cv_iag, start = start_ho_cv_iag, stop =
stop_ho_cv_iag,
            contains("weighted"))
# final name clear up
colnames(fst) <- sub("weighted_", "", colnames(fst))
# set chr type
fst <- mutate(fst, chr = factor(chr))
# sort out scaffolds
levels(fst$chr)[grep("scaffold", levels(fst$chr))] <- "scaffold"
# set factor
fst$chr_type <- factor(ifelse(fst$chr == "chrZ", "sex", "auto"))
write_csv(fst, "~/fst/sparrows.csv")
```

Plotting Fst

```
rm(list = ls())
library(tidyverse)
infile <- "~/fst/sparrows.csv"
genome <- tbl_df(read.csv(infile, header = T))
cape_verde  <-  genome %>% select(chr, start, stop, fst_ho_cv_iag_cape_verde,
fst_ho_cv_sp_cape_verde, fst_ho_cv_it_italy, fst_iag_cv_sp_cape_verde)
# first generate mean fst
cape_verde  %>%
```

```
  select(contains("fst")) %>%
  summarise_each(funs(mean))
# reorder chr levels
cape_verde$chr <- factor(cape_verde$chr, levels(cape_verde$chr)[c(1, 11, 12, 22:28, 2:10, 13, 14:21, 29, 30)])
# plot genome-wide house spanish fst
a <- ggplot(cape_verde, aes(start/10^6, fst_ho_cv_iag_cape_verde)) + geom_line()
a <- a + xlab("Position (Mb)") + ylab(expression(italic(F)[ST]))
a + facet_wrap(~chr, scales = "free_x") + theme_light()
# gather the data and plot it for all comparisons
fst <- select(cape_verde, chr, start, contains("fst"))
fst <- gather(fst, comp, fst, -chr, -start)
# clean up comp factor
fst$comp <- factor(fst$comp, labels = c("house-Iago", "house-Spanish", "house-Italian", "Iago-Spanish"))
# reorder to make it easier to visualise
fst$comp <- factor(fst$comp, levels(fst$comp)[c(1, 2, 3, 4)])
# plot genome-wide house spanish fst
a <- ggplot(fst, aes(start/10^6, fst, colour = comp)) + geom_line()
a <- a + xlab("Position (Mb)") + ylab(expression(italic(F)[ST]))
a <- a + scale_alpha_manual(values = c(0.3, 0.3, 0.3, 0.3))

a <- a + scale_colour_manual(values = c("dodgerblue", "firebrick1", "black", "burlywood"))
a <- a + facet_wrap(~chr, scales = "free_x")
a + theme_light() + theme(legend.position = "bottom", legend.title = element_blank())
```