Extensive introgression despite Haldane’s rule: insights from a grasshopper hybrid zone

## ABSTRACT

KEYWORDS: speciation, hybridisation, Haldane’s rule, hybrid zones, phylogeography

## 1. INTRODUCTION

Over-arching question (will depend on the last results): Later stages of species formation are often characterized by hybrid sterility in the heterogametic sex (Haldane’s rule) that presumably result in strong reproductive barriers. Does sterility still allow for extensive gene flow genome wide, contributing for repeatable patterns of genomic differentiation?

Specific objectives:

1. We infer the phylogeographic history of *Chorthippus parallelus* to understand the tempo and mode of species formation.

2. We measure genetic introgression across two hybrid zones at different stages of divergence, to test if Haldane’s rule and chromosomal rearrangements function as a strong barrier to gene flow.

3. We infer patterns of heterogeneity of differentiation, to test if selection and drift act in similar genomic regions during the continuum of species formation.

## 2. MATERIALS AND METHODS

### 2.1 Sampling and sequencing

To understand the phylogeographic history of *Chorthippus parallelus*, we sampled 10 localities covering most of the known range of chromosomal races described in this species, and the hybrid zones between them (Fig. 1A; Table S1). For hybrid zones, we have sampled two parental populations within their “pure range”, and a hybrid locality near the putative centre of the hybrid zone, according to results from earlier cytogenetic studies (Flanagan et al., 1999; Zabal-Aguirre et al., 2010). For the subspecies *C. p. erythropus*, we have additionally sampled a population near its putative refugium in the Iberian Central System, and a second population in Portugal. For the subspecies *C. p. parallelus*, we have additionally sampled two populations in Austria and Slovenia that putatively originated from a range expansion from a Balkan refugium (Lunt et al., 1998). In every locality we sampled 5 male individuals in close proximity. In some cases, two nearby sub-localities were sampled (3+2 individuals) to capture a representative amount of diversity (see Table S1 for details and Accession numbers). We sampled all specimens for this study, with the exception of the reference parental population of the Pyrenean hybrid zone (PAR, ERY), published in a previous study (Nolen et al., 2020; NCBI BioProject PRJNA665162).

In the field, we preserved whole body tissue in RNAlater (Qiagen) to sample the largest variety of transcripts possible. We excluded the head and upper digestive tract to avoid gut contamination and overrepresentation of eye pigments. In the laboratory, we homogenised the samples using ceramic beads (1.4/2.8 mm, Precellys) and the standard Tri-Reagent protocol (Sigma). We resuspended and purified RNA pellets with RNAeasy Mini columns (Qiagen), followed by a quantity and integrity check using an Agilent 2100 BioAnalyser. The BGI group performed mRNA enrichment, library construction and paired-end Illumina HiSeq2500 sequencing. The three populations from the Pyrenees had a sequencing average of 59,391,201 paired reads of 150 bp, while the remaining seven populations had a sequencing average of 49,410,055 paired reads of 100 bp. Additionally, we used published data from 5 individual males of *C. biguttulus* sampled in Berlin (Germany), as an outgroup (Berdan et al., 2015; NCBI BioProject PRJNA284873).

### 2.2 Mapping and filtering

We used the BAM pipeline implemented in Paleomix (Schubert et al., 2014) for processing raw data. This process first removed adapters and trimmed low-quality bases (min. quality-offset for Phred scores 33) with AdapterRemoval (Schubert et al., 2016). Overlapping pairs were collapsed into one consensus read. All reads were then mapped with BWA-MEM, discarding poorly mapped reads under a quality threshold of 30 (Li & Durbin, 2009) against the reference transcriptome previously assembled and annotated by Nolen et al. (2020). These transcripts are organised in four groups consisting of: (1) 16,970 single-copy genes, whereof 12,735 with identified open reading frames (ORFs) and 4,235 without; (2) 4,263 multi-copy genes; (3) 18,623 singleton genes; and (4) the complete mitochondrial genome. We visualised the number of nucleotides retained and coverage in partition 1 and 4 to assess the differences in coverage across localities (Fig. S1).

As phylogenetic inference requires genotype calling, we produced haplotypes for the ORFs of the 12,735 single-copy nuclear genes and for the 13 mitochondrial genes, calling the most frequent base at each position with a minimum coverage threshold of 10×, using ANGSD v0.921 (Korneliussen et al., 2014). As most of the population genetics inference can handle uncertainty of genotypes, which is particularly important with RNAseq data where coverage varies with gene expression (da Fonseca et al., 2016), we estimated genotype likelihoods using ANGSD. For this dataset, we used the 12,735 single-copy nuclear genes with identified ORFs, excluding the first and second codon positions that often correspond to nonsynonymous substitutions and hence are more affected by selection. We excluded reads with mapping quality < 15 after adjustment, base read quality < 20, or containing multiple hits. Additionally, we only considered sites present in min. 80% of individuals, and with a coverage depth greater than twice the total number of individuals (see https://github.com/lindington/ChorthPar for the detailed commands).

### 2.3 Population Structure and Admixture

To assess population structure and identify ongoing hybridisation, we used the population genetics dataset. First, we performed a non-model principal component analysis (PCA) with ngsPopgen (Fumagalli et al., 2014) to test if genetic variation reflects the spatial distribution of the samples. We considered only variable sites with a SNP *p*-value , which has been shown to accurately reflect the site frequency spectrum based on all sites (Nolen et al., 2020). We expect most of the variance (PC1) to reflect subspecies and subsequent principal components to reflect cluster chromosomal races. Second, we estimated admixture proportions in each individual, using the clustering algorithm implemented in NGSadmix (Skotte *et al.*, 2013), which maximizes Hardy Weinberg and linkage equilibria within K ancestral clusters. We considered K = 2 to the number of sampling locations (11), with 50 replicates, selecting the replicate with the highest likelihood for each K. Under a mutation-drift equilibrium, we expect each subspecies or sampling locality to form a cluster, with possible admixture near the centre of the hybrid zones, if ongoing hybridization through backcrossing is common.

### 2.4 Mitochondrial time tree

To infer the timing of diversification of the species we estimated a mitochondrial time tree. For this, we considered only the 13 protein-coding genes to assure site homology and to avoid assembly and mapping errors in non-coding parts of the mitogenome. First, we aligned the 55 complete mitogenomes, annotated the mitogenome using MITOS (Bernt et al., 2013), and extracted the alignments for the 13 protein-coding genes using a custom script (<https://github.com/lindington/ChorthPar/>). Second, we verified the absence of stop codons along the sequence with MEGA v10.1.6 (Tamura et al., 2007), and concatenated the genes. Third, we imported the concatenated alignment into IQTree v1.6.3 (Nguyen et al., 2015), estimated the best model of evolution for each mitogene using ModelFinder (Kalyaanamoorthy et al., 2017) and assessed branch support using ultrafast bootstrapping (Hoang et al., 2018). This resulted in a consensus maximum-likelihood (ML) tree that carries uncertainty on the estimation of the topology (Fig. S2). Fourth, we produced a prior chronogram calibrating the divergence between *C. p. erythropus* and *C. p. parallelus* with a minimum age of 15 kya (last glacial maximum age) and maximum age of 33.9 mya (the oldest Acrididae fossil; Song et al., 2018), using the R package *ape* (Paradis & Schliep, 2019). Finally, we inferred the time of the most recent common ancestor (TMRCA) of all *C. parallelus*, of the main mitochondrial clades, and of the split between subspecies within each clade, using the software BEAST v1.10.4 (Suchard et al., 2018). We used the uncorrelated relaxed clock model (Drummond et al., 2006), with a normally distributed prior on the substitution rate with mean of 0.0115 mutations by million years, and standard deviation of 0.001 substitutions per million years (Brower, 1994), which is appropriate given the remarkably conserved mitochondrial rates of insects (e.g., 0.0115 in butterflies and 0.0133 in beetles; Brower, 1994; Papadopoulou et al., 2010, respectively). The tree was fixed to the previously inferred maximum likelihood topology, and we used a birth-death model (Gernhard, 2008) as speciation likelihood. To facilitate convergence, tree topology, clock model and substitution model (GTR) were linked across genes. Three independent MCMC chains were run for 100 million generations, a burn-in of 10 million, and sampling every 10,000. We checked for convergence *a posteriori* using Tracer v1.7.1 (Rambaut et al., 2018), and that all ESS values were above 200. We merged the independent chains using LogCombiner, and estimated divergence times using TreeAnnotator.

### 2.5 Nuclear species tree

To assess the evolutionary relationships between sampling populations, we inferred a nuclear species trees under the multispecies coalescent model, which accommodates incomplete lineage sorting expected for recent divergence events (Degnan & Rosenberg, 2009). First, we extracted the ORFs of the 12,735 single-copy nuclear genes to avoid biases caused by assembly errors or under-expressed areas of the transcriptome, such as UTRs, that could increase the uncertainty of the estimated gene trees. Second, we retained all genes that had more than 300 calls, and that were present in a minimum of 3 individuals in each population. Third, we inferred a ML tree for every nuclear gene using IQTree v1.6.3 (Nguyen et al., 2015) to infer a ML tree for every filtered gene. For this, we used the BIC-selected models, 1,000 replicates of ultrafast bootstrapping (UFBoot), and SH-like approximate likelihood ratio tests (SH-aLRT; Guindon et al., 2010; Hoang et al., 2018). ). To account for the uncertainty in gene tree estimation, we collapsed branches with <50% bootstrap support. Fourth, we assess the information content of each nuclear gene tree and compared to the mitochondrial tree (Strimmer & Haeseler, 1997). performing a likelihood mapping test in IQTree. Finally, we estimated the species tree that maximizes the topology agreement between independent nuclear gene trees (Mirarab & Warnow, 2015), using the method implemented in ASTRAL v5.6.1 (Zhang et al., 2018). From this analysis, we output the main species tree topology, posterior probabilities for this main topology, and branch lengths in coalescent units (*T/Ne*; Degnan & Rosenberg, 2009). We used this approach to estimate an “individual species tree” where each individual is a terminal branch, and a “population species tree” where individuals sampled in the same population are merged (as in Table S1). Given the taxonomic consensus, we expect the subspecies form two reciprocally monophyletic groups with hybrid individuals or populations having an intermediate position.

### 2.6 Gene flow between populations

To test for the presence and magnitude of gene flow between all sampled populations, we used the Patterson's *D*-statistics (Durand et al., 2011). This test weights the fraction of biallelic sites that have a different topology from the previously estimated species tree. Using *C. biguttulus* as the outgroup carrying the ancestral A allele, we measured the proportion of ABBA and BABA sites for all possible combinations of three *C. parallelus* taxa that conformed to our estimated species tree (120 comparisons). We used Abbababa2 as implemented in ANGSD, which extends this analysis to comparisons among populations (Soraggi et al., 2018). We restricted the analysis to sites with a SNP *p*-value and estimated significance using a *p*-value calculated from 10 windows of ~240,000 relevant sites. A similar proportion of ABBA and BABA (D=0) sites is expected under the null hypothesis of incomplete lineage sorting driving discordance, while significantly different proportions (D<0; D>0 with p-value < 0.05) must be explained by gene flow between two populations. We visualized Patterson's *D*-statistics using a matrix of all pairwise comparisons, using the highest reached D-value for each comparison. If backcrossing in hybrid zones lead to introgression into parental populations, we expect that populations close to the hybrid zones to have higher *D*-statistics relative to more distant populations.

### 2.7 The demographic history of hybrid zones

To position the Pyrenean and the Alpine hybrid zone along the continuum of species formation, we estimated the demographic history of divergence of parental taxa at each zone separately (ERY and PAR for the Pyrenees; TAR and GOM for the Alps). We considered four nested demographic models that potentially describe the demographic history of the hybrid zones with a minimum number of parameters: (1) divergence without gene flow (“No Migration”, with three parameters: time since population split *T1*, and effective population size *NE*for each population); (2) divergence with continuous gene flow (“Migration”, with a fourth migration rate parameter *m*); (3) gene flow after secondary contact (“Secondary Contact”, with the fifth parameter time since secondary contact *T2*, after which migration starts); and (4) divergence with asymmetric gene flow after secondary contact (“Secondary Contact with Asymmetric Migration”, with the sixth *m2*parameter). First, we used ANGSD to build two-dimensional site frequency spectra (2D-SFS) for each of the 16,969 single copy nuclear genes and summed them into a single complete 2D-SFS per population pair. Second, the complete 2D-SFS was fit to all four demographic models using the diffusion approximation methods implemented in δaδi v2.0.5 (Gutenkunst et al., 2009), as described in Nolen et al. (2020; see https://github.com/zjnolen/chorthippus\_radiation for commands), with four technical replicates to guarantee convergence of the estimated parameter values. Finally, we compared nested models using a likelihood ratio test (LRT), and estimated parameter uncertainties with the Godambe Information Matrix (Coffman et al., 2016; Godambe, 1960), which uses 100 nonparametric bootstrap SFS to account with physical linkage of SNPs within the same transcript. This procedure selects the simplest model the captures most of the demographic signal contained in the 2D-SFS, and thus we chose a model that is highly supported in both hybrid zones in order to compare demographic parameters. The demographic parameters *T* and *2Nm* were estimated in reference to the constant mutation rate *μ*, as μ is probably the same for such closely related taxa. Because the subspecies hybridising in the Pyrenees are more diverged than the subpopulations hybridising in the Alps, we expect parameter estimations to reflect this with higher values of *2Nm* and more recent divergence times *T* estimated for the population pair in the Alps.

### 2.8 Heterogeneity of Gene Flow

To estimate the relative effective population size for each sampling locality, we calculated per-gene Watterson’s θ (Watterson, 1975) and π (Nei & Li, 1979). We also estimated deviation from a neutral model of constant population size by calculating per-gene Tajima’s D (Tajima, 1989). For each summary statistics, we first built the one-dimensional (1D-) SFS for each population and then calculated θ, π, and Tajima's D for each site in ANGSD. We used a custom script to combine values across linked sites within each gene and plotted their distributions for the genes sampled across all populations (see https://github.com/lindington/ChorthPar). We expect to find an excess of singletons (Tajima’s D < 0) in populations experiencing range expansions and a Tajima’s D around 0 in demographically stable populations located near refugia.

To test if the same gene show highest differentiation across the two hybrid zones, we estimated pairwise FST (Reynolds et al., 1983) between pairs of parental populations (ERY vs PAR in the Pyrenees, and TAR vs GOM in the Alps). We used ANGSD to build a 2D-SFS per population pair and to calculate per-gene FST for each comparison. We plotted the distributions for genes in the two different comparisons. Using Fisher’s exact test (Fisher, 1922) as implemented in the R package GeneOverlap (Shen, 2020), we tested if there is significant overlap between genes with the highest 5% FST in the two hybrid zones. Because high FST can be caused by selection or low recombination counteracting gene flow across the two hybrid zones (Nachman & Payseur, 2012), we estimated dXY (Nei & Li, 1979). We used ANGSD and a script adapted from Peñalba (2017; https://github.com/mfumagalli/ngsPopGen/blob/master/scripts/calcDxy.R) to estimate per-site dXY values and averaged them across sites within each gene to obtain per-gene dXY estimates. If high FST is caused by selection counteracting gene flow, we expect genes with high FST to exhibit high dXY estimates, while if high FST is caused by low recombination counteracting gene flow, we expect genes with high FST to exhibit low dXY estimates.

## 3. RESULTS

### 3.1 Mapping and filtering

Overall, trimming and mapping results were equally successful across populations (Table S2). The percentage of retained nucleotides after filtering ranged from 79% to 89% (Figure S1). The mapping was also effective for all individuals, ranging from 84% to 90% of the reads. Regarding coverage, the mitochondrial genome (chr4), had a coverage from 16,000 to 52,000 X, while the single-copy genes with identified ORF presented a coverage from 46 to 125 X (Figure S1). As the samples from the Pyrenees were sequenced with more and longer reads, they are the individuals that present the higher number of retained nucleotides and coverage of nuclear genes.

### 3.2 Population Structure

A large fraction of genetic variance (11.9 %) is explained by PC1, which separates individuals from the subspecies *C. p. erythropus* from the Iberian Peninsula, from individuals from the subspecies *C. p. parallelus* (Fig. 1B). PC2 explains 3.85% of the variance and distinguishes individuals within subspecies along a north-south gradient. PC3 and PC4 explain 3% and 2.8% of the variance, respectively, and distinguishes the Balkan samples from all others (Fig. S3).

Results from our admixture analysis show a continuous increasing likelihood with increasing K (Table S3). The first split within *C. parallelus* distinguishes the two subspecies, with some individuals from the Pyrenean hybrid zone showing admixture (Fig. S4). Next, the European and Italian chromosomal races of *C. p. parallelus* separate with admixture found in the Alpine hybrid zone and the two locations closer to the Balkans (DOB and SLO). The Austrian population (DOB) becomes distinct at K=5 and does not show admixture with any of the neighbouring localities. At K=11 each sampling locality forms its own cluster, with admixture found only in two individuals collected closer to the reference population of pure *C. p. erythropus* (Table S1).

### 3.2 Mitochondrial time tree (Enrique)

The maximum likelihood mitochondrial tree (Fig. S2) shows a well-supported topology (BP > 95), where the two subspecies and most of the populations are not reciprocally monophyletic. The mitochondrial lineages form six major clades (A-F), most of them containing lineages found in a single subspecies. The exception are clades A and F, which contain internal nodes that separate *C. p. parallelus* and *C. p. erythropus*. We used those internal nodes to estimate the lower boundary for the divergence time between subspecies (Fig. 2).

The TMRCA of all *C. parallelus* is estimated to be around 380 ka (95% HPD 432-322). The diversification of the six major clades ranges from 388 ka to 218 ka (Table S4). The lowest boundary of divergence between subspecies is estimated to be around 100 ka (128-66) in both internal clades.

### **3.3 Nuclear species tree** (Enrique)

After filtering, we obtained 5,929 genes that were used for gene trees and species trees estimation. The likelihood mapping analysis shows that most gene trees have enough power to resolve 60-80% of the quartets, reflecting a high informativity of nuclear gene trees, yet below that of the mitochondrial tree (Fig. S5). The individual gene trees show a wide range of sorting between alleles from each subspecies, from high incomplete lineage sorting similar to what was observed in the mitochondrial genome, to two reciprocally monophyletic clades (Fig. S6).

The individual species tree (Figure S7) shows that *C. p. erythropus* and *C. p. parallelus* are reciprocally monophyletic. In addition, the majority of the populations were monophyletic. Within the *erythropus* clade, individuals cluster in three major subclades: Portugal, Central Spain and Spanish Pyrenees. The two sampling localities of the Pyrenees *erytrophus* population (ERY) were reciprocally monophyletic and clustered together (PP = 1). However, the sampling localities of the hybrid zone (PHZ), were not reciprocally monophyletic and did not cluster together. One of the localities was sister of the Spanish Pyrenees (ERY) (PP = 1), while the other is at the base of the clade, suggesting that individuals from this locality experience gene flow. On the *parallelus* clade, individuals clustered geographically in three subclades. 1) The oldest split within the *parallelus* clade separates the East Austrian individuals (DOB) from the remaining *parallelus*. 2) the Italian (GOM), Slovenian (SLO) individuals together with the Alpine hybrid zone (AHZ) individuals. 3) the West Austrian (TAR) and French (PAR) individuals. Regarding the French localities, one did not form a monophyletic group while individuals from the other locality clustered together and formed a monophyletic group (PP = 1).

Accordingly, the “population species tree” (Fig. 3A) shows that *C. p. erythropus* and *C. p. parallelus* are reciprocally monophyletic (posterior probability, PP > 0.95). The Pyrenean hybrid zone (PHZ) is at the base of the *erythropus* clade, as expected if there is introgression of *parallelus* alleles into this population. The reference population of *erythropus* outside of the Pyrenean hybrid zone (ERY) is sister of the population located in the putative glacial refugium at the Iberean Central System (CSY). Within the *parallelus* clade we obtained the same topology as the individual species tree: 1) the very distinct East Austrian population (DOB) 2) a clade of the populations to the South of the Alps, Italy (GOM) and Slovenia (SLO), together with the Alpine hybrid zone (AHZ). The Slovenian population is at the base at this subclade, also consistent with introgression from other localities. 3) a clade of the populations to the North of the Alps, West Austria (TAR) and France (PAR).

### 3.4 Gene flow between populations

We have estimated Patterson's *D*-statistics as a proxy for gene flow in 115 population comparisons. Out of these, 121 significantly deviate from the null hypothesis of no gene flow (p-values < 0.05; Table S5), consistent with gene flow between 34 out of 42 possible population pairs (Fig. 3). All populations experience gene flow with at least one non-sister taxon. Notably, populations closer to the Pyrenean hybrid zone show the highest values of *D*-statistics across comparisons, i.e., the reference *parallelus* population in France (PAR) consistently shows high gene flow with all populations of *erythropus* from Iberia, and the Pyrenean hybrid population (PHZ) consistently shows high gene flow with all populations of *parallelus*. The Slovenian population (SLO) also shows high values of *D*-statistics with the Italian population (GOM and AHZ), also consistent with gene flow.

### 3.5 The demographic history of hybrid zones

In our demographic models, all technical replicates converged to similar likelihoods and parameter estimations. The most likely model is the most parameter rich, i.e., the “SC with Asymmetric Migration” model. Yet, when penalizing the likelihoods for an increased number of parameters using the adjusted likelihood ratio test (LRT), we find that the simplest model that can explain the observed SFS is the “Symmetric Migration” model for the Pyrenean hybrid zone, and in the “Secondary Contact” model for the Alpine hybrid zone. This suggests that our data does not have enough power to estimate complex demographic models with many parameters.

To allow comparisons across hybrid zones, we thus chose the simpler “Symmetric Migration” model for parameter estimation. Estimated *NE* are similar in the Alps, but in the Pyrenees the size difference is roughly two-fold. As expected, the parameter estimation converged on more recent divergence time and higher *2Nm* in the Alpine hybrid zone relative to the Pyrenean hybrid zone (Table 1).

### 3.6 Heterogeneity of Gene Flow

Distributions of per-gene π were similar across populations, with most genes showing values close to zero, and populations differing mainly in the length of the tail. We used the mean to summarise the tail length and found that the hybrid zone populations have rather high mean per-gene π, suggesting that they have larger effective population sizes relative to the other populations (Fig. 4A). The Pyrenean hybrid zone PHZ shows an intermediate mean compared to its parental populations ERY and PAR, while the Alpine hybrid population AHZ has a higher mean than both its parental populations TAR and GOM. Accordingly, these populations also have the high means of per-gene Watterson’s θ, suggesting that they have high mean genetic diversity compared to the other populations (Fig. S8).

Per-gene Tajima’s D distributions were generally negative in all populations, with tails spanning between -1 and 1 (Fig. 4B). We used the mean to summarise trends in the tails. Populations located in proximity to putative refugia, (CSY and ERY near the Iberian refugium, and GOM near the Italian refugium) have means closest to 0, suggesting no deviation from neutrality and thus demographic stability. Notably, in the Austrian population close to the Balkans (DOB), we find a mean close to 0, conforming to expectations of demographic stability found in populations near putative refugia. We found the most negative means of Tajima’s D in populations separated from the refugia by a geographic barrier (POR by the Central System; PAR by the Pyrenees and Alps), suggesting that they experienced a range expansion. Notably, the Slovenian population SLO has a low mean of Tajima’s D, conforming to expectations of demographic range expansion found in regions that would have been uninhabitable for *C. parallelus* during glacial maxima. The hybrid zone populations PHZ and AHZ have means of Tajima’s D that are intermediate compared to their parent populations (ERY and PAR, and TAR and GOM, respectively).

The per-gene FST distributions are unimodal with a positive skew. The distributions significantly differ between hybrid zones with a higher mean and less positive skew in the Pyrenees (Fig. 5A). Genes in the highest 5% per-gene FST have values between 0.62938 and 1, and between 0.7375 and 1 in the Alps and Pyrenees, respectively. Out of 575 genes in each hybrid zones’ highest 5%, 117 overlap. Using Fisher’s exact test, we find overlap of genes in the highest 5% FST between hybrid zones to be significant, with a p-value of 1.1e-41. Per-gene dXY distributions of both hybrid zones are unimodal and strongly positively skewed, and very similar in the two hybrid zones. We found a significant negative correlation (slope = -0.0150551; p-value < 2e-16) between dXY (response variable) and FST (predictor variable; Fig. 5B) in both hybrid zones, suggesting that high FST values are driven by other factors than gene flow, such as linked selection due to low recombination.

## 4. DISCUSSION

**4.1 Population structure, admixture, and topology of *C. parallelus* (test the phylogeographic history of speciation)**

We find that the time of split of *C. p. erythropus* and *C. p. parallelus* coincides with the last glacial maximum. Our mitochondrial time tree suggests multiple cycles of allopatry, which may correspond to glacial cycles. The order of splitting events within *C. parallelus* first separates the two subspecies, and then notably the East Austrian population DOB from all other *C. p. parallelus* populations, before chromosomal races and sampling localities split. This is supported by our admixture analysis, in which the East Austrian population DOB shows no admixture and forms a cluster at K=5. Together, these results establish the phylogeographic history under multiple cycles of allopatry of *C. parallelus*.

These findings are supported by results from our PCA and our estimations of Tajima’s D. Most of the genetic variance in our samples is explained by subspecies and the second largest component reflects north-south gradient (Fig. 1). Individuals from the same localities cluster together, suggesting that geographic distance is the driving factor of variance within subspecies. Estimated means of Tajima’s D, which can be interpreted as a proxy for demographic range expansion or bottleneck (ref), correspond to signals of demographic stability in the East Austrian population (DOB). This, together with the established phylogeographic history, suggests its location may have been a microrefugium during at least one of the glacial cycles, explaining its early split from the *parallelus* clade.

**4.2 Gene Flow between *C. parallelus* subspecies and populations (test if Haldane’s rule is a strong barrier to gene flow vs late stages)**

Using the ABBA-BABA test, we find that gene flow is pervasive across all populations, and using demographic modelling, that gene flow occurs across both hybrid zones, confirming that full reproductive isolation has not been achieved (Fig. 4; Table S6). Together, these results show that despite hybrid male sterility (Haldane’s rule), introgression can occur across hybrid zones.

In results from our admixture analysis, we find support for hierarchical population structure, as subspecies, chromosomal races and hybrid zones, and then sampling localities cluster together at increasing Ks (Fig S3). However, signals of ubiquitous gene flow may be “absorbed” by the allelic frequencies of the populations, if all individuals within it share the same fraction of admixture. This is because allele frequencies of the inferred ancestral population are shifted to reflect the proportion of admixture that is shared by all individuals (Lawson et al., 2018). Our time tree estimations reveal 6 major mitochondrial clades that split earlier than the two subspecies, leading to extensive ILS of mitochondrial genes. We find strong discordance among mitochondrial and nuclear gene trees, which can be caused both by gene flow and/or ILS. While terminal branches present higher discordance than internal ones (suggesting ILS), the hybrid zone population PHZ is the most distinct within the *erythropus* clade (Fig), which is likely due to gene flow from the *parallelus* clade. We found, using the ABBA-BABA-test, that much of the gene tree discordance cannot be explained by ILS alone, and that high levels of gene flow are found even between population pairs that are geographically distant (Fig. 4, Table S1). Notably, populations in or bordering the Pyrenean hybrid zone (PHZ, PAR) showed high levels of gene flow with all populations of their reciprocal sister-subspecies, suggesting that this hybrid zone acts as a passage for introgression. Our demographic model’s estimated values for *2Nm* show that while gene flow is more pervasive across the Alpine hybrid zone, all populations in both hybrid zones have values of population migration rates < 1, suggesting that distinct populations will be maintained.

Together, these results suggest that Haldane’s rule is not a strong enough barrier to gene flow for speciation to occur in *C**. parallelus*. But that gene flow, if maintained at this level, cannot completely homogenise the distinct populations.

**4.3** **Heterogeneity of within and between population diversity (test if selection and drift act in similar genomic regions)**

Selection vs Drift -> selection

We find signatures of stable demographic ranges in (CSY, GOM) while (PAR, POR, SLO) exhibit signals of range expansions (Fig.). This is consistent both with locations of known refugia during the last glacial maximum and with the fact that northern slopes would have been uninhabitable for *C. parallelus* during the last glacial maximum. Notably, our findings that DOB exhibits signals consistent with other known refugia (Fig.). and that this clade is deeply divergent (Fig.), suggest that this location may have acted as a cryptic microrefugium for *C. parallelus* during the Pleistocene.

## 5. ACKNOWLEDGEMENTS

## 6. REFERENCES

## 7. FIGURES

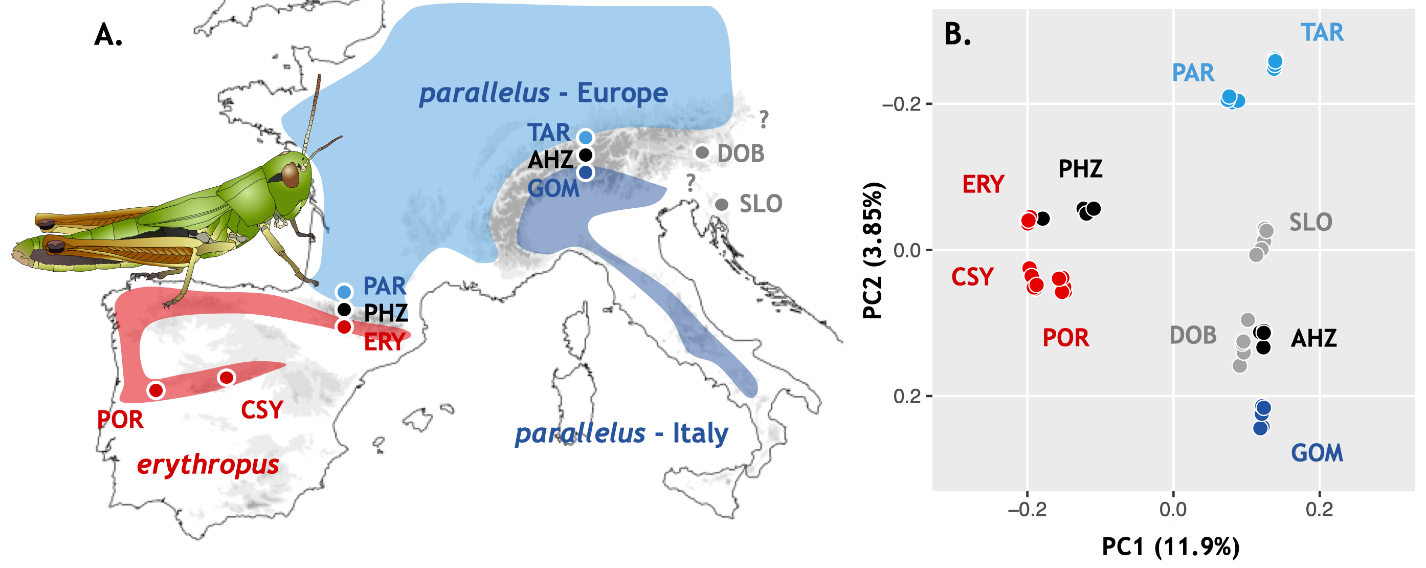
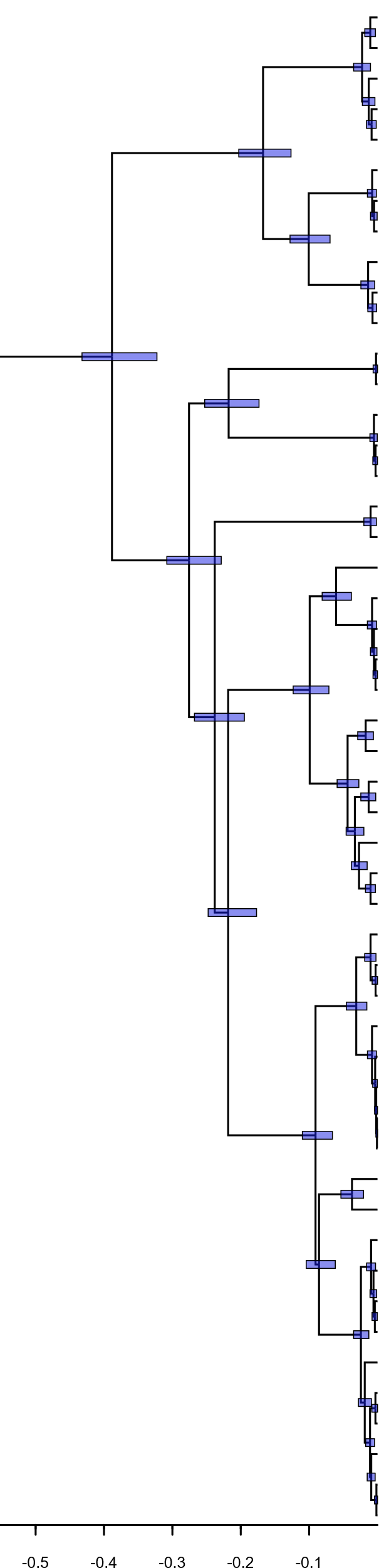
Figures: 

Fig. 1. Sampling and PCA

Fig. 2. Phylogenetics: Time tree and Nuclear species tree



A

B

C

F

D

E

mya

par-ery A

par-ery F

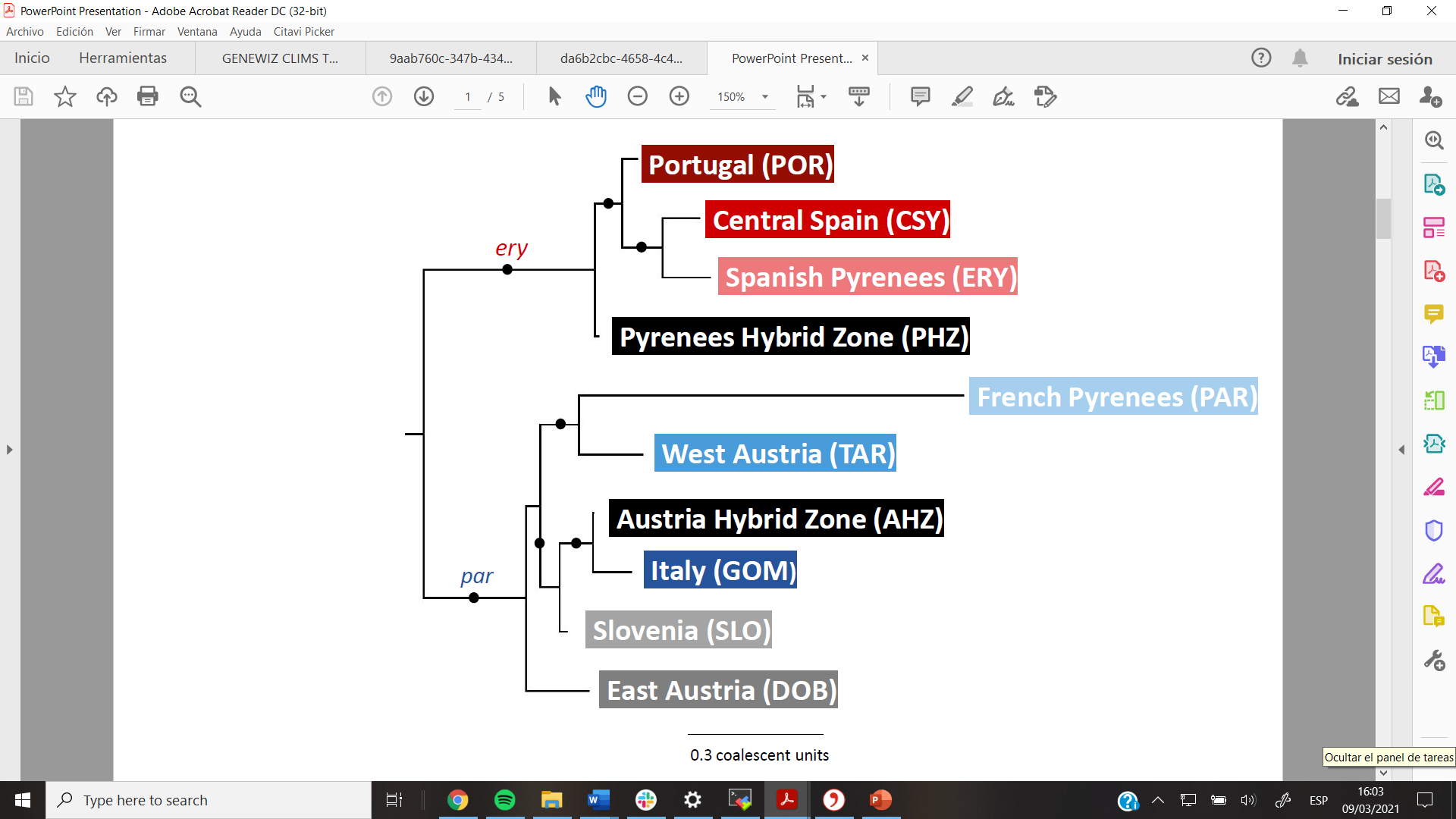


Fig. 3 Gene flow: heat map ABBA-BABA

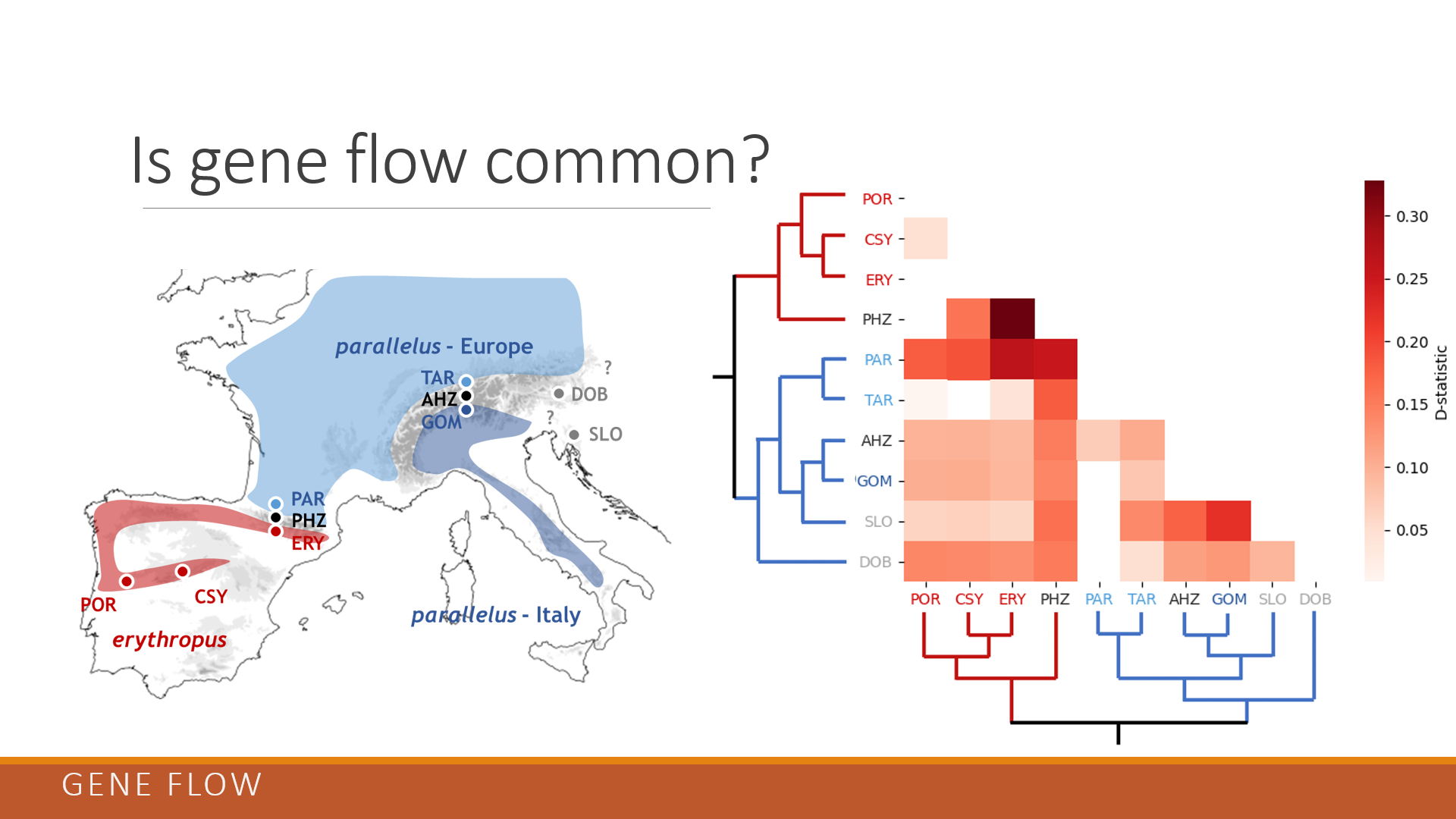
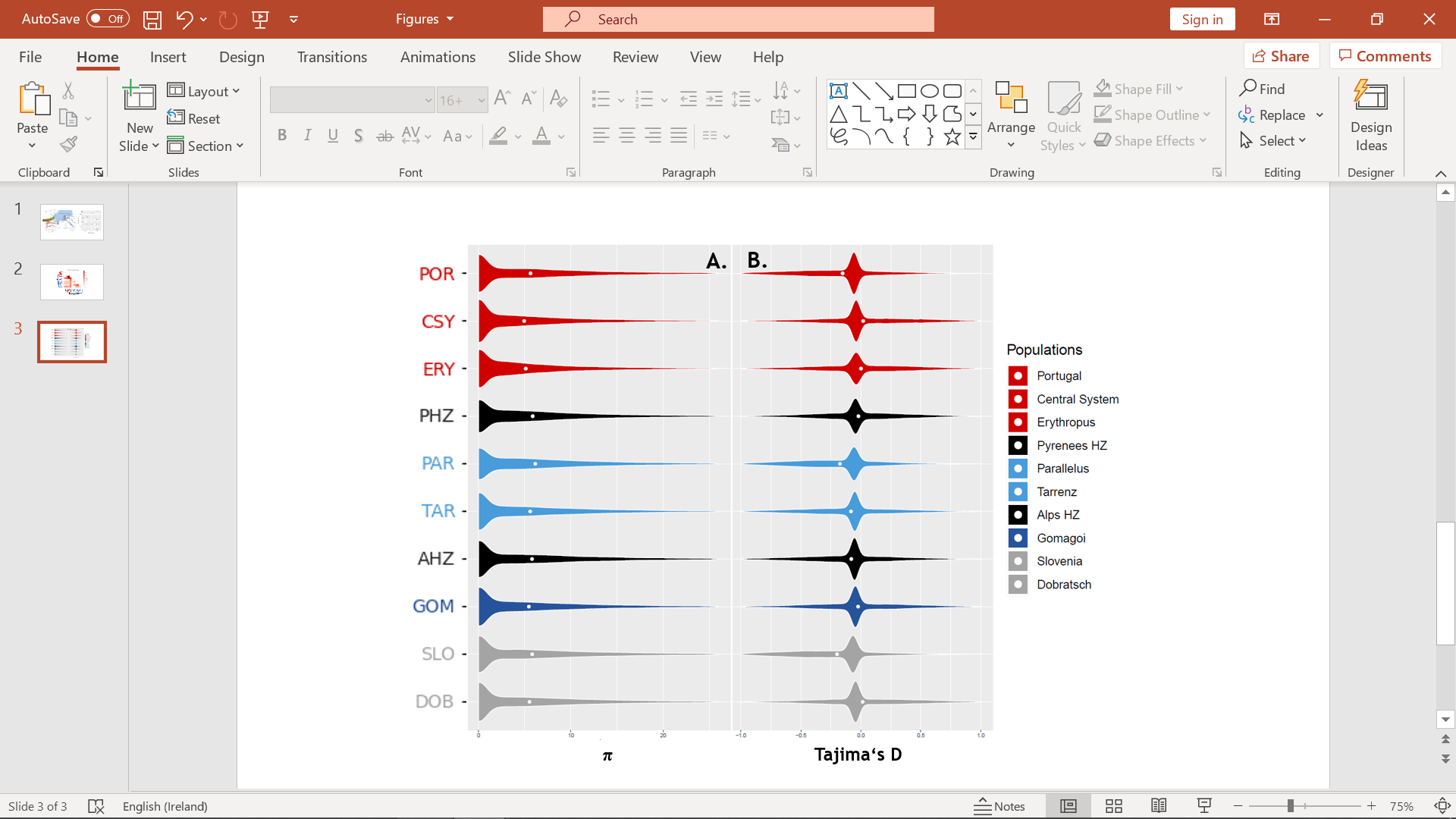
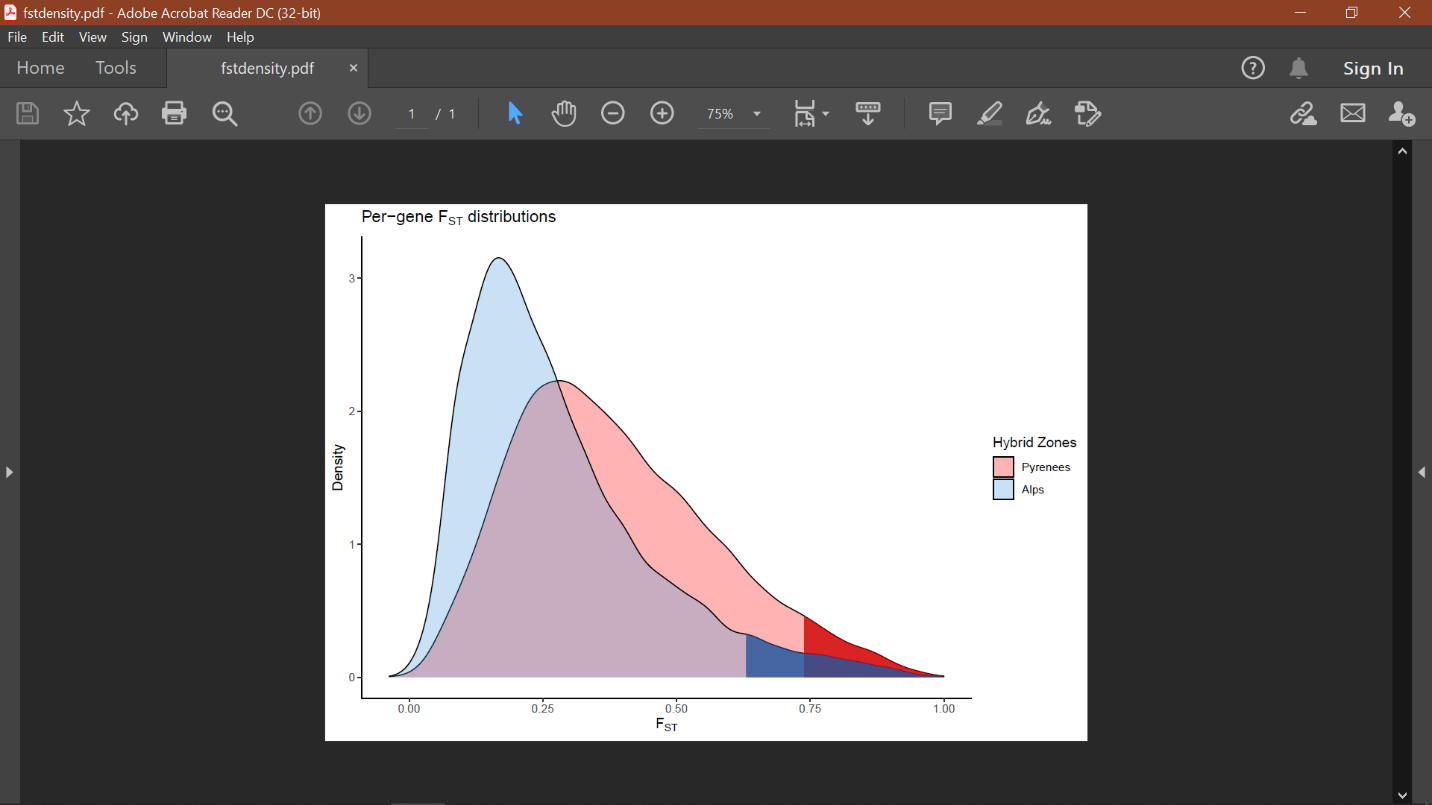


Table 1. dadi results

Fig. 4. Per-gene pi (A) & tajima’s D (B), Fst density plot





## 7. SUPPLEMENTARY MATERIAL

Table S1: Node heights and intervals of confidence of the relevant nodes, chronologically ordered.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Node** | **Age (Brower rate)** | | **Sup. 95 (Brower rate)** | **Inf. 95 (Brower rate)** | **Age (Papadopoulou rate)** | **Sup. 95 (Papadopoulou rate)** | **Inf. 95 (Papadopoulou rate)** |
| **Parallelus / A** | | 0.3883 | 0.4322 | 0.3226 | 0.3547 | 0.3887 | 0.2922 |
| **B / C** | 0.2757 | | 0.3082 | 0.2285 | 0.2521 | 0.2766 | 0.2058 |
| **D** | 0.238 | | 0.2677 | 0.1948 | 0.2175 | 0.2411 | 0.176 |
| **E / F** | 0.2184 | | 0.2477 | 0.1769 | 0.1995 | 0.1602 | 0.2235 |
| **hyb A (AHZ)** | 0.1674 | | 0.2029 | 0.1265 | 0.1528 | 0.1814 | 0.1132 |
| **par-ery A** | 0.1005 | | 0.1279 | 0.0694 | 0.0919 | 0.1153 | 0.063 |
| **par – E - east**  **(DOB)** | 0.0992 | | 0.1234 | 0.0711 | 0.0906 | 0.111 | 0.0641 |
| **par-ery F** | 0.0907 | | 0.1099 | 0.0659 | 0.0824 | 0.0983 | 0.059 |
| **Portugal (POR)** | 0.0311 | | 0.0456 | 0.0156 | 0,0283 | 0.041 | 0.0141 |
| **Italy (GOM)** | 0.0109 | | 0.0173 | 0.0043 | 0.01 | 0.0156 | 0.004 |

Figure S1: Data quality assessment

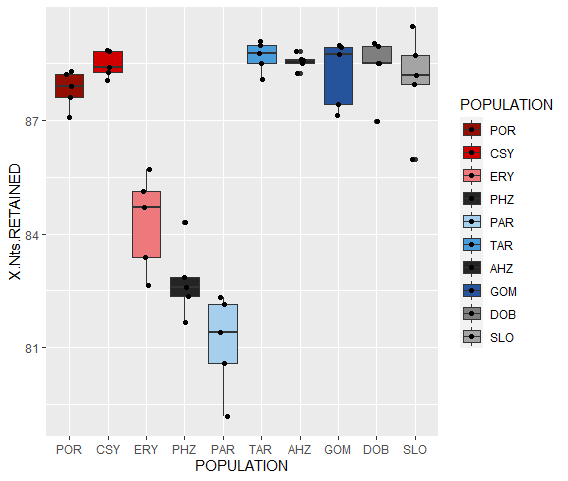
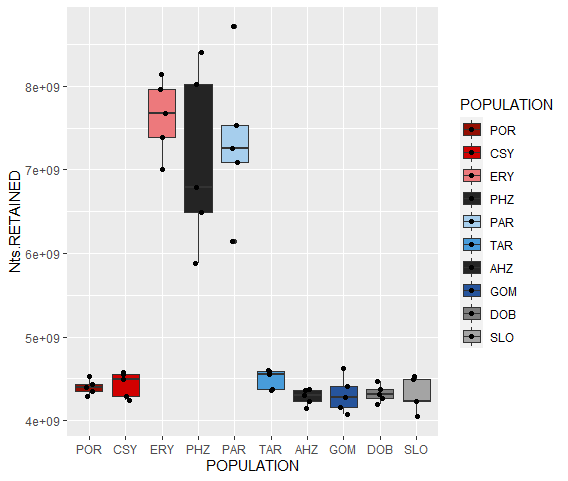
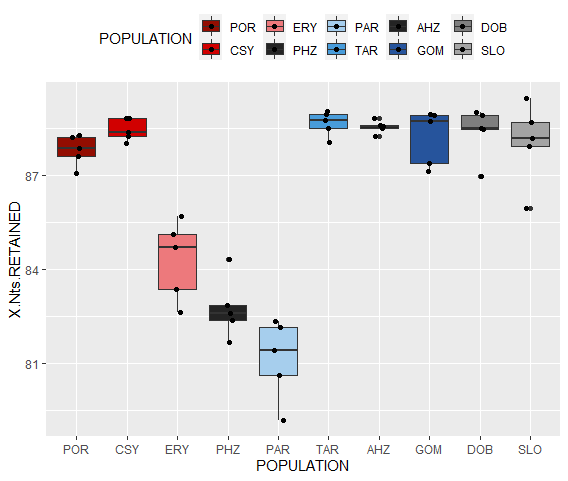
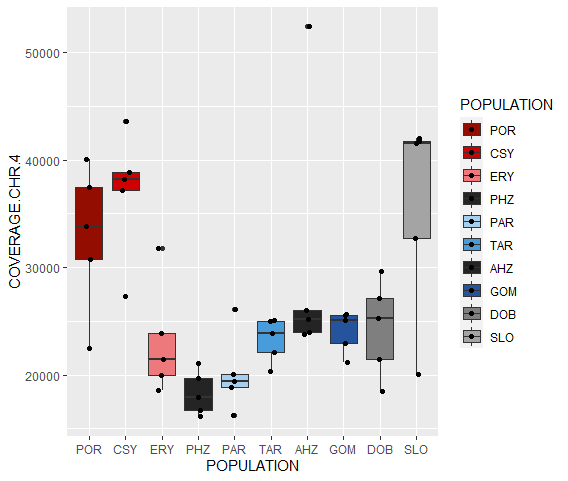
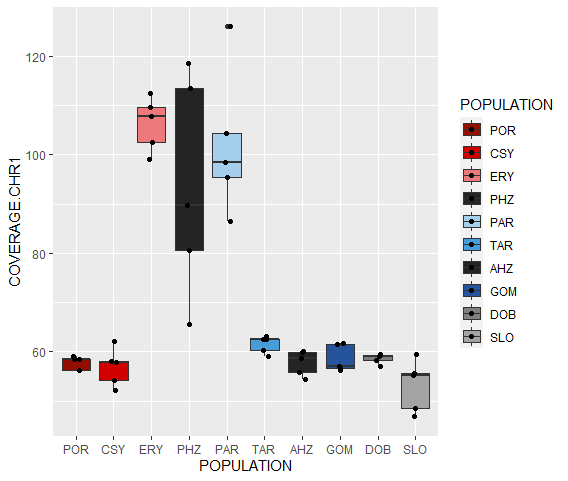


Figure S2: PCA Components 3&4

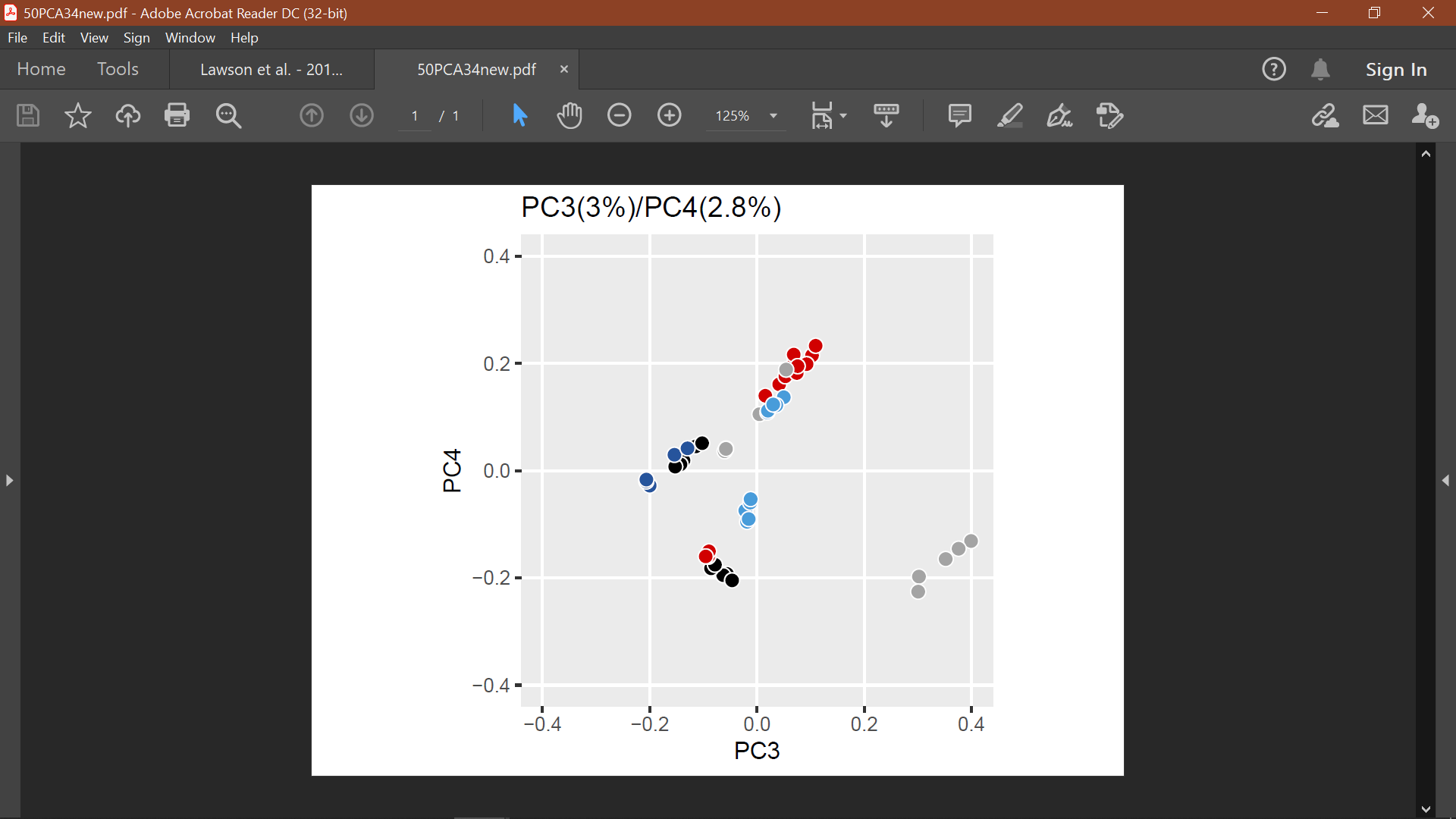


Figure S3: Population structure from K=2 to K=11

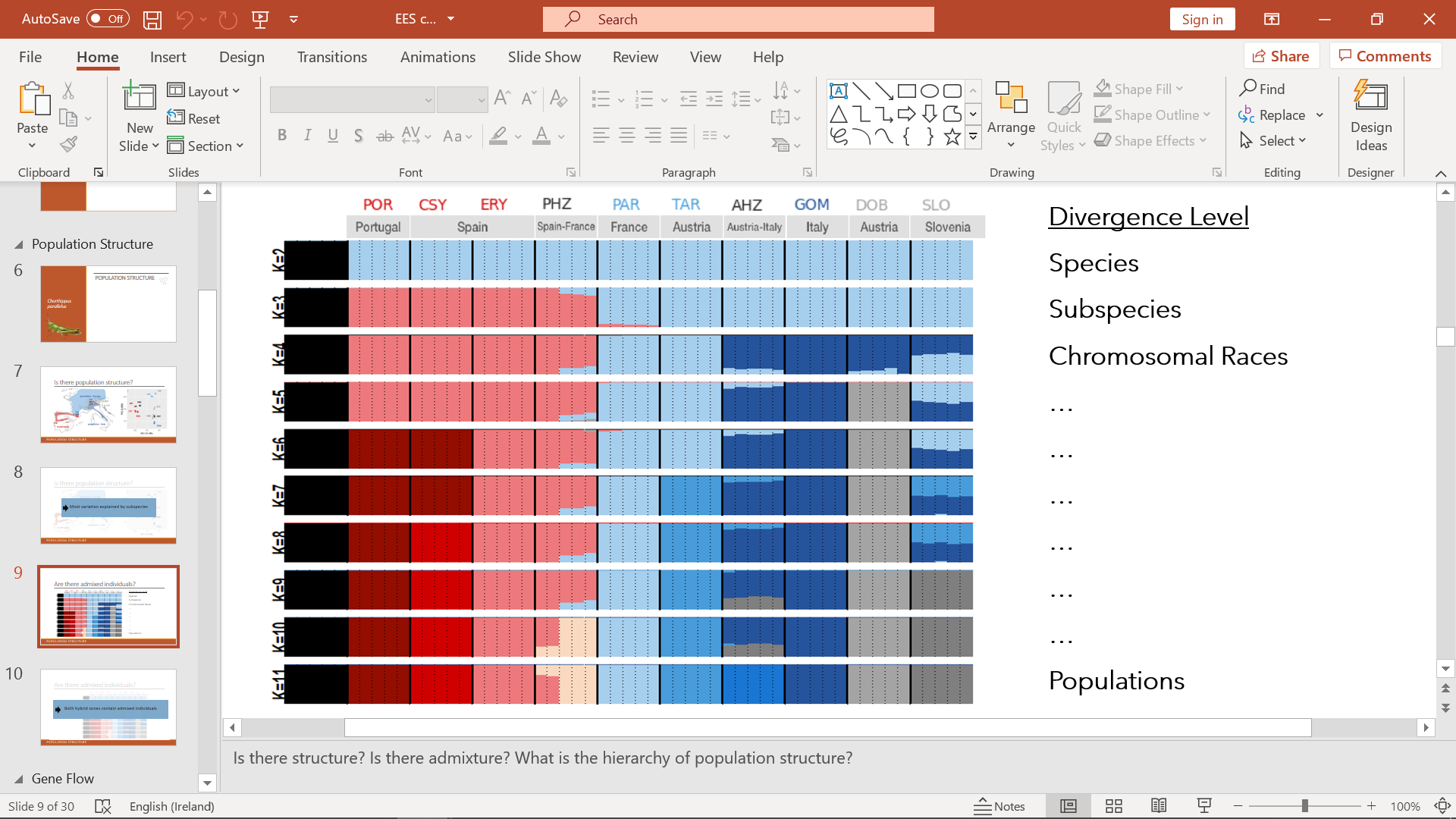


Figure S4: Maximum likelihood mitochondrial tree

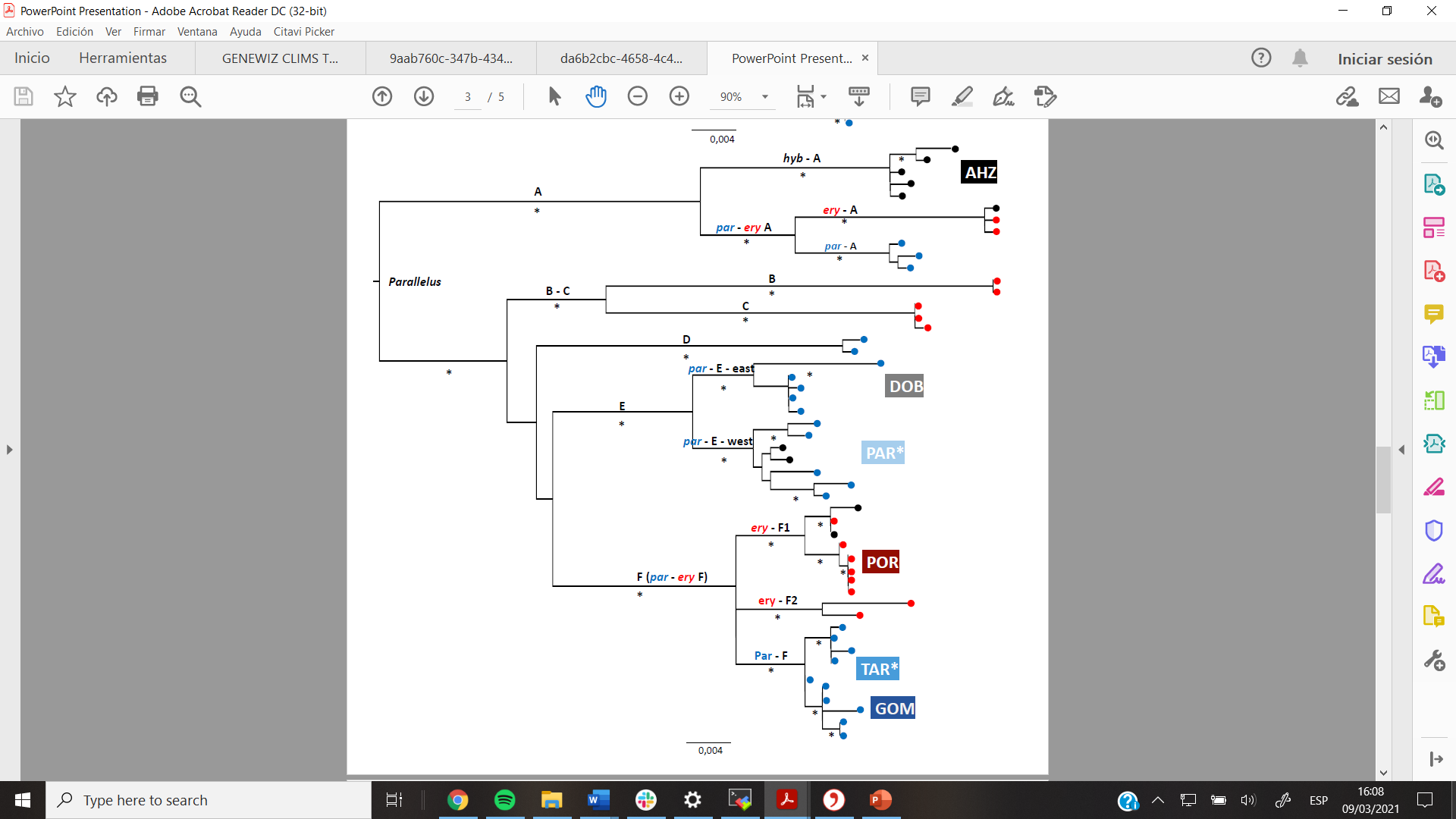


Figure S5: Ternary plots with mapped dots showing the percentages of fully, partly and unresolved quartets for each gene alignment. Red dot represents mitogenome while black dots represent single nuclear gene trees

*A close up of a map

Description automatically generated*

Figure S6: Species tree at locality level

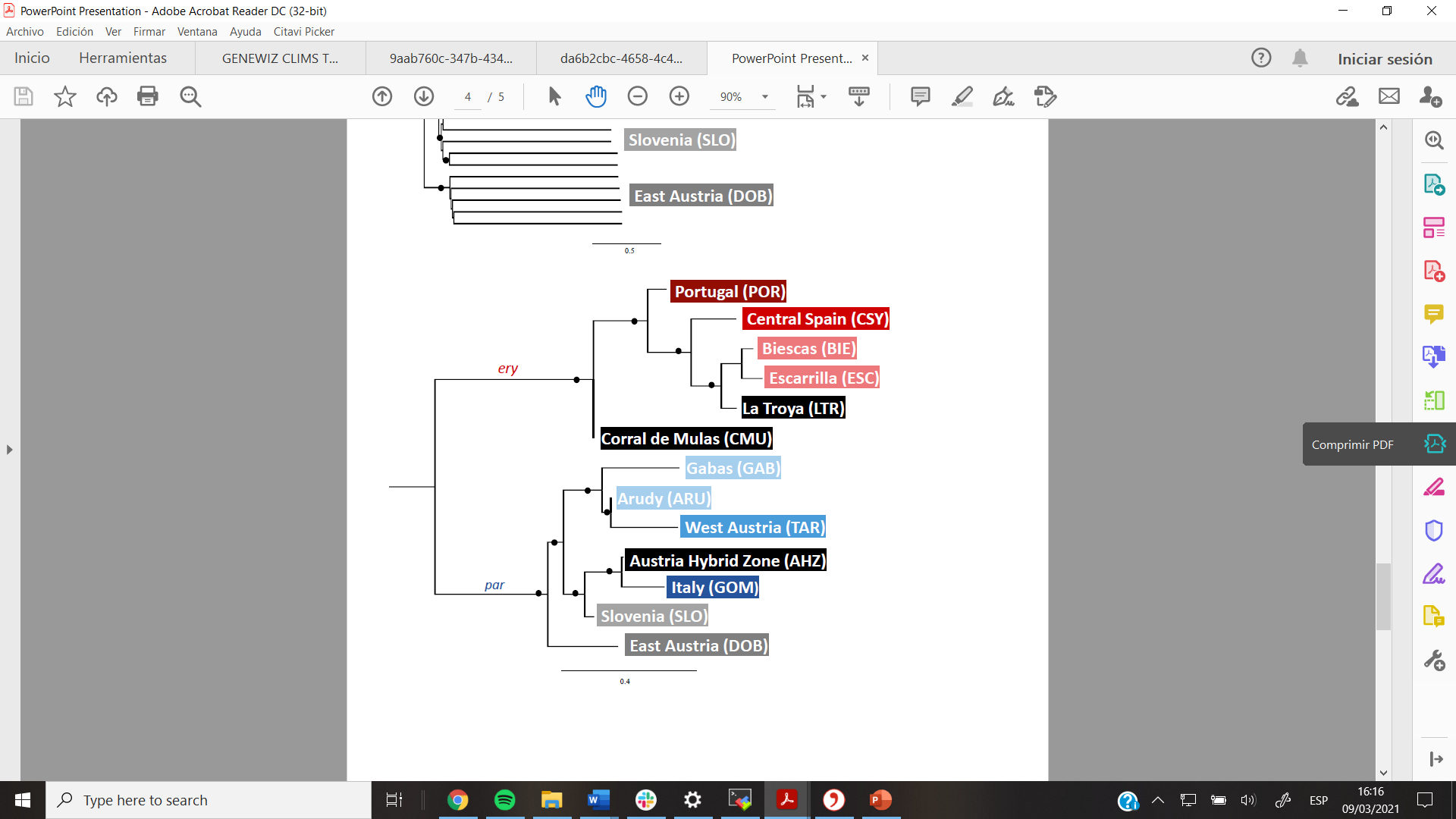


Figure S7: Species tree at individual level

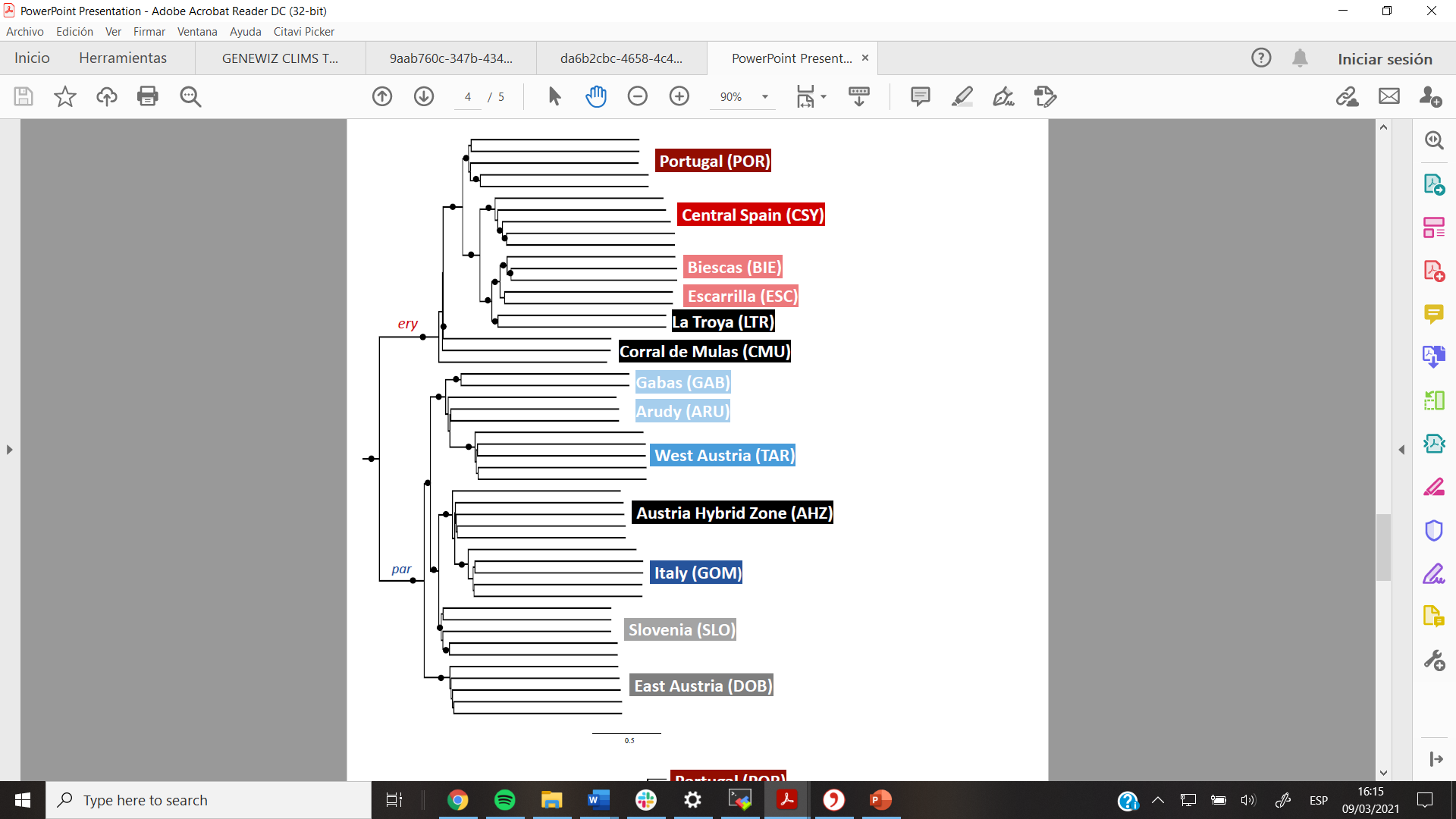
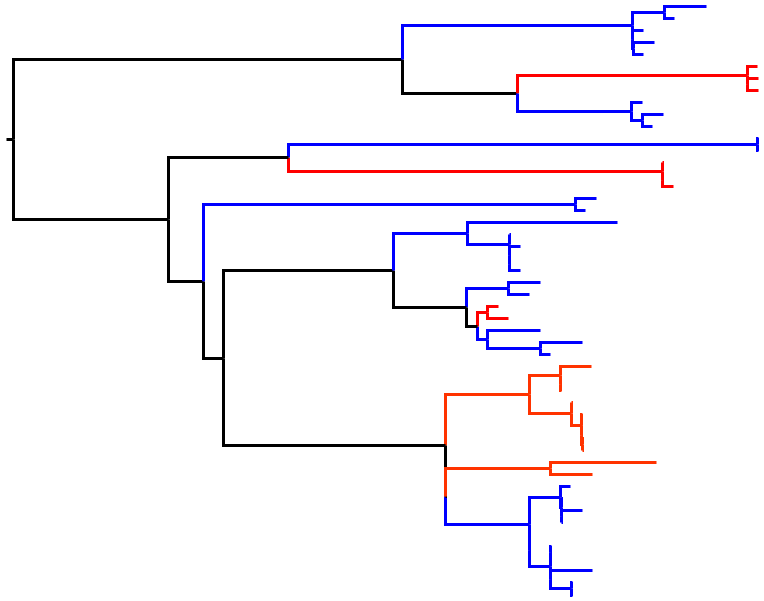
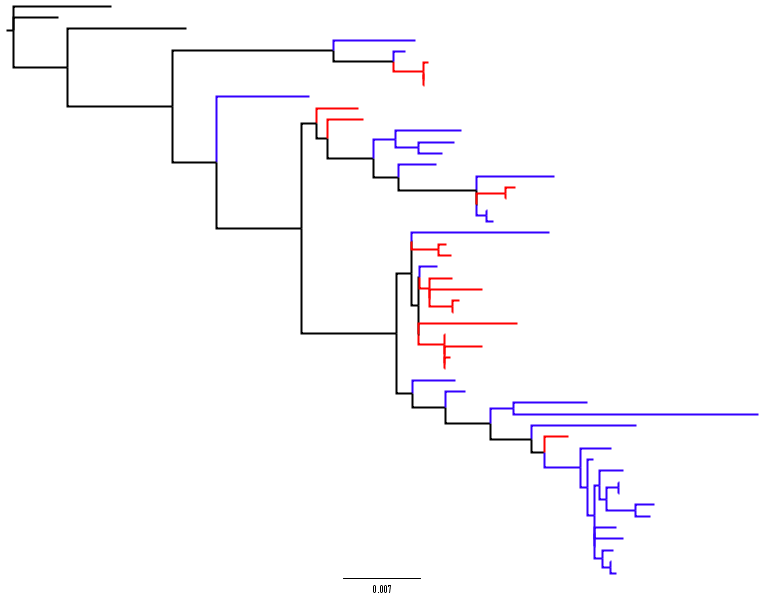
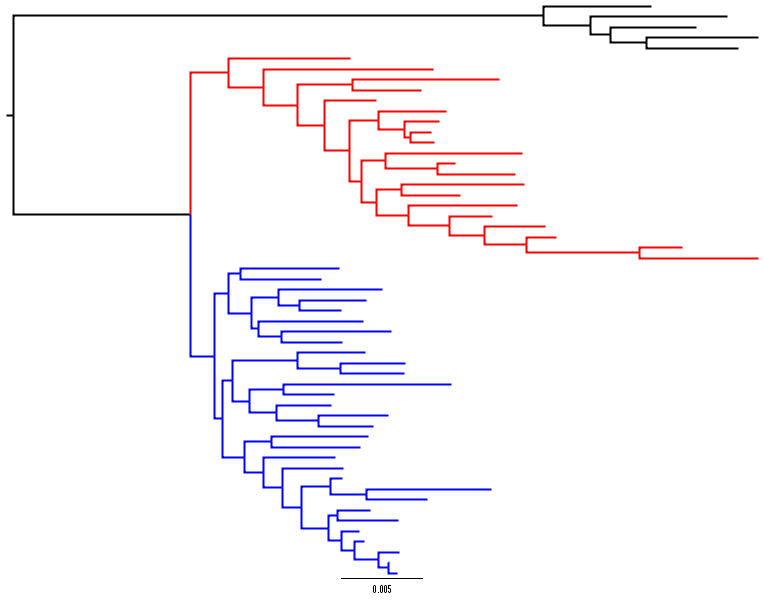
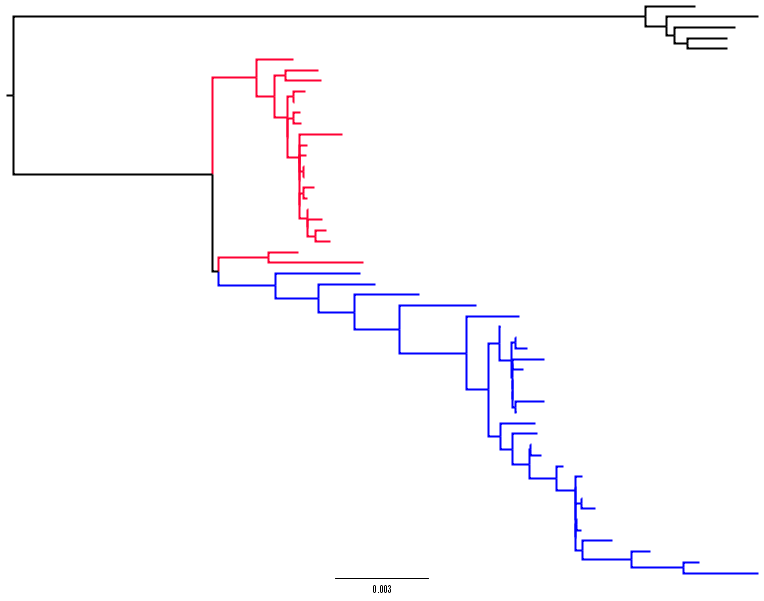
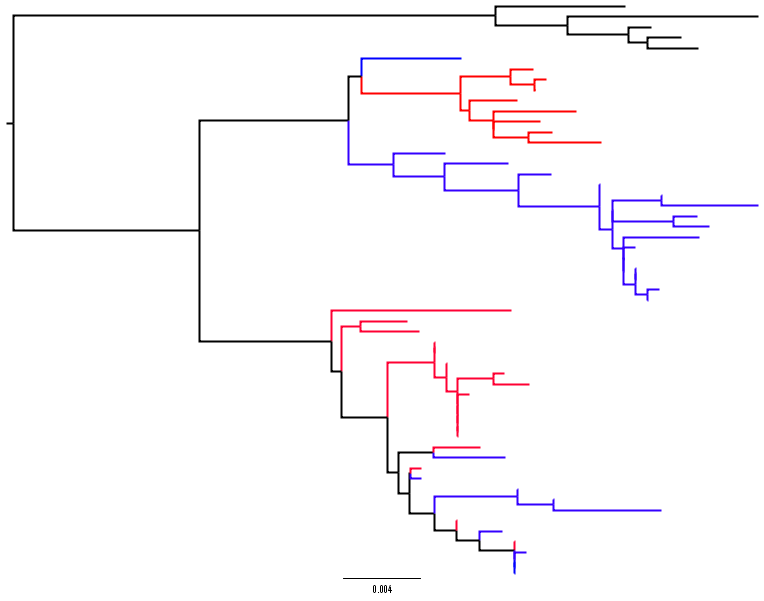


Figure S8: Mitochondrial tree topology vs. nuclear gene trees topology.



Mitogenome

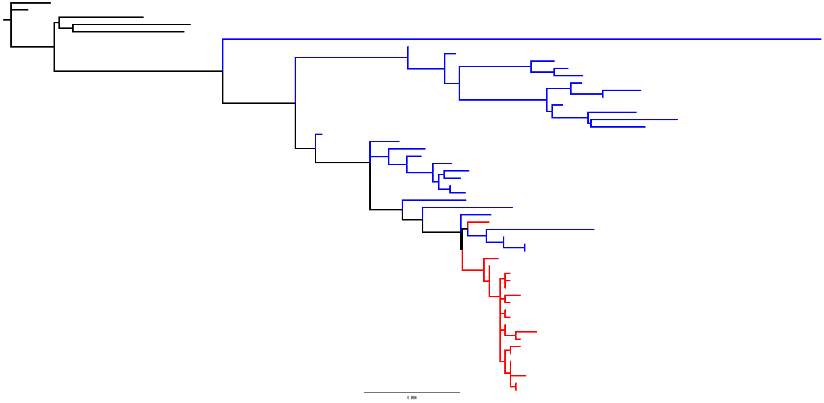
Nuclear gene 18,134

Nuclear gene 20,356

Nuclear gene 17,086

Nuclear gene 18,210

Nuclear gene 19,303



0.004

0.004

0.007

0.006

0.005

0.005

Figure S9: per-gene dXY per FST with regression lines (red: pyrenees, blue: alps)

