MOLECULAR ECOLOGY

Supplemental Information for:

Pleistocene climate fluctuations drove demographic history of African golden wolves (*Canis lupaster*)

Carlos Sarabia[1†], Juan C. Larrasoaña[2], Vicente Uríos[3], Bridgett vonHoldt[4], Jennifer A. Leonard[1†]

[1] Conservation and Evolutionary Genetics Group, Estación Biológica de Doñana (EBD-CSIC)

[2] Instituto Geológico y Minero de España (IGME)

[3] Vertebrate Zoology Research Group, University of Alicante

[4] Faculty of Ecology and Evolutionary Biology, University of Princeton

[†] Corresponding authors: Carlos Sarabia (cdomsar@gmail.com); Jennifer A. Leonard (jleonard@ebd.csic.es)

Table of Contents:

• **Table S2.** Total reads, mappability, percentage of PCR duplicates and coverage of 26 canid genomes mapped against CanFam3.1 • **Table S3.** Total reads, mappability, percentage of PCR duplicates and coverage of 23 canid genomes mapped against African hunting dog • **Table S4.** Comparison of % reads mapped, %PCR duplicates and coverages between genomes mapped against CanFam3.1 and African hunting dog reference genomes. • **Table S5.** Standard deviations of genome wide Fst values within populations of African golden wolves, gray wolves and coyotes. • **Table S6.** Genome wide inbreeding coefficient (Fi) vs different coverages of downsampled sets of genomes. • **Table S7.** Table of splitting events between the Algerian and Kenyan (various coverages) lineages. • **Table S8.** Divergence time estimations between two runs of MiSTI: using even and uneven time segments • **Table S9.** Divergence time estimations with the Cavalli-Sforza (1969) equation. • **Table S10.** Table of splitting events, polynomial equations of adjusted graphs to the data points, R2 and curve maximum calculated with the Newton-Raphson approach. • **Table S11.** Migration rates vs time segments per pair of lineages as calculated in MiSTI.	Sheet 3 Sheet 4 Sheet 5 Sheet 6 Sheet 7 Sheet 8 Sheet 9 Sheet 10
Supplementary Code • **00.cutadapt.sh:** Adapter trimming for raw data genomes • **00.reference.genomes.index.sh:** Indexing our reference genomes • **01.preprocessing.autoXMTY.sh:** Pre-processing pipeline for wild canids genome mapping • **02.a.ANGSD.distribution.qscores.sh:** Generating a distribution of quality scores • **02.b.ANGSD.genotype.likelihoods.sh:** Estimating genotype likelihoods with ANGSD • **02.c.ANGSD.PCA.sh:** PCA using genotype likelihoods with ANGSD and ngsTools • **02.d.ANGSD.ngsAdmix.sh:** Admixture proportions using ANGSD and ngsAdmix • **02.e.ANGSD.plink.merge.sh:** SNP merging and filtering • **02.f.ANGSD.plink.flashPCA.sh:** SNP-based PCA • **02.g.ANGSD.plink.Admixture.sh:** SNP-based Admixture • **03.SFS.Het.Fst.thetas.sh:** SFS calculation, heterozygosity, genomewide Fst and thetas • **04.PSMC.sh:** Pairwise Sequentially Markovian Coalescent (PSMC) • **05.ngsPSMC.sh:** Genotype likelihood-based PSMC (ngsPSMC) • **06.ROHs.Fi.sh:** Runs of Homozygosity (ROHs) and Inbreeding coefficient (Fi) • **07.a.MiSTI.sh:** Estimating divergence with PSMC-based Migration and Split Time	

Inference (MiSTI) • **07.b.MiSTI.replicability.cov.sh:** Replicability test: a. High and low genomic coverages • **07.c.MiSTI.replicability.times.sh:** Replicability test: b. Definition of time segments • **07.d.MiSTI.replicability.Cavalli.sh:** Replicability test: c. Cavalli-Sforza (1969) equation	

Supplementary Methods

Materials and methods

An African golden wolf roadkill from a previous study (Urios, Donat-Torres, Monroy-Vilchis, & Idrissi, 2015) from which the mitochondrial genome has been already published (KT378605) was found in 32º 33' 21.8'' N, 5º 50' 50.9'' W at the Issoulrhene area, 12 km southeast of Zaouiat Cheikh, in the Moroccan Atlas mountains. The location is hill slope at about 2000 meters of altitude, with dense patches of olive trees and sparse shrub vegetation around the valley. The sample was extracted from a roadkill and immediately put in 96% ethanol, and conserved at -20ºC until DNA extraction. Protocols for DNA extraction and library preparation followed Camacho-Sanchez et al. (2018). The library was sequenced on an Illumina NovaSeq at Johns Hopkins Genetic Resources Core Facility. The sample is referred to as "west Morocco" in this study to differentiate it from a previously published genome (Gopalakrishnan et al., 2018) from another Moroccan individual (referred to here as "east Morocco"). An additional 26 genomes were obtained from the literature: six African golden wolves (Kenya, Ethiopia, Egyptian Sinai, Senegal, east Morocco and Algeria), seven domestic dogs (two Nigerian village dogs, two African breeds -Saluki and Basenji-, and three Eurasian dogs from China, Qatar and India), six gray wolves (Saudi Arabia, China, Canada, Iran, Spain, Syria), two Eurasian golden jackals (Syria and Israel) one Ethiopian wolf, one African hunting dog (south Africa) and three coyotes (California, Illinois-Midwest, Mexico) (**Table S1**).

Pre-processing pipeline

We used cutadapt (Martin, 2011) to trim adapters and low quality base pairs (-q 20 option) in all raw reads files (.fastq) . Quality of the reads was evaluated visually with FastQC (Andrews, 2010). Reads were mapped using bwa mem v1.3 (Li & Durbin, 2010) to the reference genome of *Canis familiaris* (domestic dog) CanFam3.1 (Lindblad-Toh et al., 2005), with an assembled Y-chr (kindly provided by Dr. Krishna Veeramah). With the aim to compare admixture plots and Principal Component Analyses (PCA) and discard possible ascertainment biases introduced by the dog reference, reads were also mapped to an assembled reference genome of *Lycaon pictus* (African hunting dog) (Campana et al., 2016). We used samtools v1.9 (Li, 2011) to sort .bam files and to eliminate singletons and sequences with complementary reads in other chromosomes or with very low mapping quality (MAPQ<5). Duplicates and realigned reads around indels were removed simultaneously in all genomes using GATK v3.7 (McKenna et al., 2010) and the output .bam files were checked for mean read depth per chromosome. All reads had a sequencing quality higher than 20 and had a complementary read in the same chromosome.

Variant calling and quality filters

Reads mapped to autosomes in bam format were separated from reads mapping to sex chromosomes and mitochondrial DNA and were used for subsequent analyses. Before calling genotype likelihoods, reads of low quality and multiple hits were filtered out to print a distribution of quality scores using ANGSD (Korneliussen, Albrechtsen, & Nielsen, 2014) following Matteo Fumagalli's tutorial for ngsTools (Fumagalli, n.d.). We computed base quality alignment using option -baq (Li, 2011) and filtered reads with low mapping quality (-minMapQ 20). We then called genotype likelihoods in autosomes using ANGSD filtering out sites with depth <5X, reads with mapping quality less than 20, sites represented more than twice the mean coverage depth as in Freedman et al., (2014) and non-uniquely mapped reads. Since some of the genomes in our dataset had a mean low genome depth (around 5X), we used genotype likelihood frequencies to account for uncertainty in inferred genotypes in downstream analyses and called SNPs using the -doplink option in ANGSD. We downloaded the .refseq annotation file of CanFam3 from the UCSC Genome browser (Kent et al., 2002) and using in-home scripts, defined an exclusion zone of 10kb upstream and downstream the genes in order to exclude selective sweeps and select for neutral markers (Freedman et al., 2014).

Genetic structure

In order to study population genetic structure, genotype posterior probabilities of genotypes of the five Old World species (African golden wolves, dogs, gray wolves, Ethiopian wolf and Eurasian golden jackals) were generated with ANGSD (-dogeno 32 option) from genomes both mapped to the CanFam3.1 and African hunting dog reference genomes. The ANGSD output was used to perform a Principal Component Analysis (PCA) using the ngsCovar package in ngsTools (Fumagalli, Vieira, Linderoth, & Nielsen, 2014) and Rscript v3.4.4 (The R Core Team, 2017). We also estimated admixture proportions from genotype likelihoods from genomes mapped to both reference genomes using NGSadmix (Skotte & Albrechtsen, 2013), setting number of clusters (K) between 5 and 14. In order to estimate the best-fit K, NGSadmix was run until K=25. Further attempts to run the program with higher K values failed in our server. To avoid biases in local minima for K generated by NGSadmix, we ran the software five times, obtaining the same plot (data not shown).

Since we wanted to minimize ascertainment bias due to sites under linkage disequilibrium, we also used .glf.gz files from genomes mapped to both reference genomes to run ANGSD with the -doPlink option to call a number of SNPs based on the genotype likelihoods with a p-value=0.00001, using fixed frequencies (-doMaf 1) and using frequencies as prior (-doPost 1). These SNPs were filtered for genic regions (Freedman et al., 2014) and filtered for deviations from Hardy-Weinberg (HW) equilibrium (--hwe 0.001), minimum allele frequencies (--maf 0.05) and linkage disequilibrium (--indep-pairwise 50 5 0.5) using PLINK v1.9 (Purcell et al., 2007). With this curated set of unlinked SNPs in HW equilibrium, we performed a PCA using flashPCA (Abraham & Inouye, 2014) and estimated admixture composition using ADMIXTURE (Alexander, Novembre, & Lange, 2009) with a number of clusters (K) between 5 and 14. Visual comparisons were made for plots with genotype likelihoods and with SNP calling and for those reads mapped against CanFam3.1 and the *Lycaon pictus* genome. All PCA and admixture tests (both run

by NGSadmix and Admixture) were estimated excluding genic regions with 10 kb upstream and downstream.

Demographic history

We estimated a consensus sequence of our mapped .bam files using bcftools v1.9 (Li et al., 2009) without excluding genic regions. Calling a consensus sequence from low coverage samples (e.g., 7X-15X) can be misleading since the SNP calling process can misinterpret a heterozygous sites as homozygous and thus mask the real genomewide heterozygosity and dramatically modify estimations of demographic history (Nadachowska-Brzyska, Burri, Smeds, & Ellegren, 2016). To avoid this, we repeated the process with the actual coverage and downsampled .bam files of the Kenyan African golden wolf (7X, 9X, 11.12X, 15X and 24X) to visually estimate the best False Negative Rate (FNR) as suggested (Li & Durbin, 2011) and previously done in other studies (Hawkins et al., 2018; Kim et al., 2014). Plots of the Kenyan African golden wolf genome at these different mean genomewide coverages were corrected visually after a round of iterations with the psmc_plot.pl program in the PSMC package up to the third decimal of FNR and plotted together with and without the FNR correction (Fig. S1A,B). After determining the FNR per individual, we called consensus sequences applying the recommended coefficient to downgrade mapping duality for reads with excessive mismatches in bcftools mpileup (-C 50) and minimum and maximum coverage thresholds of 5 and 100, respectively (-d 5, -D 100) to generate .fq.gz files. PSMC was called using 64 atomic time intervals divided in a parameter of 6 and 58 parameters of one atomic interval each (-p "1*6+58*1") as specified in previous studies with African golden wolves (Freedman et al., 2014), and a correction for initial theta per individual and coverage following the README file of PSMC (Li & Durbin, 2011). PSMC files were generated by correcting initial theta using the -r option as suggested at the README file of PSMC (Li & Durbin, 2011). Since the option -r determines the theta_0/rho ratio and has a default value of 5, we estimated each original theta with the provided equation:

theta_FNR = theta_0/(1-FNR),

where theta_0=5 and FNR was calculated visually per each genome according to their coverage. FNR per each coverage (7X, 9X, 11.2X and 15X) was 0.35, 0.21, 0.18 and 0.11, respectively. Therefore, -r option was set as 7.6923, 6.3291, 6.0975 and 5.6179, respectively. We generated the .psmc file from the normal coverage (24X) with default parameters for -r. An example command would be:

$psmc -N20 -t10 -r6.3291 -b -p "1*6+58*1" -o Kenya.11.2X.psmc Kenya.11.2X.psmcfa

A round of 50 iterations of bootstrapping per genome was applied and all genome plots were overlaid to draw a multisample PSMC plot. Mutation rate was defined as $4.5 * 10^{-9}$ (Skoglund et al., 2015; Koch et al., 2019) and generation time as 3 years (Chavez et al., 2019; Freedman et al., 2014; Gopalakrishnan et al., 2018; Koepfli et al., 2015; Y. H. Liu et al., 2018). We took into consideration both ends of the mutation rate estimation by Koch et al. (2019) ($2.7\text{-}7.1 * 10^{-9}$), and plotted them without bootstrapping (Fig. S1C,D).

We explored a recently developed program to infer demographic population history in individual genomes with low coverage, ngsPSMC (Shchur, Korneliussen, & Nielsen, 2017). We generated files in ANGSD as input for ngsPSMC with the option -dopsmc and filtering for a minimum depth of 5X per genome. In ngsPSMC we used the same design for defining atomic time intervals as in PSMC and ran ngsPSMC using 50 iterations and initial popsizes per individual at 10^5 years ago as observed at the PSMC plot. We used the calculated genomewide nucleotide diversity per individual as theta (θ) (see "Methods: Summary Statistics") and calculated genomewide rho from a recombination map for dogs from a previous study (Auton et al., 2013). Mutation rate and generation time were defined as in PSMC. NgsPSMC is still under development, so bootstrapped plots could not be provided.

There is a robust collection of proxies to understand past climatic variability over the Sahara. We have considered both low- and high-latitude climate mechanisms influencing past environmental variability in the Sahara back to 1.5 million years ago (Drake, Breeze, & Parker, 2013; Ehrmann, Schmiedl, Beuscher, & Krüger, 2017; Larrasoaña, Roberts, & Rohling, 2013; McClymont, Sosdian, Rosell-Melé, & Rosenthal, 2013; Rohling, Mayewski, & Challenor, 2003; Smith, 2012) that could have affected the demographic history of African golden wolves in our PSMC and ngsPSMC plots, and also the speciation event that led to African golden wolves as previously proposed (Chavez et al., 2019; Koepfli et al., 2015) with a confidence interval of 400kyr. . We compared the timing of these events with the PSMC and ngsPSMC maxima and minima and observed any possible correlations between climatic events and increases or decreases of population.

Summary statistics

Our sample sizes (north African golden wolves (AGW) – 5 individuals, east AGW – 2 individuals) were too small for detection of more recent changes in population sizes using IBD-based methods such as SNeP (Barbato, Orozco-terWengel, Tapio, & Bruford, 2015) and IBDNe (Browning & Browning, 2015), so we needed an indirect approach to estimate recent changes in Ne. We inferred changes in Ne of non-genic regions with a series of thetas neutrality tests provided by the ANGSD package

Since most of our samples had a low coverage, we aimed to estimate or thetas and Fst taking genotype uncertainty into account. We relied on a likelihood-based estimation of site frequency spectrum (SFS) using ANGSD (Nielsen, Korneliussen, Albrechtsen, Li, & Wang, 2012). Briefly, a SFS calculation is an estimation of the proportion of sites at different allele frequencies and ANGSD is able to do so by computing genotype likelihoods first and calculate posterior probabilities of Sample Allele Frequency (SAF) for each site (Fumagalli, 2017). Previous studies have suggested that African hunting dogs have been evolving as an independent lineage from other canids since at least 1.7 Myr ago, with no detected recent admixture (Chavez et al., 2019), so we decided to use that reference genome (Campana et al., 2016) as ancestral to call unfolded SFS. SAF files were generated from .bam alignment files assuming HW equilibrium and using a multisample GL estimation option (-dosaf 1), with an upper depth filter of 2.5 times the mean read depth per sample. We also estimated SAF files per population as defined in **Table S1**. Using the realSFS program from the ANGSD package (Korneliussen, Moltke, Albrechtsen, & Nielsen, 2013; Nielsen et al., 2012), we generated SFS files from the SAF files and calculated their genomewide heterozygosity using the fraction of singletons from the whole SFS as in Gopalakrishnan et al., (2018). Although a measure of SFS is robust for low coverage samples (Han,

Sinsheimer, & Novembre, 2015), we wanted to study if lower coverages for some of our samples could decrease the measure of heterozygosity. We down sampled the Kenyan African golden wolf genome (24X) to mean depths of 7X, 9X, 11.2X and 15X using samtools view -bs (Li, 2011) and repeated the steps for SFS calling. We plotted mean genome wide observed and corrected heterozygosity for all samples.

In order to estimate mean Fst among individuals, a joint SFS between pairs of populations (2DSFS) was calculated. Using a 50kb sliding windows scan with realSFS (Fumagalli et al., 2014), we estimated region-based Fst values and calculated genome wide average Fst and 95% confidence intervals with custom scripts in bash. Finally, we computed a series of nucleotide diversity indexes (Tajima's D (Tajima, 1989), Fu and Li's F and D (Fu & Li, 1993), Fay's H (Zeng, Fu, Shi, & Wu, 2006), Zeng's E (Zeng et al., 2006)) and thetas (Θ_W, $\Theta\pi$, Θ_{FL}, Θ_H, Θ_L) (Durrett, 2008; Fay & Wu, 2000; Fu & Li, 1993; Tajima, 1989; Watterson, 1975; Zeng et al., 2006) using the -doThetas 1 option in ANGSD with population-based SFS as prior information (-pest), divided in 50-kb windows across the genome and excluding genic regions to avoid biases over genes under selection.

While Tajima's D and Fu and Li's D estimate ratios of rare variants as compared to common ones, Fay and Wu's H take into consideration the abundance of very high-frequency variants relative to intermediate-frequency variants. Zeng's E estimates the abundance of high-frequency variants relative to low-frequency variants. Fay and Wu's H and Zeng's E are the estimators with the highest sensitivity to changes in high-frequency variants as compared to the other estimators (Zeng et al., 2006). Zeng's E is the most sensitive test to population growth, since high-frequency variants reach equilibrium later than rare variants (Zeng et al., 2006).

Heterozygosity

We evaluated heterozygosity in our samples at a genome wide level through the estimation of genome wide inbreeding coefficient and runs of homozygosity (ROHs), comparing within populations of African golden wolves (north and east) and with other canids (coyotes, and middle eastern gray wolves). We defined four populations: afr_north (AGW from Algeria, Egypt, east Morocco, west Morocco, Senegal), afr_east (AGW from Ethiopia, Kenya), coyote (coyotes from California, Midwest and Mexico) and gwolf_me (gray wolves from S. Arabia, Iran and Syria). We used ngsF (Vieira, Fumagalli, Albrechtsen, & Nielsen, 2013) to calculate the inbreeding coefficient or F_i per individual. ngsF is part of the ngsTools package (Fumagalli et al., 2014) which works well with low coverage data. We estimated genotype likelihoods per individual (-doglf 3) with a p-value threshold of 0.001 in ANGSD. Then, we extracted the number of sites from the .mafs.gz file and computed 20 iterations in ngsF to find the best starting point to calculate F_i. F_i per individual was calculated in PLINK with --het. This method uses a number of called SNPs based on genotype likelihoods using the -doplink option in ANGSD that served as dataset for the SNP-based PCA and admixture plots from previous section (see "Methods: Variant calling and quality filters"). While PLINK needs genotypes to be called by ANGSD or other softwares, ngsF uses genotype likelihoods and is able to downstream the uncertainty of called genotypes if our samples have a low coverage. This approach could be more reliable for samples below 15X (Nadachowska-Brzyska et al., 2013).

In order to calculate ROHs across the whole genome we used two different approaches. The first method uses PLINK and makes use of the SNP dataset from the F_i calculation. In each population we removed SNPs in close linkage disequilibrium in 200-basepairs (bp) windows with a step size of 100 bp and a R^2 of 0.9 using option –indep-pairwise 200 100 0.90 in PLINK and generated ROHs as in Sams & Boyko, (2019). The second method uses the software ROHan (Renaud, Hanghøj, Korneliussen, Willerslev, & Orlando, 2019), which is especially adapted to work with low coverage genomes and uses Bayesian statistics to estimate local rates of heterozygosity, infers ROHs and computes local heterozygosity values inside and outside of ROHs. We ran ROHan in windows of 500kb, using the .bam mapped and filtered files from all African golden wolves, coyotes and gray wolves of the Middle East. Expected theta in ROHs was set to $2*10^{-5}$ and we used the default error rate of Illumina platforms as provided by the program. Plots of local heterozygosity were computed across the whole genome and a summary of ROHs was calculated. Finally, we calculated inbreeding coefficients from ROHs (F_{ROH}) as in (McQuillan et al., 2008; Sams & Boyko, 2019):

$$F_{ROH_j} = \frac{\sum_k lengt\,h\left(ROH_k\right)}{L};$$

where ROH_k is the *k*th ROH in individual *j*'s genome and L is the total length of the genome.

Divergence dating

The Sahara region has been subjected to cycles of desertification and regreening for the last 8 million years (Drake et al., 2013; Ehrmann et al., 2017; Larrasoaña et al., 2013; Smith, 2012). We used MiSTI (Shchur, 2019) to estimate times of divergence between local lineages represented by our seven individuals. We used a table of green Sahara periods (GSPs) in the last million years from Larrasoaña et al., (2013) to define time segments of potential connectivity among lineages. A list of cooler stadials associated with increased aridity of the Sahara region was considered as potential times for divergence (Ehrmann et al., 2017; Heinrich, 1988; Rohling et al., 2003). A pairwise time scale was generated using PSMC and 2DSFS files from previous sections. GNU Parallel (Tange, 2018) was used to model simultaneously different scenarios of divergence among lineages with different migration rates in dry periods and GSPs, using an automatized optimization round for migration rate per time per segment. We extracted a table of splitting times from MiSTI and plotted log likelihoods per proposed splitting time against time. Finally, a polynomial curve was fitted per group of data where $R^2 \geq 0.99$ to estimate the maximum point of the curve using the Newton-Raphson approach and a confidence interval of the upper 5%, 1% and 0.1% of log likelihood points.

Replicability of results in divergence time estimation

We have estimated divergence times within seven AGW lineages using a site frequency spectrum-based novel software, MiSTI (Shchur, 2019). Since this study represents, to our knowledge, the first publication where this software has been used, we wanted to ensure the replicability of our results by testing the program. We attempted to observe if there was a difference in divergence times estimations by changing the coverage of one of the genomes involved and the definition of time segments where MiSTI automatically estimates migration rates. We also wanted to compare results with estimations of time divergences using the Cavalli-Sforza (1969) equation.

a. High and low genomic coverages. To compare divergence times of two genomes with different coverages, we first used the .psmc files of the Kenyan genome at the original coverage (24X) and four downsampled coverages (7X, 9X, 11.2X and 15X, equivalent to the coverages of AGW from Senegal, Algeria, west Morocco and Egypt – East Morocco, respectively) from the "Demographic history" section (see above). Previous to generate the .psmc files, we accounted for the theta_0 correction as in section "Demographic history"

We estimated the unfolded SFS of each downsampled and normal coverage Kenyan .bam file (7X, 9X, 11.2X, 15X, 24X) using ANGSD, and used the .saf files to calculate the joint SFS (2DSFS) using realSFS as in section "Summary Statistics" (see above) separatedly between the Algerian genome and each coverage of the Kenyan genome. Following MiSTI's instructions (Shchur, 2019), we estimated joint site-frequency spectrum in MiSTI format for each Algerian-Kenyan combination using the script ANGSDSFS.py. Also, time scale files were calculated using calc_time,py of the MiSTI package joining time scales from both .psmc files in each case. Although some time steps were different, we gathered the ones that coincided between the five time scale files to run the program several times. An example of this can be see below:

time step	algeria.kenya.7X	algeria.kenya.9X	algeria.kenya.11.2X	algeria.kenya.15X	algeria.kenya.24X
0	0	0	0	0	0
1	1237	1487	1551	1776	1919
2	2205	2205	2205	2205	2205
3	2579	3100	3234	3703	4003
4	4037	4585	4585	4585	4585
5	4585	4852	5060	5793	6263
6	5618	6752	7042	7154	7154
7	7154	7154	7154	8060	8715
8	7334	8814	9192	9928	9928
9	9197	9928	9928	10519	11375
10	9928	11052	11525	12921	12921

In this table we see how several time steps coincide (0, 2205, 4585, 7154, 9928), which are the time steps extracted from the Algerian .psmc file.

We used these time steps to run MiSTI in parallel to optimize migration rates and possible time of divergences across the five combinations: algeria.kenya.7X, algeria.kenya.9X, algeria.kenya.11.2X, algeria.kenya.15X, algeria.kenya.24X. Loss of heterozygosity was calculated per coverage as in section "Heterozygosity" (see above). An example script is below:

time parallel --header : -j 20 $misti -uf --bsSize 10 --hetloss 0.21472 0.23025 --funits $MiSTI/setunits.canid.txt $in/psmc/afr_wolf.algeria.psmc $in/psmc/afr_wolf.kenya.7X.psmc $in/misfs/algeria.kenya.7X.mi.sfs {per} -o $in/mi/algeria.kenya.7X.opt.mi -mi 1 0 4 00.0 1 -mi 2 0 4 00.0 1 -mi 1 5 11 00.0 1 -mi 2 5 11 00.0 1 -mi 1 12 40 00.0 1 -mi 2 12 40 00.0 1 -mi 1 41 42 00.0 1 -mi 2 41 42 00.0 1 -mi 1 43 44 00.0 1 -mi 2 43 44 00.0 1 -mi 1 45 46 00.0 1 -mi 2 45 46 00.0 1 -mi 1 47 48 00.0 1 -mi 2 47 48 00.0 1 -mi 1 49 50 00.0 1 -mi 2 49 50 00.0 1 -mi 1 51 55 00.0 1 -mi 2 51 53 00.0 1 ::: per 17 18 19 20 21 22 23 24 25 26 27 28 >> $output/algeria.kenya.7X.out

Split times and log likelihoods were extracted from each output file, plotted against each other in Excel and a polynomial curve was fitted per group of data with R2>0.99 as in section "Divergence dating". Results are presented in Table S7.

b. Definition of time segments. In this section we define time segments as those segments between time steps generated using calc_time,py of the MiSTI package as in "High and low genome coverages" (see above). These time segments define either "humid" or "dry" conditions in Sahara. For example, using the time scale file from Algeria – Kenya (7X) above, two time segments could be 0-4 (0-4037 yr ago) and 5-11 (4585-11218 yr ago). The software architecture of MiSTI does not allow to repeat a time step (0,1,2,3,4…) from one time segment to another and the computation time to explore each one of the time segments possible is too much to explore all 21 divergence times at the same time. Since inputting manually these time segments could introduce a bias, we ran MiSTI twice in the same pair of genomes using paired time segments: 0-1; 2-3; 4-5… and 0-2; 3-4; 5-6… . Times and log-likelihood values were extracted from the MiSTI output files, plotted against each other and a polynomial curve was fitted per group of data as in section "Divergence dating". Results are presented in Table S8.

C. Comparison to Cavalli-Sforza estimation. We compared our results of MiSTI with an estimation of divergence times following the Cavalli-Sforza (1969) equation:

$T = - \log (1-\dot{F}_{ST})$ (1),

Where $\dot{F}_{ST}$ is an estimate of F_{ST} and T is scaled time, which is a measure of divergence and can be related to number of generations/years:

$T = t/2N$ (2), where t is number of generations/years since divergence and N is the population effective size of the two populations.

Finally, we can relate the mean genomewide estimations of Watterson's theta of section "Summary statistics" (see above) with a rough estimation of the population effective size:

$\theta_W = 4*Ne*\mu$ (3)

where θ_W is a measure of Watterson's theta, Ne is the population effective size and μ is the mutation rate per site and generation. Combining equations (1), (2) and (3) we have:

$$t = (- \log (1-\dot{F}_{ST}))*2*(\theta_W/(4*\mu)) \quad (4)$$

t is a mathematical measure of divergence time between two populations that, although robust, presents a great standard deviation (Nielsen et al., 1998). In our estimation we have firstly calculated genomewide $\dot{F}_{ST}$ and θ_W including genic regions between pairs of lineages (for example, Algeria – Kenya), but filtering out the sections of the curve with the lowest 5% number of sites per genomic window in R and retaining windows with a consistent number of sites. A range of values for the mutation rate per site and generation has been proposed in previous literature (Lindblad-Toh et al., 2005; Freedman et al., 2014; Skoglund, Ersmark, Palkopoulou & Dalen, 2015; Frantz et al., 2016; Koch et al., 2019). We chose $\mu = 4.0\text{-}4.5*10^{-9}$ as these values were more consistently used in previous studies (Skoglund et al., 2015; Frantz et al., 2016; Koch et al., 2019). Although Koch et al. (2019) proposes a wider range of $2.7\text{-}7.1*10^{-9}$ mutations per site and generation, we have observed that values close to $4.0*10^{-9}$ are more consistent with the estimated speciation time of African golden wolves (Koepfli et al., 2015, Chavez et al., 2019) (Figs. 3 and S1C,D). The results of the estimations for divergence times are presented in Table S9.

Replicability tests: results

a. High and low genomic coverages. We have observed generally consistent estimations of divergence times regardless of the coverage used, except for when one of the genomes was 9X and the other 7X (Table S7), when the divergence time is shown to be slightly later.

Such a situation is presented only in the Algeria (9X) – Senegal (7X) divergence, which is part of the EMAS cluster (see Figures 2, S2,S3,S5, Table 1 and main text Results: "Divergence during glacial periods"). In all three lineages from the EMAS cluster, divergence times (Algeria-Senegal, East Morocco-Algeria and East Morocco-Senegal) are lower than the earlier time step and have possibly diverged less than 2500-3000 yr ago. The estimation of the Algeria-Senegal divergence would not necessarily be underestimated since both the Senegal (7X) – East Morocco (11.2X) and the East Morocco (11.2X) – Algeria (9X) divergence times are close to 0, with no evidence of underestimation. Furthermore, all heterozygosity losses were accounted for in our scripts (see Supplementary Code Files 7a-d). For this reason, we trust the MiSTI results even in the Algeria-Senegal divergence time estimation.

b. Definition of time segments. We have found exactly the same estimations of divergence times regardless of what time segments were predefined. However, slightly different estimations of divergence times were found when using more or less time steps to draw the polynomial curve (see Table S8). For this reason, we decided to include all time steps between 0 and 150kyr ago for all MiSTI estimations of divergence times. Subsequently, we allowed MiSTI to automatically calculate migration rates between pre-defined time segments.

c. Comparison to Cavalli-Sforza estimation. We found a rather good consistency between results by MiSTI and by the Cavalli-Sforza equation (1969) (Table S9). As previously reported, the Cavalli-Sforza equation presents wide standard deviations (Nielsen et al., 1998), even though we used only significant 95% data under the curves of genome wide $\dot{F}_{ST}$ and θ_W.

However, the divergence time estimates with the Cavalli-Sforza (1969) equation did not coincide with our MiSTI estimates for those divergences involving either the Ethiopian or West Moroccan AGW.

These genomes present the lowest genome wide heterozygosities (see Table 2). The Cavalli-Sforza equation is heavily dependent on a proper estimation of genome wide Fst and assumes that with divergence times are so small that mutation is not as important as genetic drift (Nielsen et al., 1998). If the Ethiopian and west Moroccan lineages have been relatively isolated in mountainous ranges in small population sizes for a long time (as suggested by our results, see Tables 1 and 2 and Fig.S8), we expect to find a higher impact of genetic drift and therefore the Cavalli-Sforza estimate could be more affected than the MiSTI estimates, which rely upon both PSMC and 2DSFS more robust estimations.

Challenges and opportunities of working with low coverage genomes

In this study we are working with a number of medium and low coverage genomes (below 15X) that have posed certain challenges for demographic studies. Genotype-based studies rely upon SNP calling techniques that are mostly inefficient when using coverages lower than 15X and PSMC plots needed to be corrected according to false negative rates, which alter the shape and structure of the demographic history curves (Hawkins et al., 2018; Kim et al., 2014). A number of IBD-based programs (Barbato et al., 2015; Browning & Browning, 2015) that can detect more recent changes in population size require both SNP calling and bigger datasets. Although SFS-based methods (Harris & Nielsen, 2013; X. Liu & Fu, 2015) are mostly reliable if used in low coverage samples (Han et al., 2015) and has been used extensively (Nielsen et al., 2012), we observed different SFS-based measures of genomewide heterozygosity when downsampling a 24X genome to coverages of 7X, 9X, 11.2X and 15X. In line with this result, we propose correcting genomewide heterozygosity whenever using low and high coverage samples.

The most reliable method to study genotypes in medium and low coverage genomes is the discovery of genotype likelihoods. Softwares like ANGSD (Korneliussen et al., 2014) are able to call genotype likelihoods in a fairly big number of low coverage genomes and downstream the uncertainty of called genotypes to further analyses. Packages like ngsTools (Fumagalli et al., 2014) and softwares of recent development (ROHan (Renaud et al., 2019), ngsPSMC (Shchur et al., 2017), MiSTI (Shchur, 2019)) used in this study are bound to empower researchers working with low coverage whole genome sequences and help in the development of conservation policies for elusive or cryptic species. These techniques will be pivotal in studies where technical challenges to extract high coverage genomes or funding capacities are a limiting factor. In summary, future projects working with African golden wolves may benefit from the use of museum collections, low coverage genomes from opportunistic samples and possibly genomic libraries from noninvasive samples (Hernandez-Rodriguez et al., 2018) to respond evolutionary and ecological questions of one of the least studied canids in the world.

Supplementary Bibliography

Abraham, G., & Inouye, M. (2014). Fast principal component analysis of large-scale genome-wide data. *PLoS ONE*, *9*(4), 1–5. https://doi.org/10.1371/journal.pone.0093766

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc*.

Auton, A., Rui Li, Y., Kidd, J., Oliveira, K., Nadel, J., Holloway, J. K., … Boyko, A. R. (2013). Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. *PLoS Genetics*, *9*(12). https://doi.org/10.1371/journal.pgen.1003984

Barbato, M., Orozco-terWengel, P., Tapio, M., & Bruford, M. W. (2015). SNeP: A tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Frontiers in Genetics*, *6*(MAR), 1–6. https://doi.org/10.3389/fgene.2015.00109

Browning, S. R., & Browning, B. L. (2015). Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. *American Journal of Human Genetics*, *97*(3), 404–418. https://doi.org/10.1016/j.ajhg.2015.07.012

Camacho-Sanchez, M., Quintanilla, I., Hawkins, M. T. R., Tuh, F. Y. Y., Wells, K., Maldonado, J. E., & Leonard, J. A. (2018). Interglacial refugia on tropical mountains: Novel insights from the summit rat (Rattus baluensis), a Borneo mountain endemic. *Diversity and Distributions*, *24*, 1252–1266. https://doi.org/10.1111/ddi.12761

Campana, M. G., Parker, L. D., Hawkins, M. T. R., Young, H. S., Helgen, K. M., Szykman Gunther, M., … Fleischer, R. C. (2016). Genome sequence, population history, and pelage genetics of the endangered African wild dog (Lycaon pictus). *BMC Genomics*, *17*(1), 1–10. https://doi.org/10.1186/s12864-016-3368-9

Cavalli-Sforza, L. L. (1969). Human diversity. *Proc. 12th Int.Congr. Genet*. 2:405-416

Chavez, D. E., Gronau, I., Hains, T., Kliver, S., Koepfli, K.-P., & Wayne, R. K. (2019). Comparative genomics provides new insights into the remarkable adaptations of the African wild dog (Lycaon pictus). *Scientific Reports*, *9*(1), 8329. https://doi.org/10.1038/s41598-019-44772-5

Drake, N. A., Breeze, P., & Parker, A. (2013). Palaeoclimate in the Saharan and Arabian Deserts during the Middle Palaeolithic and the potential for hominin dispersals. *Quaternary International*, *300*, 48–61. https://doi.org/10.1016/j.quaint.2012.12.018

Durrett, R. (2008). Probability Models for DNA Sequence Evolution, Second Edition by Richard Durrett. In *Springer*. https://doi.org/10.1111/j.1751-5823.2009.00085_5.x

Ehrmann, W., Schmiedl, G., Beuscher, S., & Krüger, S. (2017). Intensity of african humid periods estimated from saharan dust fluxes. *PLoS ONE*, *12*(1), 1–18. https://doi.org/10.1371/journal.pone.0170989

Fay, J. C., & Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, *155*(3), 1405–1413.

Frantz, L.A.F., Mullin, V.E., Pionnier-Capitan, M., Lebrasseur, O., Ollivier, M., Perri, A., Linderholm, A., Mattiangeli, V., Teasdale, M.D., Dimopoulos, E.A. (2016). Genomic and archaeological evidence suggest a dual origin of domestic dogs. *Science* 352(6290):1228–1231.

Freedman, A. H., Gronau, I., Schweizer, R. M., Ortega-Del Vecchyo, D., Han, E., Silva, P. M., … Novembre, J. (2014). Genome Sequencing Highlights the Dynamic Early History of Dogs. *PLoS Genetics*, *10*(1). https://doi.org/10.1371/journal.pgen.1004016

Fu, Y. X., & Li, W.-H. H. (1993). Statistical Tests of Neutrality of Mutations. *Genetics*, *133*(3), 693–709.Fumagalli, M. (2017). A tutorial for some basic analyses using ngsTools/ANGSD. Available online at: https://github.com/mfumagalli/ngsTools/blob/master/TUTORIAL.md.

Fumagalli, M., Vieira, F. G., Linderoth, T., & Nielsen, R. (2014). NgsTools: Methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*, *30*(10), 1486–1487. https://doi.org/10.1093/bioinformatics/btu041

Gopalakrishnan, S., Sinding, M. H. S., Ramos-Madrigal, J., Niemann, J., Samaniego Castruita, J. A., Vieira, F. G., … Gilbert, M. T. P. (2018). Interspecific Gene Flow Shaped the Evolution of the Genus Canis. *Current Biology*, *28*(21), 3441-3449.e5. https://doi.org/10.1016/j.cub.2018.08.041

Han, E., Sinsheimer, J. S., & Novembre, J. (2015). Fast and accurate site frequency spectrum estimation from low coverage sequence data. *Bioinformatics*, *31*(5), 720–727. https://doi.org/10.1093/bioinformatics/btu725

Harris, K., & Nielsen, R. (2013). Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. *PLoS Genetics*, *9*(6). https://doi.org/10.1371/journal.pgen.1003521

Hawkins, M. T. R., Culligan, R. R., Frasier, C. L., Dikow, R. B., Hagenson, R., Lei, R., & Louis, E. E. (2018). Genome sequence and population declines in the critically endangered greater bamboo lemur (Prolemur simus) and implications for conservation. *BMC Genomics*, *19*(1), 1–15. https://doi.org/10.1186/s12864-018-4841-4

Heinrich, H. (1988). Origin and consequences of cyclic ice rafting in the Northeast Atlantic Ocean during the past 130,000 years. *Quaternary Research*, *29*(2), 142–152.

Hernandez-Rodriguez, J., Arandjelovic, M., Lester, J., de Filippo, C., Weihmann, A., Meyer, M., … Marques-Bonet, T. (2018). The impact of endogenous content, replicates and pooling on genome capture from faecal samples. *Molecular Ecology Resources*, *18*(2), 319–333. https://doi.org/10.1111/1755-0998.12728

Kent, J., Charles, S., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., … Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, *12*(6), 996–1006. https://doi.org/10.1101/gr.229102. Article published online before print in May 2002

Kim, H. L., Ratan, A., Perry, G. H., Montenegro, A., Miller, W., & Schuster, S. C. (2014). Khoisan hunter-gatherers have been the largest population throughout most of modern-human demographic history. *Nature Communications*, *5*. https://doi.org/10.1038/ncomms6692

Koch, E. M., Schweizer, R. M., Schweizer, T. M., Stahler, D. R., Smith, D. W., Wayne, R. K., & Novembre, J. (2019). De Novo Mutation Rate Estimation in Wolves of Known Pedigree. *Molecular Biology and Evolution*, *36*(11), 2536–2547. https://doi.org/10.1093/molbev/msz159

Koepfli, K. P., Pollinger, J., Godinho, R., Robinson, J., Lea, A., Hendricks, S., … Wayne, R. K. (2015). Genome-wide evidence reveals that African and Eurasian golden jackals are distinct species. *Current Biology*, *25*(16), 2158–2165. https://doi.org/10.1016/j.cub.2015.06.060

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, *15*(1), 356. https://doi.org/10.1186/s12859-014-0356-4

Korneliussen, T. S., Moltke, I., Albrechtsen, A., & Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics*, *14*(1). https://doi.org/10.1186/1471-2105-14-289

Larrasoaña, J. C., Roberts, A. P., & Rohling, E. J. (2013). Dynamics of Green Sahara Periods and Their Role in Hominin Evolution. *PLoS ONE*, *8*(10). https://doi.org/10.1371/journal.pone.0076514

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496. https://doi.org/10.1038/nature10231

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. https://doi.org/10.1093/bioinformatics/btp352

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., … Lander, E. S. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, *438*(7069), 803–819. https://doi.org/10.1038/nature04338

Liu, X., & Fu, Y. X. (2015). Exploring population size changes using SNP frequency spectra. *Nature Genetics*, *47*(5), 555–559. https://doi.org/10.1038/ng.3254

Liu, Y. H., Wang, L., Xu, T., Guo, X., Li, Y., Yin, T. T., … Zhang, Y. P. (2018). Whole-Genome sequencing of African dogs provides insights into adaptations against tropical parasites. *Molecular Biology and Evolution*, *35*(2), 287–298. https://doi.org/10.1093/molbev/msx258

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(10).

McClymont, E. L., Sosdian, S. M., Rosell-Melé, A., & Rosenthal, Y. (2013). Pleistocene sea-surface temperature evolution: Early cooling, delayed glacial intensification, and implications for the mid-Pleistocene climate transition. *Earth-Science Reviews*, *123*, 173–193. https://doi.org/10.1016/j.earscirev.2013.04.006

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Proceedings of the International Conference on Intellectual Capital, Knowledge Management & Organizational Learning*, *20*, 254–260. https://doi.org/10.1101/gr.107524.110.20

McQuillan, R., Leutenegger, A. L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., … Wilson, J. F. (2008). Runs of Homozygosity in European Populations. *American Journal of Human Genetics*, *83*(3), 359–372. https://doi.org/10.1016/j.ajhg.2008.08.007

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). https://doi.org/10.1371/journal.pgen.1003942

Nadachowska-Brzyska, K., Burri, R., Smeds, L., & Ellegren, H. (2016). PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. *Molecular Ecology*, *25*(5), 1058–1072. https://doi.org/10.1111/mec.13540

Nielsen, R., Mountain, J.L., Huelsenbeck, J.P., Slatkin, M. (1998). Maximum-Likelihood estimation of population divergence times and population phylogeny in models without mutation. *Evolution,* 52(3). 1998. pp. 669-677

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE*, *7*(7). https://doi.org/10.1371/journal.pone.0037558

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Renaud, G., Hanghøj, K., Korneliussen, T. S., Willerslev, E., & Orlando, L. (2019). Joint Estimates of Heterozygosity and Runs of Homozygosity for Modern and Ancient Samples. *Genetics*, *212*(July), 587–614. https://doi.org/10.1534/genetics.119.302057

Rohling, E. J., Mayewski, P. A., & Challenor, P. (2003). On the timing and mechanism of millennial-scale climate variability during the last glacial cycle. *Climate Dynamics*, *20*(2–3), 257–267. https://doi.org/10.1007/s00382-002-0266-4

Sams, A. J., & Boyko, A. R. (2019). Fine-scale resolution of runs of homozygosity reveal patterns of inbreeding and substantial overlap with recessive disease genotypes in domestic dogs. *G3: Genes, Genomes, Genetics*, *9*(1), 117–123. https://doi.org/10.1534/g3.118.200836

Shchur, V. (2019). *MiSTI: PSMC-based Migration and Split Time Inference from two genomes*.

Shchur, V., Korneliussen, T. S., & Nielsen, R. (2017). ngsPSMC: genotype likelihood-based PSMC for analysis of low coverage NGS. Retrieved from https://github.com/ANGSD/ngsPSMC

Skoglund P, Ersmark E, Palkopoulou E, Dalen L. 2015. Ancient wolf genomereveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Curr Biol.* 25(11):1515–1519.

Skotte, L., & Albrechtsen, A. (2013). Estimating Individual Admixture Proportions from. *Genetics*, *195*(November), 693–702. https://doi.org/10.1534/genetics.113.154138

Smith, J. R. (2012). Spatial and temporal variations in the nature of Pleistocene pluvial phase environments across Africa. In *Modern Origins A North African perspective* (pp. 35–77).

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595. https://doi.org/PMC1203831

Tange, O. (2018). *GNU Parallel 2018*. https://doi.org/10.5281/zenodo.1146014

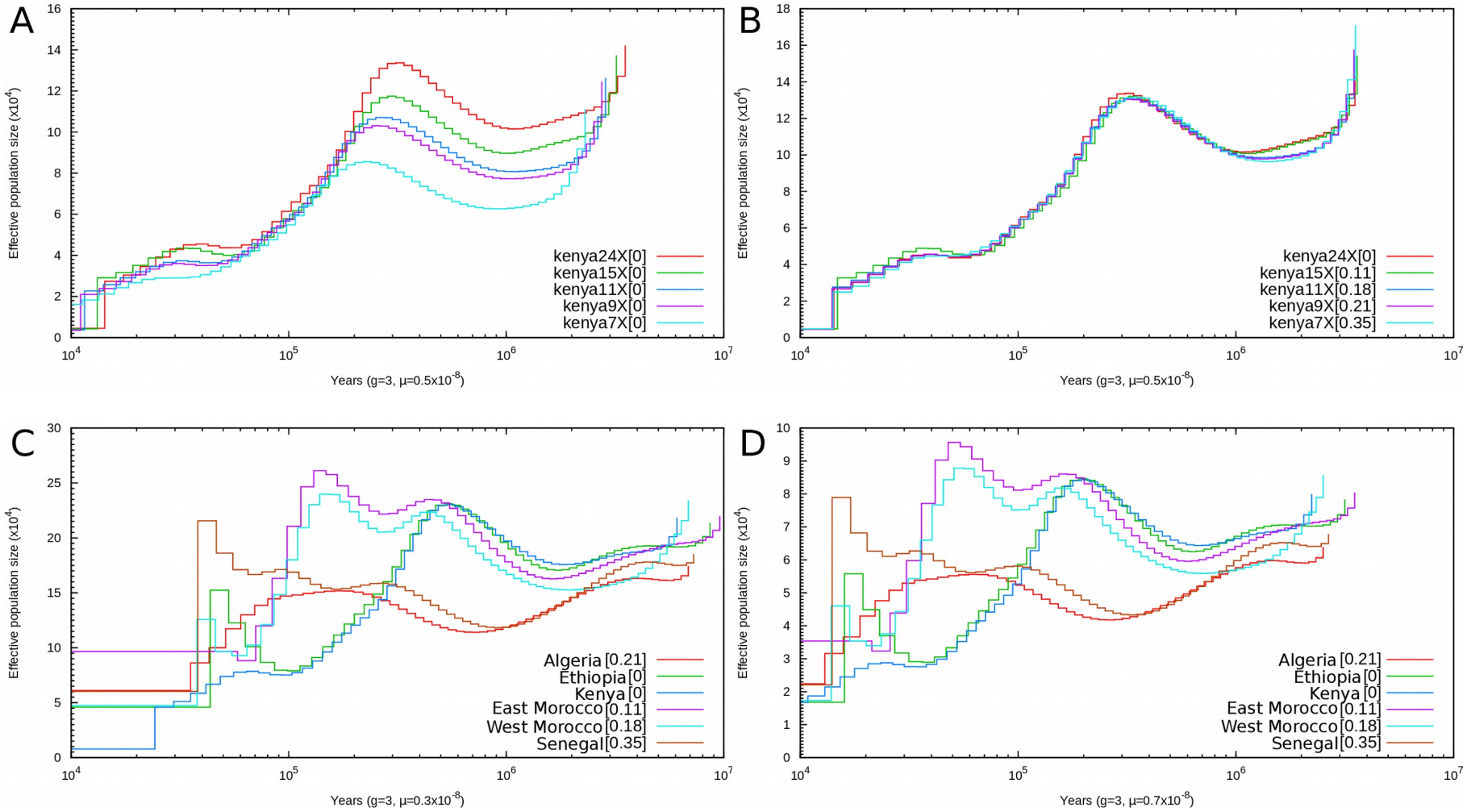
Team, R. C. (2017). *R: A language and environment for Statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/*.

Urios, V., Donat-Torres, M. P., Monroy-Vilchis, C. R. O., & Idrissi, H. R. (2015). El análisis del genoma mitocondrial del cánido estudiado en Marruecos manifiesta que no es ni lobo (Canis lupus) ni chacal euroasiático (Canis aureus). *Altotero*, *3*.

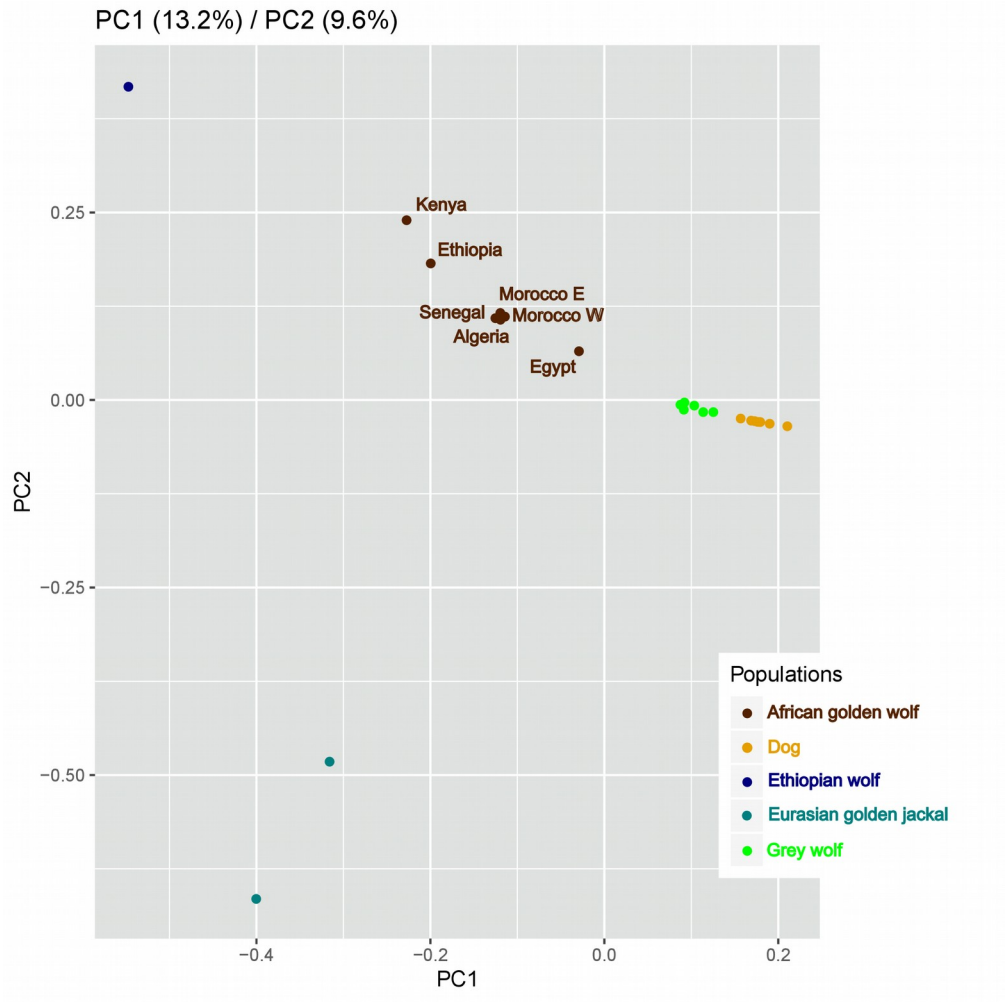
Vieira, F. G., Fumagalli, M., Albrechtsen, A., & Nielsen, R. (2013). Estimating inbreeding coefficients from NGS data: Impact on genotype calling and allele frequency estimation. *Genome Research*, *23*(11), 1852–1861. https://doi.org/10.1101/gr.157388.113

Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, *7*, 256–276.

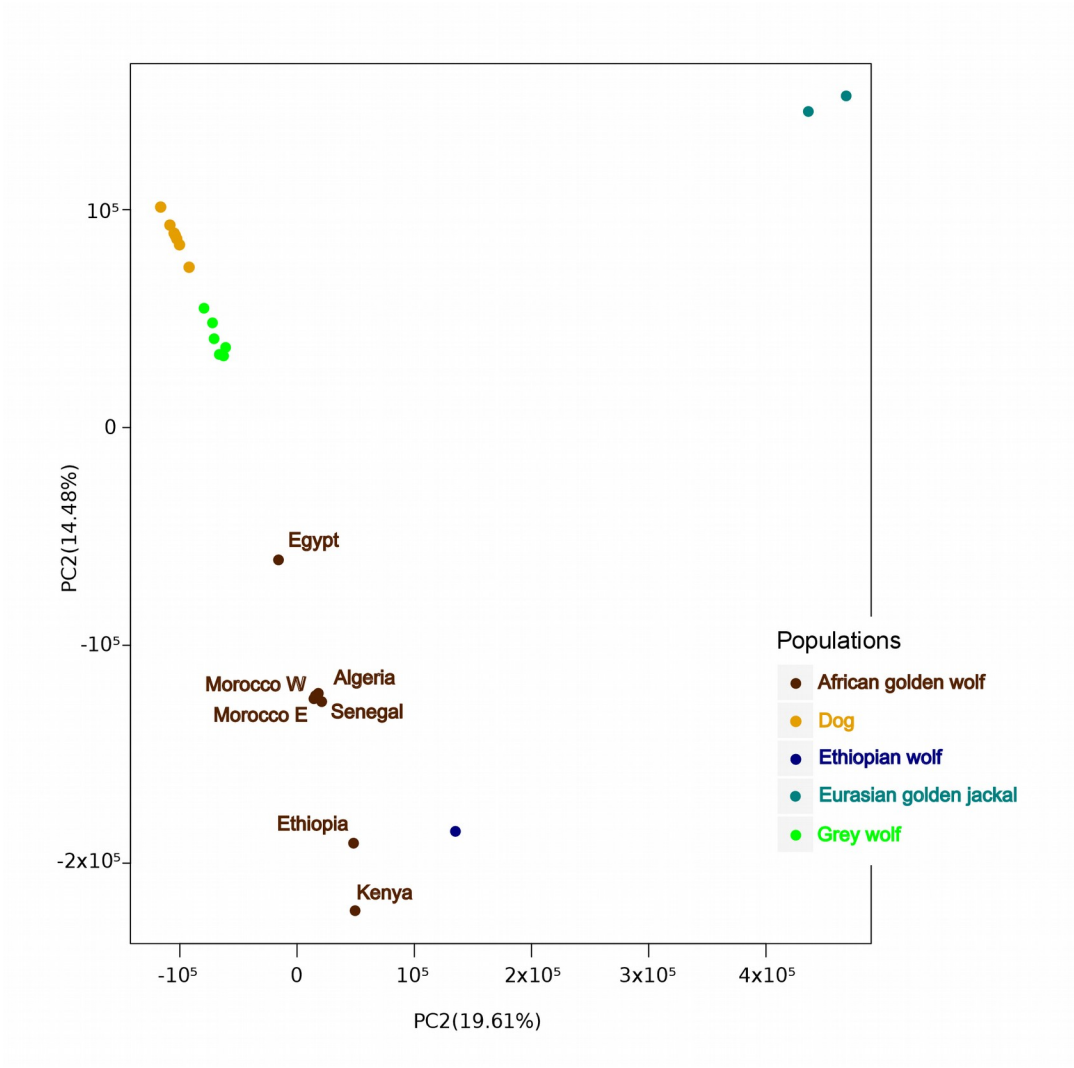
Zeng, K., Fu, Y. X., Shi, S., & Wu, C. I. (2006). Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*, *174*(3), 1431–1439. https://doi.org/10.1534/genetics.106.061432



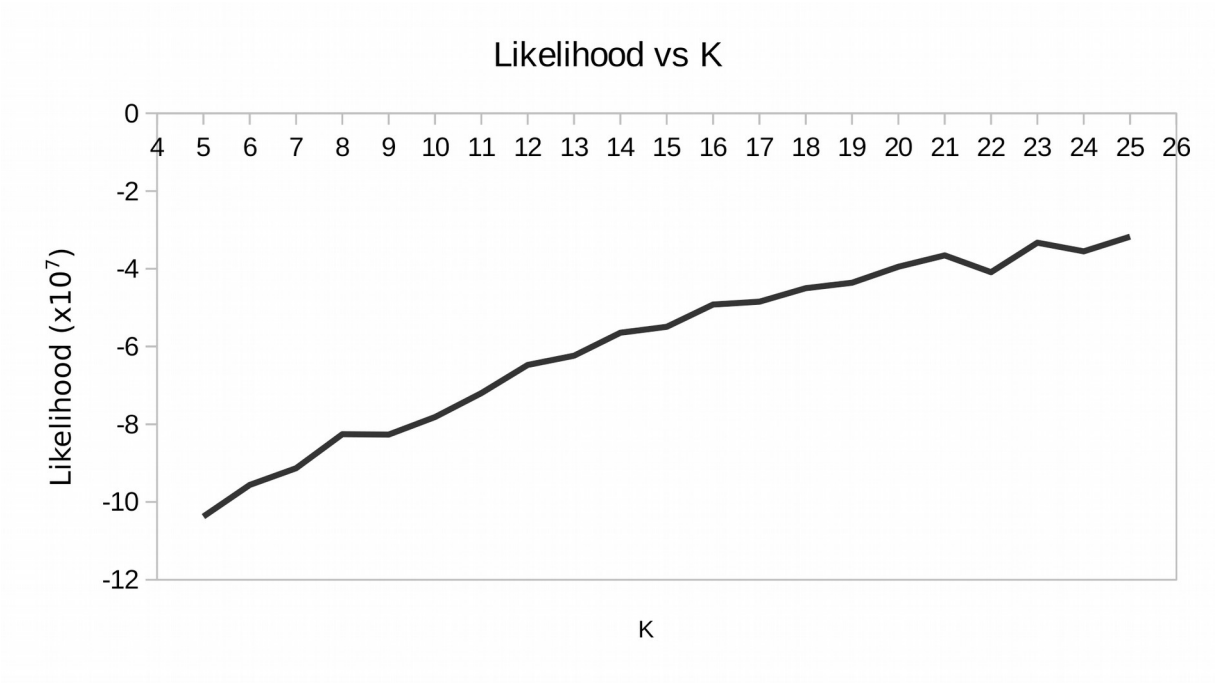
Supplementary Figure 1. PSMC plots of African golden wolf (AGW) genomes under different conditions. A, B represent PSMC plots of the Kenyan AGW with the normal (24X) and downsampled coverages (15X, 11.2X, 9X, 7X) without (A) and with (B) False Negative Rate correction for low heterozygosity due to low coverages. C, D represent PSMC plots of six AGW (Algeria, Ethiopia, Kenya, East Morocco, West Morocco, Senegal) with lower (C) and upper (D) bounds of the mutation rate estimation by Koch et al., (2019) ($2.7\text{-}7.1*10^{-9}$).



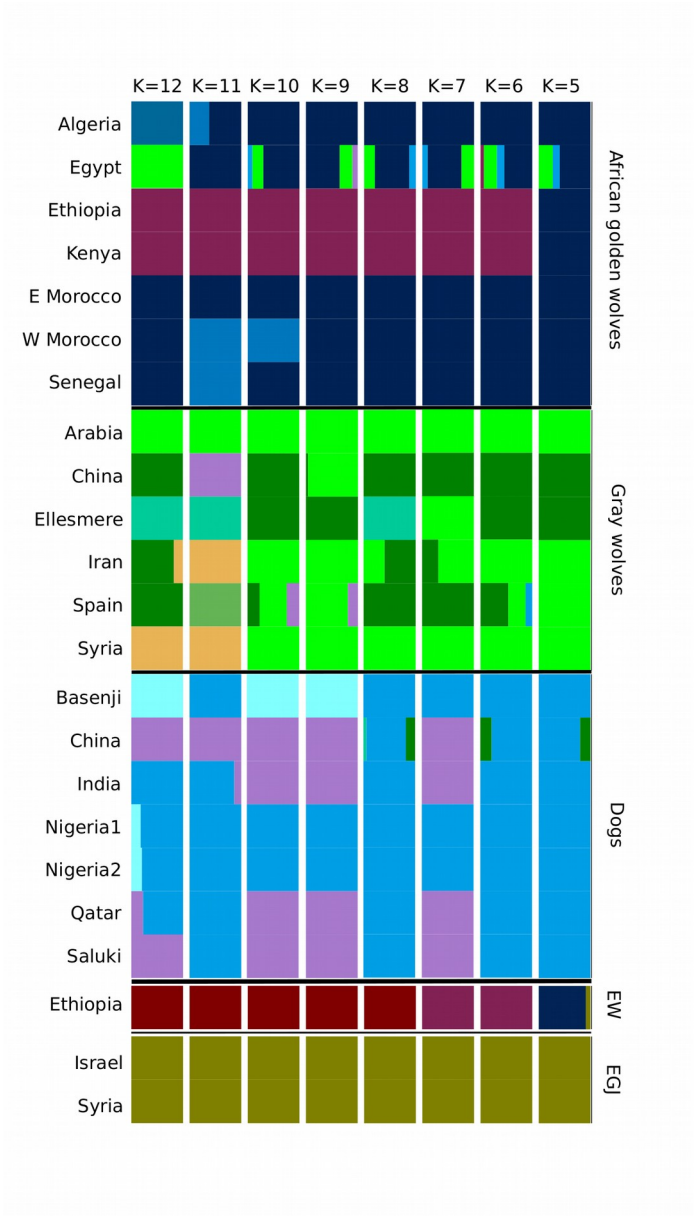
Supplementary Figure 2. Genotype likelihood-based Principal Component Analysis (PCA) generated by ngsCovar from the ngsTools package. PCA was called using 2.54 million sites in 16 genomes of wild Old World canids (African golden wolves, gray wolves, Ethiopian wolves, Eurasian golden jackals) and 7 genomes of domestic dogs.



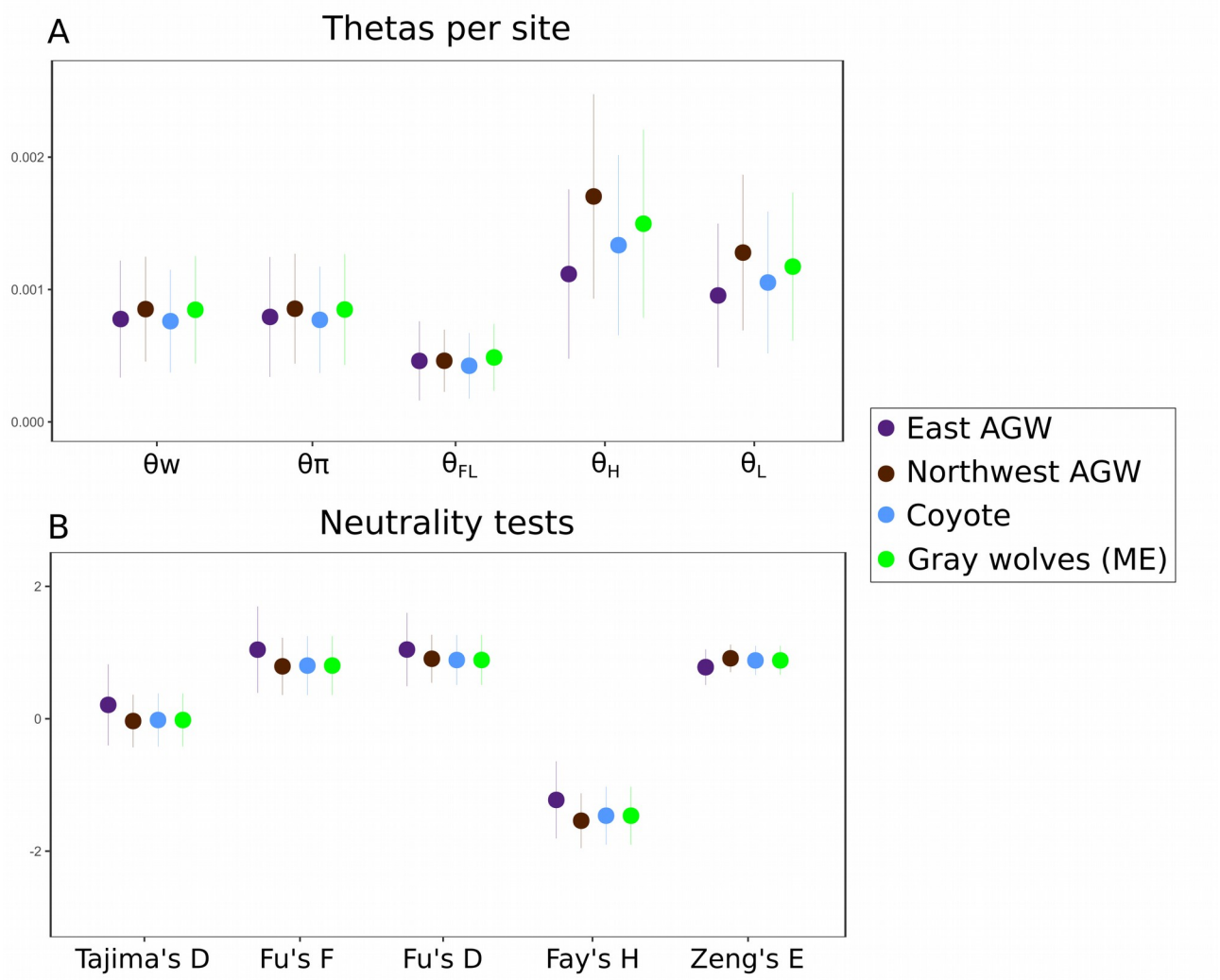
Supplementary Figure 3. SNP-based Principal Component Analysis (PCA) of 23 canid individuals. SNPs were called based in genotype likelihood using ANGSD with the -doPlink option, curated and filtered for Hardy-Weinberg equilibrium and linkage disequilibrium using PLINK v1.9. PCA was generated by flashPCA using 625,000 sites in 16 genomes of wild Old World canids (African golden wolves, gray wolves, Ethiopian wolves, Eurasian golden jackals) and 7 genomes of domestic dogs.



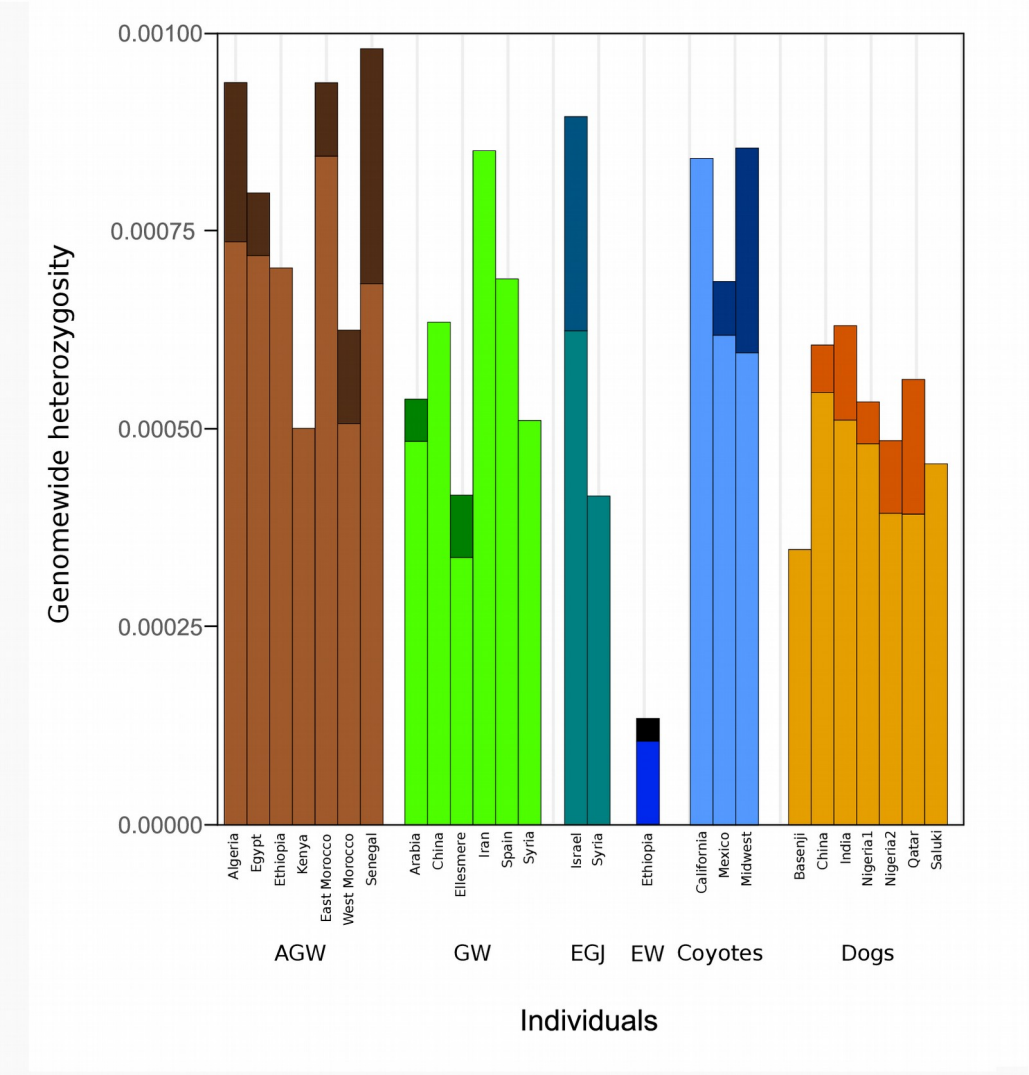
Supplementary Figure 4. Best-fit calculation of K (likelihood) vs values of K as calculated by NGSadmix using 23 genomes o
Old World canids mapped against African hunting dog (admixture plot in **Figure 2**).



Supplementary Figure 5. SNP-based Admixture plots of Old World canids mapped against African hunting dog showing admixture proportions, including the 23 individuals used at this study. SNPs were called as in **Supplementary Figure 3**. Eastern (Kenya, Ethiopia) African golden wolves cluster in a different group from those from the north (Egypt, Algeria, East Morocco, West Morocco, Senegal). EW: Ethiopian wolf. EGJ: Eurasian golden jackal. This plot is based in 625,000 unlinked sites.

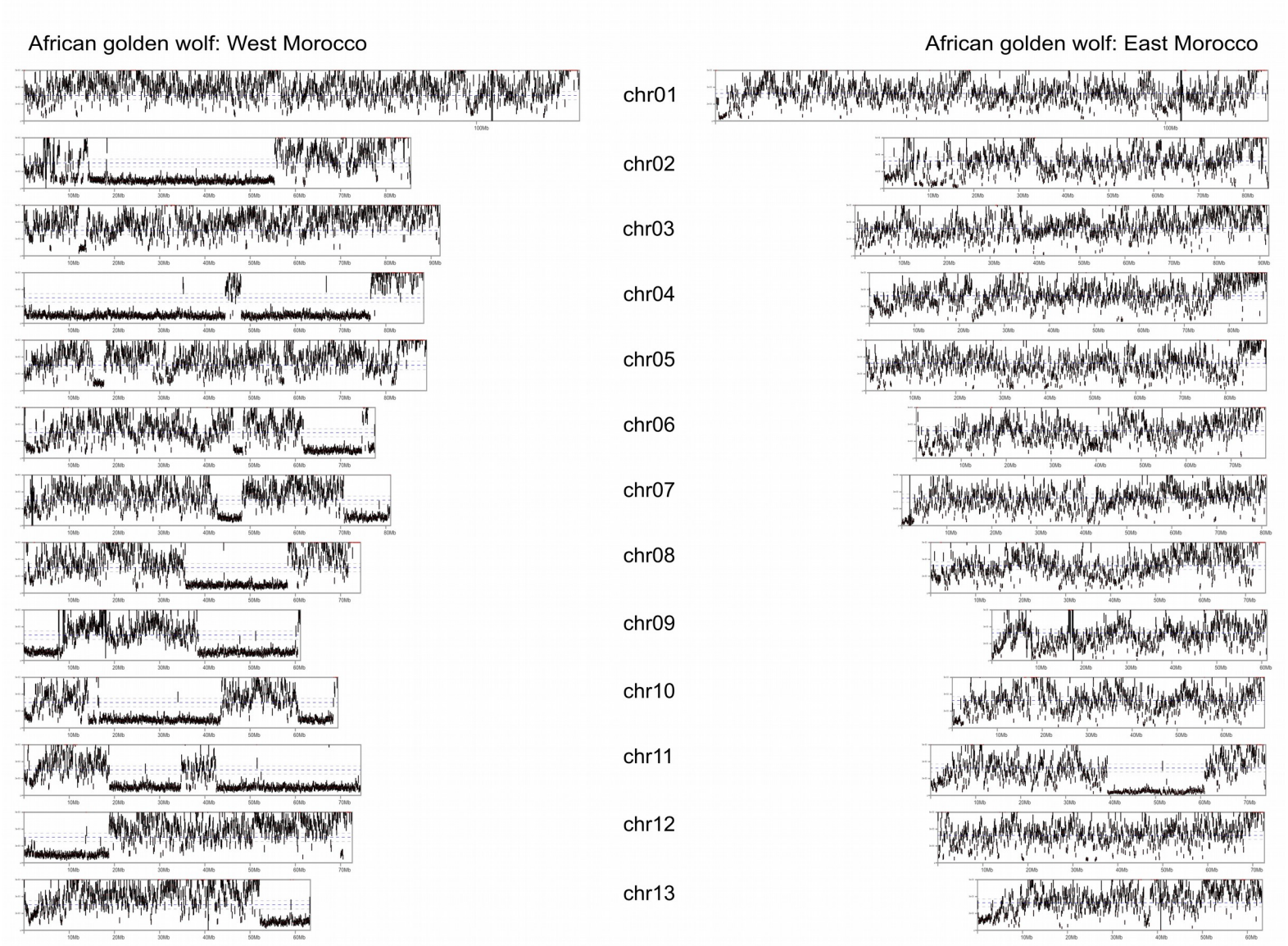


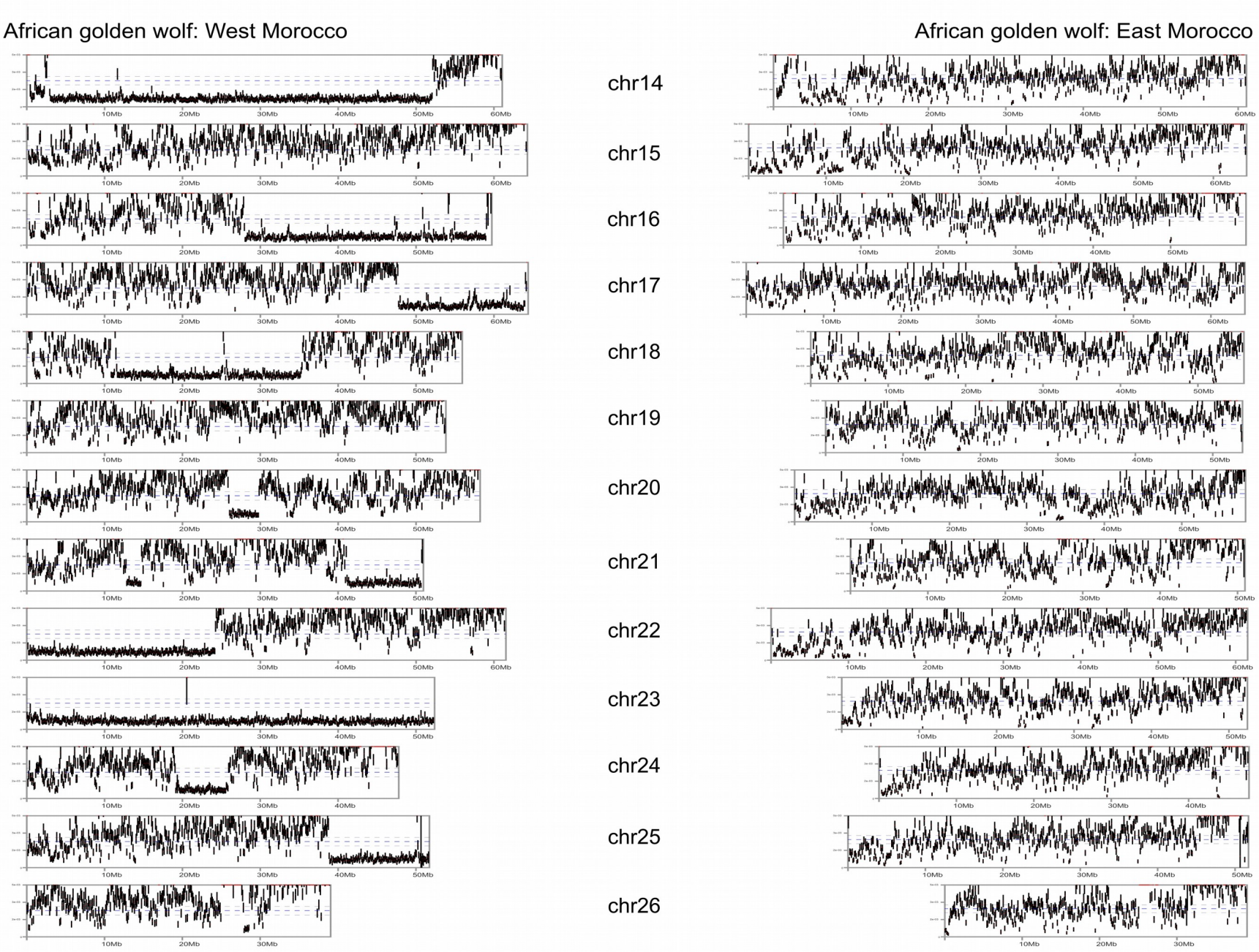
Supplementary Figure 6. Thetas per site and neutrality tests of four populations: east African Golden Wolves (AGW) (Ethiopia, Kenya), northwest AGW (Algeria, east Morocco, west Morocco, Senegal), Coyote (California, Midwest, Mexico), Gray wolves of the Middle East (ME) (S. Arabia, Iran, Syria). We considered 50-kb non-overlapping windows across the whole genome and filtered out those windows with a number of sites outside the 99.7% of the distribution (mean +/- 3 standard deviations). Theta statistics were calculated dividing by the total number of sites. Neutrality tests were averaged per window.



Supplementary Figure 7. Genome wide heterozygosity calculated per individual and population. Heterozygosity was calculated using the fraction of singletons from the unfolded Site-Frequency Spectrum (SFS). Genome wide heterozygosities were corrected using the Kenyan African golden wolf genome (24X) and down sampling it to each coverage, calculating proportion of lost heterozygosity for each coverage. Corrections are marked in darker colors. AGW: African golden wolves. GW: gray wolves. EGJ: Eurasian golden jackals.

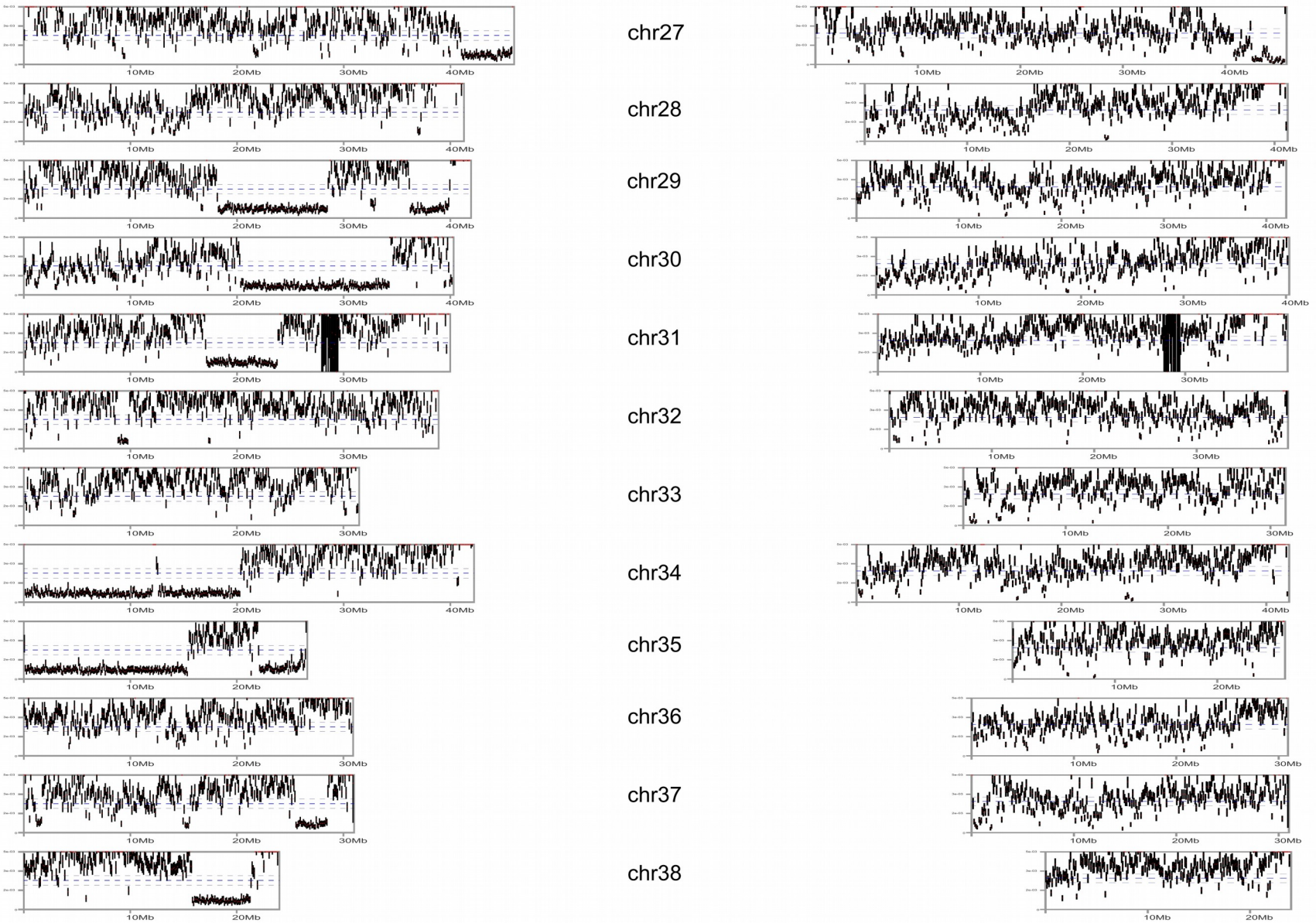
Supplementary Figure 8. Heterozygosity plots of east and west Moroccan individuals per chromosome. Plots were generated with ROHan using 500kB windows, minimum coverage of 5X and maximum coverage of 2.5 times the mean coverage per genome. --rohmu option were set as 2e-5. All other settings were left as default.

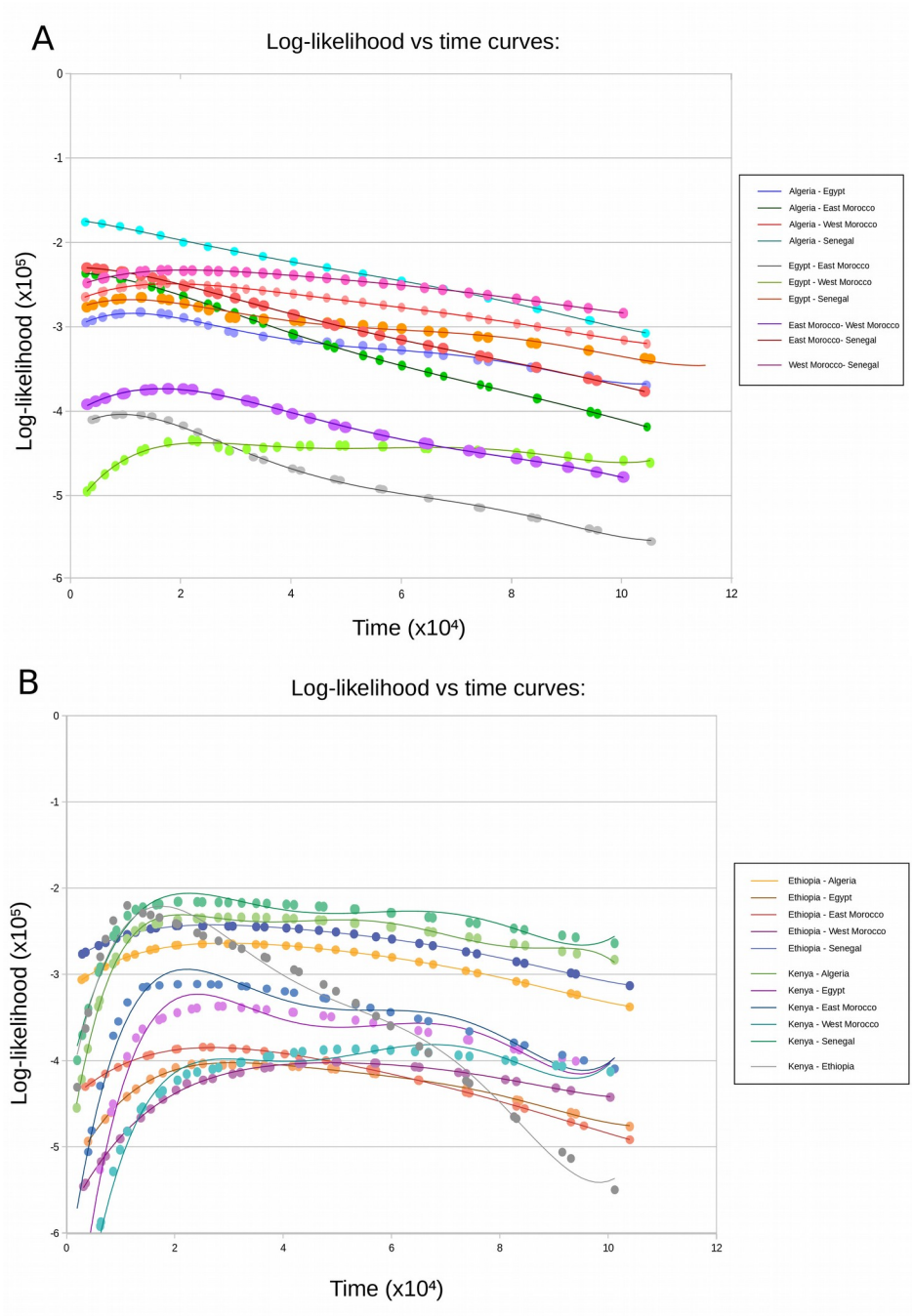



African golden wolf: West Morocco
African golden wolf: East Morocco
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chr23
chr24
chr25
chr26
10Mb
20Mb
30Mb
40Mb
50Mb
60Mb

African golden wolf: West Morocco

African golden wolf: East Morocco





Supplementary Figure 9. Log-likelihood of divergence between members of the north African golden wolf cluster (A) and north vs. east African golden wolf cluster (B) vs time. Likelihood of divergence times was calculated paralellizing MiSTI with GNU Parallel using the default optimization round. Time segments were defined using likely aridization / regreening Sahara periods defined in the literature. A polynomial curve equation was adjusted to the fifth degree and plotted when $R^2 > 0.99$.