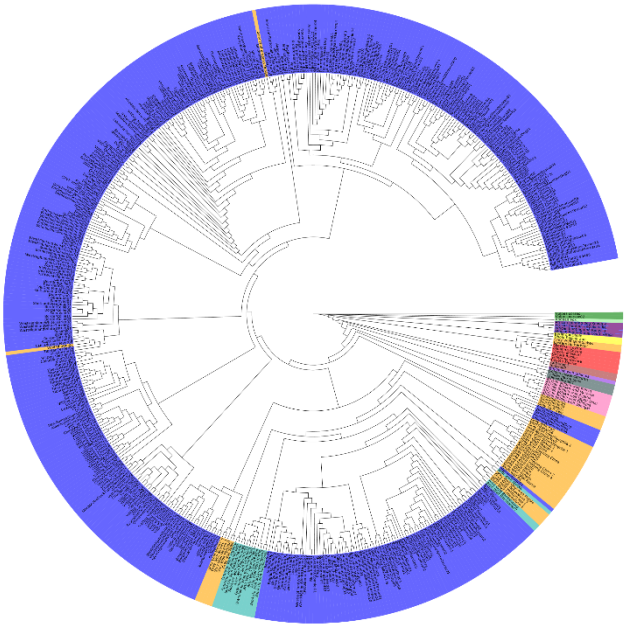
Master of Science in Omics Data Analysis

Master Thesis

Evolution of mtDNA variation during canid's domestication

By

Víctor Manuel González Basallote



Supervisor: Tomàs Marquès-Bonet, Institut de Biologia Evolutiva - University of Pompeu Fabra

Co-supervisor: Narcis Fernández Fuentes, Biosciences Department, University of Vic

Department of Systems Biology

University of Vic – Central University of Catalonia

18/09/2017

ACKNOWLEDGEMENTS

I would like to thank Tomàs Marquès for giving me the opportunity to participate in the Comparative Genomics Lab (IBE-CNAG) and to make this project from the first moment we got in touch. From the first day, I have felt surrounded by a totally calm atmosphere with a very good relationship between my colleagues, which has contributed enormously in the accomplishment of my project, and what is more important, in my learning. Such a great environment has surrounded me where I have met incredible people who have helped me when I needed it. Specially, I would like to give my most heartfelt thanks to Aitor Serres and Irene Lobón for their incredible support, which has allowed me to solve all the encountered problems I have had and to learn much more than I would have done without them during all these months.

I would like also to thank my family, especially my parents, for having been my cornerstone during the last 5 years. For wanting me and missing me so much from a distance. For making me, even if I do not prove it, to miss you every day. For letting me study and develop a little more as a person. There really are no words to describe the admiration I feel for you. Thank you very much.

Abstract

The close coexistence between human beings and dogs and the efficiency in their domestication have made dogs a very interesting target for basic research. By studying the history and evolution of dogs, much can be learned about human health, history and social behavior. Moreover, dogs show some astounding diversity patterns which make them mystifyingly unique; to date, no other species has ever been found to spawn such a huge morphological variability (dog breeds) from a very little initial effective population size in a very short period of time.

In this project, we have put all our efforts into two noteworthy aspects. First, we completely redesigned an in-house mitochondrial DNA reconstruction pipeline which, after having assessed its efficiency and performance, we used it to reconstruct a large dataset containing almost all the available dog breeds and well as other samples from different canids. Then, the reconstructed sequences were used to study crucial aspects of dog evolution: A massive scale phylogeny was built to see whether the current morphology-based classification of dogs truly correlates with their genetic distances as well as other aspects. In addition to that, we have inferred the population history of dogs, which showed a population bottleneck and posterior expansion attributable to a domestication event, and we got results about the Time to the Most Common Ancestors (TMRCAs) between divergent species. Furthermore, we assessed some evolutionary statistics which showed the strong purifying selection acting on canids, but also that dogs could have a relaxation in this selection due to the artificial breeding and selection made by humans.

Finally yet importantly, we encountered a discrepancy in the starting codon of the ND4L gene in the reference genome of dogs (*CanFam 3.1.)* which suggest that the gene does not start in the proposed codon but in a nearby one.

Key words: bioinformatics, mtArchitect, dogs, canids, wolves, phylogenesis, mitochondrial DNA.

Table of Contents

List of Figures

List of Tables

List of Codes

1. Introduction

1.1. Genus Canis

The genus *Canis* is a genus of the mammal family *Canidae* and is represented by many well known species and subspecies such as dogs (*Canis lupus* familiaris), grey wolves (*Canis* lupus), African golden wolves (*Canis* anthus), Ethiopian wolves (*Canis simensis)*, eurasian golden jackals (*Canis aureus*) and coyotes (*Canis latrans*). It belongs to the tribe *Canini* which is represented by the genus *Canis* and by dholes (*Cuon alpinus*), African hunting dogs (*Lycaon pictus*) and the south american foxes (from the genus *Lycalopex*) (1). It is the sister group of the genus *Vulpes* (foxes) which shares a common ancestor from which they diverged around 9-11 millions years ago (2) and together with genus *Vulpes* they constitute the family *Canidae* (3).

In particular, dogs (*Canis lupus familiaris*) are a subspecies of the grey wolf. Experts suggest that they originated as product of domestication of the grey wolves 15000-40000 years ago (4–6), although the number of events as well as their geographical location are in doubt (7–9). Despite such disparity, the data suggests that dogs appeared while humans were hunter-gatherers (6,10). Dogs have helped us in crucial tasks since their domestication e.g. hunting, help with grazing, protection of the household, companionship, etc., Modern dogs comprise a mixture of sizes, shapes, abilities among other characteristics since people have selectively bred them to find the optimal and desirable skills for each breed. For example, as stated in Parker et al, 2017 (11), we can see a separation between herding dog and hunting dog breeds, which suggest selection based on those characteristics. In fact, the current dog classification is based on its function in the society, such as companion dogs, guard dogs, hunting dogs or working dogs. Nevertheless, most of the existing breeds are quite recent, based for example on physical traits nowadays. Moreover, as often highlighted (12), dogs show some astounding diversity patterns which make them mystifyingly unique; to date, no other species has ever been found to spawn such a huge morphological variability (dog breeds) from a very little initial effective population size in a very short period of time. Thus, the leading role that dogs have played in human lives has become clear, making it interesting to study the evolution of dogs to infer the characteristics of human society over the history and learn more from it.

Like dogs, the history of grey wolves (*Canis lupus*) is a bit unclear, given their migratory movements (13), and their relationship with other species like coyotes (13,14). Moreover, there is a debate with the current subspecies classification as there are some subspecies of grey wolves which have been suggested to be distinct species according to their genetic, physical and historical differences (13,15–19).

In fact, to demonstrate the doubts that lie within this genus, Koepfli et al, 2015 established only two years ago that what was believed to be an unique specie, the golden jackal, resulted in two distinct species, the African golden wolf (*Canis anthus*), and the Eurasian golden jackal (*Canis aureus*) (20).

We wanted to explore all these issues by reconstructing the evolutionary history of almost all dog breeds as well as other canids and trying to clarify some unclear aspects of the history of the genus *Canis* using mitochondrial DNA sequences (mtDNA). See Section 2 for more details-

1.2. Mitochondrial DNA

The mitochondrial dna is maternally inherited (21–23) being as consequence a haploid genome. The mammalian mtDNA has a unique genetic code different than nuclear genetic code (23), and is composed by the two ribosomal RNAs (12 and 16S), 22 tRNAs and 13 genes from the family OXPHOS related to phosphorylation: ND1-6 and ND4L are part of complex I, CYTB of complex III, COX1-3 of complex IV and finally, ATP6 and ATP8 of complex V (24,25). The mammalian mtDNA also contains a control region (D-loop), which is though to evolve neutrally and is composed of promoters and a hypervariable region. This control region is widely used in phylogenetic approaches (16,26,27). The length of the sequence of the domestic dog is thought to be 16727 base pairs (28,29) its control region is composed of about 1300 base pairs containing a repetitive region formed by small tandem repeats, as observed in the last dog genome assembly, *CanFam 3.1* (29).

Due to the haploid condition of the mitochondrion, deleterious mutations can accumulate due to the lack of recombination. Therefore, it suffers such a strong selective pressure that its genes are progressively lost. Those genes have been migrating to the nucleus where sexual reproduction and recombination mechanisms are available allowing a protection against deleterious mutations (30). This handicap is also a key aspect why mtDNA is highly used in evolutionary studies: it has a high mutational rate representing a good genetic marker for these studies (31).

1.3. mtArchitect

MtDNA sequences are reconstructed by two different technologies. Traditionally, it has been reconstructed by long-range PCR protocols. However, due to its interest in many fields such as the field of evolutionary biology, it became necessary to develop bioinformatic tools in order to do fast, cheap and feasible reconstructions, making it easier to have larger datasets of

mitochondrial sequences. Thus, scientists have taken advantage of newly technologies such as Whole Shotgun Sequencing (WGS), which is a High-throughput sequencing (HTS) method, to develop those bioinformatics tools. However, when dealing with WGS data, we are faced with the problem of separating the reads that belong to mtDNA from those that belong to the nucleus.

An in-house mitochondrial sequence reconstruction pipeline was published on Lobon et al. 2016 (32). The tool was developed to accurately reconstruct mitochondrial sequences tested in samples from the genus *Pan*. However, testing the tool on canid samples lead to some limitations of a different nature which had to be fixed though a redesign of the pipeline, incorporating new software, changing parameters, etc. Overall design and profound implementation of the pipeline is explained in Section 3.1.

1.4. Evolutionary analysis

In addition to redesigning the mtArchitect pipeline, we used it to reconstruct a large dataset containing almost all dog breeds as well as other canid species to conduct some evolutionary studies detailed in the conclusions. We hope that these analyses might shed light on some of the unresolved doubts about the evolutionary history of canids, and that they could be extrapolated to their historical relationship with humans (see Section 1.1 for details).

2. Objectives

1. Redesigning of an in-house mitocondrial DNA reconstruction pipeline. Assessing its performance and efficiency.

2. Constructing a large scale phylogenetic tree of dog breeds along with other canid species and analysing it.

3. Finding signatures of selection in the mitochondrial DNA of dogs.

4. Finding signatures of a relaxation of the selective constraint in dog mitochondrial DNA.

5. Estimating the time to the most recent common ancestor between dogs and wolves.

6. Estimating the effective population size during the domestication of dogs.

3. Methods

3.1. Mitochondrial DNA reconstruction

3.1.1. Background

The process of mitochondrial DNA (mtDNA) reconstruction from paired-end Whole Genome Shotgun (WGS) sequencing data, presents many challenges to be overcome: the proportion in which the mtDNA is found in the samples, the ability to distinguish between nuclear and mitochondrial DNA, the balance between the discovery of new variants and the accuracy of the resulting sequences, etc. All of these can be critical factors to the process of quality mtDNA sequence reconstruction.

Our pipeline tries to find a compromise between the conservative and the inclusive approaches to mtDNA reconstruction by featuring two stages, one where an improved, sample-specific reference is created by iteration and the second where the definitive sequence is assembled 'de novo' using the updated reference (Figure 3.1):

- First, the WGS data is aligned to a reference mtDNA sequence with very loose stringency parameters to create a starting read pool. In this case, all the paired-end reads that map to the mitochondria -including those where one end maps to the reference but the other one does not- are kept. Then an iterative process is started where the reference sequence is aligned to the read pool with the same – and in the three last iterations more stringent- parameters and variants are called. The resulting sequence is used as a template for a new iteration round until no more variants can be called and the updated reference is optimal. In order to account for mtDNA circularity, the process is mimicked using an 8 kbp origin-shifted reference.
- Second, the WGS data is mapped to the updated reference sequence and a *'de novo'* assembler is used to produce the final reconstruction. Since the sequence coverage is normally too high for the software to work correctly, the reads have to be sub-sampled to the correct coverage and, again, an iterative approach has to be taken where, for each iteration, a different subset of reads is used by the assembler. A total of 40 reconstructions are assembled at the end of the process. The resulting contigs of all assemblies are aligned together and their consensus sequence is regarded as the optimal reconstruction.

This method has been proven successful in reconstructing 'de novo' mtDNA sequences even when using very distant reference genomes. It shows an accuracy rate bigger than 99.9% (less

than 20 errors per assembly). MtArchitect can be executed on command line of UNIX operating systems.

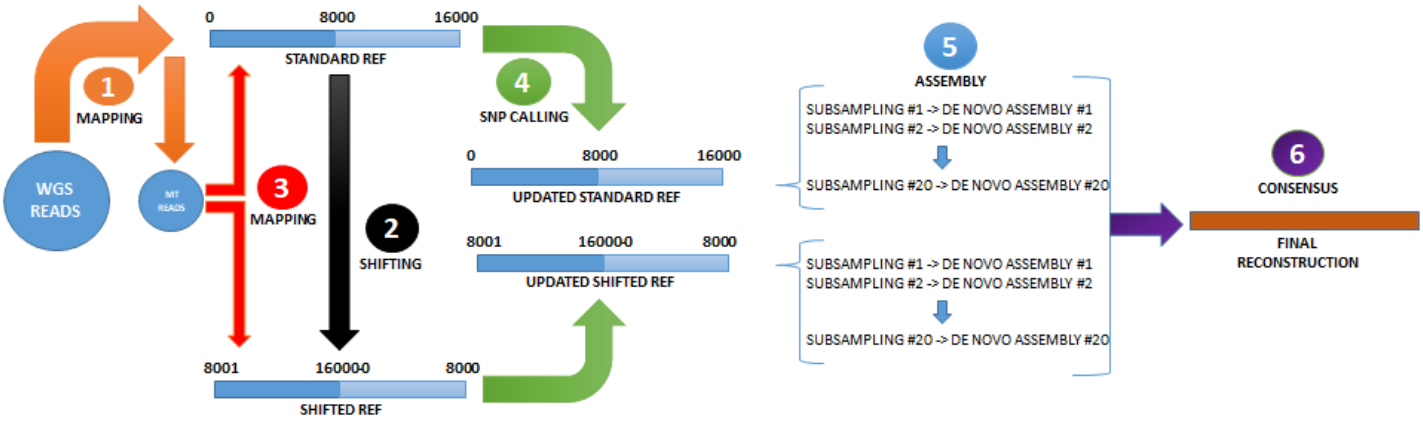


Figure 3.1. mtArchitect Overview. 1) Whole genome sequencing reads are mapped to a standard mitochondrial reference sequence with low stringency parameters to retrieve mitochondrial reads. 2) The standard reference sequence is shifted 8 kb so that the highly polymorphic D-loop is centred and more reads covering it can be retrieved. 3) and 4) The mitochondrial reads are mapped with regular parameters to the reference as well as the shifted reference and then SNPs are called and incorporated into the reference, creating a new specific sequence (updated standard reference sequence and updated shifted reference sequence). This step is iterative in order that the newly incorporated SNPs favour the mapping of more reads at each iteration. 5) All whole genome-sequencing reads are mapped to both modified references. 6) The final set of reads is subsampled approximately to 150x and a de novo assembly is performed 20 times for each modified reference. (7) The final sequence is constructed from the consensus of the 40 assemblies. Scheme adapted from Lobon et al. 2016 (32).

3.1.2. Implementation

The Burrows-Wheeler Aligner (BWA) software (33), in parcticular bwa mem (34), was used to perform the WGS mapping. Briefly, bwa mem takes into consideration reads that map to the reference sequence with a minimum seed length, in other words, only reads with a minimum of consecutive matches will be mapped (-k parameter). Many other parameters are taken into consideration, highlighting the next key parameters for the whole mapping process in the pipeline in the table 3.X.

3.1.2.1. Lax mapping stage

In the initial stage (lax mapping stage), WGS reads are mapped to the mtDNA reference sequence of dogs (*CanFam 3.1.*) (29). Different lax parameters were chosen depending on the evolutive distance between dogs and the corresponding sample to be mapped against, as distant

species will have a more divergent sequence than closely related species and thus more mismatches and/or insertions and deletions (indels) will be present. In fact, it was strictly necessary to give looser parameters to very distant species and more conservative parameters to closely related ones. Otherwise, reads that came from those distant samples would not have been incorporated to the starting read pool due to the high divergence with respect to the reference sequence. On the other hand, using loose parameters to map closely related species would have incorporated many nuclear reads to the starting read pool which might incorporate false variants to the modified sequence during the iterative stage.

Table 3.1. BWA mem key parameters.

Parameter	-A	-B	-O	-E	-L	-T
Meaning	Matching score	Mismatch penalti	Gap open penalty	Gap extension penalti	Clipping penalty	Minimum score to output
Default	1	4	6	1	5	30

Thus, closely related species, such as coyotes, golden jackals, grey wolves, etc., were given a concrete lax parameter (bwa mem -A 4), while distant species such as hunting dogs or foxes were given even more permisive parameters (bwa mem –A 6 –B 2). Paired-end reads with only one of the reads mapped were also retrieved and included in the set using SAMtools (35) view tag parameters. Firstly, all secondary, supplementary and unmapped reads were removed retaining only mapped reads (samtools view –F 2308) into a read pool set. Secondly, unmapped reads (samtools view –f 4) were put into another pool and then, unmapped reads whose read pair maps were rescued filtering them by their read barcodes. PCR duplicates were also removed (samtools rmdup). After that, reads were quality trimmed and we removed paired reads with a median quality base below 26 using Perl.

3.1.2.2. Iterative mapping stage

For each sample, the resulting BAM files were used to perform the sample-specific reference upgrading by iteratively re-mapping to the reference up to 10 times (bwa mem –A 3 –O 7 –L 10 for closely related species and –A 4 –B 2 –O 7 –L 10 for distant species) and calling variants with (samtools mpileup -L 1000 -d 1000 -I |bcftools –V indels –ploidy 1 –c | vcf-consensus(36)). Specially important is the increase of clipping penalty (bwa mem -L) during mapping to the reference genome. It is plausible that reads could have some mismatches at their 5’ and 3’ ends which causes BWA to softclip those ends (they are ‘invisible’ to the aligment

and so are anotated in CIGAR code created by SAMTools) due to the high concentration of mismatches, avoiding counting those nucleotides during variant calling and thus missing important variants which should be anotated. By increasing clipping penalty, BWA would prefer to align 5' and 3' end mismatches instead of clipping them. In addition to that, during those iterations, the variant calling does not incorporate indels, as it has been reported that samtools mpileup does not call indel variants effectively (it displays low sensitivity and specificity values as shown by Laurie et al, 2016 (37)). Three last iterations are performed with stricter parameters (bwa mem [default parameters], bwa mem –B 6 –O 8 –E 4 –L 20 and bwa mem –B 8 –L 10). During the last two iterations another variant caller is used, in this case Freebayes(38) (freebayes –standard-filters --min-coverage 5 -read-max-mismatch-fraction 0.30 --min-alternate-fraction 0.4 [Require at least this fraction of observations supporting an alternate allele in the in order to evaluate the position] --ploidy 1), which incorporates all kind of variants very accurately. The chosen parameters act as filters to avoid variant calling from reads that map on multiple locations (mapping quality 0), and those that have a very large quantity of mismatches so that they are suspected to be nuclear reads. Last but not least, the read pool which is used in the mapping process at the last iteration is made with only the reads whose mates fully map to the last modified sequence, removing read pairs where only one mate maps. The same process is done with a 8 kbp origin-shifted reference. This way, only well-supported variants are introduced into the reference, preventing a non-ending cycle of variants being called at the same position as well as the inclusion of nuclear mitochondrial DNA segments (NUMTs) into the final sequence. Eventually, we will have a modified reference sequence with accurately incorporated variants that will be used during the assembly stage.

For a better comprehension of the steps, a pseudocode with all the steps and program used in the iterative stage is providen below (Code 3.1). Some programs and parameters have not been commented as their are less crucial. For example, among others, samtools sort is used to sort reads based on their first based leftmost mapping position, samtools index to index BAM (binary representation of a Sequence Alignment Map file (SAM files)) files, bgzip to compress VCF (Variant Call Format) files, and tabix to index VCF files.

Code 3.1. Part of Lax Mapping Stage.

```
#Lax mapping stage
i=1 #First iteration
##Strict mapping of fastq obtained in first step
${bwa} mem -A 4 -O 7 -L 10 -p [Reference Sequence] [Initial_Read_Pool.fastq] | ${samtools} view  -F 2316 -Su - | ${samtools} sort - -o [Output_${i}.bam] -T [Temporary_File.tmp]
${samtools} index [Output_${i}.bam]
#SNP calling, create consensus an align to standard ref
${samtools} mpileup -L 10000 -d 10000  -I -gf [Reference Sequence [Output_${i}.bam] | ${bcftools} call -V indels --ploidy 1  -c - | ${mtArchitect}/Parse_Homozygous.pl | ${bgzip} > [output_{i}_vcf.gz]
tabix -p vcf $ItPath/iteration${i}_${Sample}.vcf.gz
cat [Reference Sequence] | ${vcf}/vcf-consensus [output_{i}_vcf.gz] > [output_{i}_MT.fasta] #created updated reference sequence in iteration i
${bwa} index [output_{i}_MT.fasta]

##Iterations 2 to 10
for i in $(seq 2 1 10)
do
    k=$((i-1))
    ${bwa} mem -A 4 -O 7 -L 10 -p [output_{k}_MT.fasta] [Initial_Read_Pool.fastq]| ${samtools} view  -F 2316 -Su - | ${samtools} sort - -o [Output_${i}.bam] -T [Temporary_File.tmp]
    ${samtools} index [Output_${i}.bam]
    ${samtools} mpileup -L 10000 -d 10000  -I -gf [output_{k}_MT.fasta] ][Output_${i}.bam] | ${bcftools} call -V indels --ploidy 1  -c - | ${mtArchitect}/Parse_Homozygous.pl | ${bgzip} > [output_{i}_vcf.gz]
    ${tabix} -p vcf $ItPath/iteration${i}_${Sample}.vcf.gz
    cat [Reference Sequence] | ${vcf}/vcf-consensus [output_{i}_vcf.gz] > [output_{i}_MT.fasta] #created updated sequence in iteration ${i} based on updated sequence from last iteration
    ${bwa} index [output_{i}_MT.fasta]

## Three last with stricter bwa parameters:
#Iteration 1:
    # bwa mem [standard parameters]
# Iteration 2:
    # bwa mem –B 6 –O 8 –E 4 –L 20
    # freebayes –f [Last_updated_reference_sequence.fasta] –standard-filters --min-coverage 5 [require at least this coverage to count variant calling] -read-max-mismatch-fraction 0.30 [exclude read with more than 30% of mismatches with respect to its read length] --min-alternate-fraction 0.4 [Require at least this fraction of observations supporting an alternate allele in the in order to evaluate the position] --ploidy 1
# Iteration 3
    # Only with paired-end reads who have mapped to the last updated sequence, excluding those pairs whose mate has not map.
    # bwa mem –B 8 –L 10
    #freebayes [same parameters] -> We obtain our final updated sequence
```

3.1.2.3. Assembly stage

For the assembly stage, the starting read pool was mapped to the updated reference (bwa mem –L 10) and the coverage sampling was achieved by filtering a fix number of reads each time. The final set of reads is sub-sampled to have approximately 150-fold depth of coverage. Before making the sub-sampling, the initial mapped reads were divided in two separated read pools: a pool set for reads aligned in high coverage zones, and another for reads aligned in low coverage zones. Withough going into details, this was necessary in order to include in each subsampling the reads that map in low coverage zones. For example, a very high mean coverage (i.e. 2000x) would cause most of the reads covering areas of low coverage not to be subsampled. Briefly, BEDTools(39) suite was used to compute coverages and create a BED file (bedtools genomecov –d and –bga in distinct steps) with only those positions that have a coverage less than the global median coverage. Then, we used bedtools merge –i to merge intervals that fall under the coverage threshold (for example, a range of coordinates of 1-50 with a coverage of 20 and another of 51-60 with coverage 10, would be fused into a single interval of 1-60). In addition, we intersected and captured reads that either are overlapped a 50% by low coverage regions or that overlaps at least the 50% of a low coverage region (bedtools intersect –f 0.50 –F 0.50). Once the two read pools are separated, they are both subsampled at 150-fold depth of coverage and merged together.

Velvet (40) was the software of choice to perform the '*de novo*' assembly. Two Perl scripts were used to automatically choose the optimal parameters. First, VelvetK can estimate the best k-mer size to use for Velvet. It needs two inputs: the estimated genome size, and a read pool. Then, VelvetOptimiser was used to find the optimal parameters as it searches a supplied kmer value range for the optimum one, estimates the expected coverage and then searches for the optimum coverage cutoff and outputs the contigs that are product of the assembly. The kmer interval was set between the one gotten with Velvetk, and the median read length (Velvetoptimiser.pl –s [starting kmer obtained with Velvetk] –e [final kmer obtained from median read length] –f "-ShortPaired –fastq [subsampled read pool] –long –fasta [updated reference sequence]"). Apart from adding the subsampled reads to the assembly process, the updated reference sequence was also added as a long read to guide the assembly and build larger contigs (-long –fasta [updated reference sequence]). The subsampling process was repeated up to 20 times per updated reference sequence as well as per shifted updated reference and the resulting contigs obtained from all the subsamplings were aligned with the updated reference sequence with a R script retrieving a consensus sequence which was the final

reconstructed mitocondrial sequence. A pseudocode (Code 3.2) is shown below with Velvet part:

Code 3.2. Part of Assembly Stage. Subsampling and Velvet assembly.

```
#Assembly Stage. Velvet part
#Note: random_pe.pl is used to subsample reads. It needs a integer number (threshold) which will used to subsample reads up to that number
        for i in $(seq 1 1 20)
        do
            #Subsample fastq
            cat [highercoverage.fastq] | ${mtArchitect}/random_pe.pl [threshold] > [subsample_{i}.fastq]
            cat [lowercoverage.fastq] | ${mtArchitect}/random_pe.pl [threshold] >> [subsample_{i}.fastq]
            #Use velvet to create contigs
            contigSize=$(perl velvetk.pl --size 16727 --best [subsample_{i}.fastq])
        Readlength_subsample=$(cat [subsample_{i}.fastq] | sed -n '2~4p' | awk '{ print length($0); }' | awk -F : '{sum+=$1} END {printf("%d\n",sum/NR - 0.5)}') #Obtain median read length
        perl VelvetOptimiser.pl -s $( echo $((contigSize))) -e $(echo $((Readlength_subsample))) -f "-shortPaired -fastq [subsample_{i}.fastq] -long -fasta [Updated_reference_sequence.fasta]" -p "${Sample}_norm" -d "Assembly/${Sample}/norm/$i" #Obtain optimal parameters for velvet and executes it.
```

3.2. Datasets and annotation

Our mtDNA reconstructions were pooled and aligned using MAFFT (41) with default parameters except that missing nucleotides (Ns) were treated like another wildcard to allow better alignments. The alignment file was manually annotated in accordance to an NCBI-ENSEMBL consensus. A large dataset of 515 samples was drawn from the aligned data, both for exome statistics and for phylogenetic tree construction. See Table 3.2 for a summary of the dataset and data gathering.

Table 3.2. Summary table of the datasets employed in the analysis. * Comprised by many subspecies. ** Common foxes and south american fox.

Species	African Golden Wolves	Golden Jackals	Dogs	Wolves*	Red Wolves	Foxes**	Coyotes*	Ethiopian Wolves	Hunting Dogs	Dholes	Total
Thomas Gilbert's group	5	2	-	19	-	-	3	1	2	1	33
Wang et al. 2016 (42)	-	-	47	-	-	-	-	-	-	-	47
Wang et al. 2013 (43)	-	-	6	4	-	-	-	-	-	-	10
Freedman et al. 2014. (6)	-	1	1	4	-	-	-	-	-	-	6
Koepfli et al. 2015 (20)	1	-	-	-	-	-	-	-	-	-	1
Auton et al. 2013. (44)	-	-	9	-	-	1	-	-	-	-	10
Fan et al. 2016. (45)	-	-	-	9	-	-	-	-	-	-	9
Campana et al. 2016	-	-	-	-	-	-	-	-	2	-	2
vonHoldt et al. 2016 (46)	-	-	-	-	2	-	2	-	-	-	4
Unknown origin	-	-	373	16	-	2	1	-	-	1	393
Dataset	6	3	436	52	2	3	6	1	4	2	515

3.3. Neighbor Joining Tree

3.3.1. Background

Neighbor joining (NJ) (47) Tree method is an agglomerative approach to the problem of phylogenetic tree creation that has recently become quite popular due to its notorious time-performance relation. The NJ algorithm is based on iteratively pairing the closest neighboring taxa in a DNA difference (distance) matrix, and subsequently rearranging the aforementioned matrix so that it fits the newly paired taxa into one element. In a little more detail, three steps are taken for each iteration of the NJ algorithm:

- First the starting distance matrix is corrected for distances among all taxa; because for two neighbours to be grouped together not only is it important that they are similar to each other, but they must collectively be as different as possible from the rest of the taxa.

$$Q_{ij}=(n-2)d_{ij}-\sum_{k=1}^{n}d_{ik}-\sum_{k=1}^{n}d_{jk} \quad i,j\in\mathbb{N}\ ;\ n\geq 3$$

  Where Q_{ij} is the corrected matrix, n is the number of individuals and d_{ij} is the starting distance matrix (notice that d_{ik} and d_{jk} correspond to the summation of the distances from one matrix element to the rest). The neighbors that have the smallest value in the Q_{ij} matrix are joined together into a new node.

- Second, the branch lengths to the new node are calculated. The distance between the two neighbors is split in two branches using the least-squares approach and the taxa that differs the most from all the others will have a greater branch length.

$$\delta_{iu}=\frac{1}{2}d_{ij}+\frac{1}{2(n-2)}\left[\sum_{k=1}^{n}d_{ik}-\sum_{k=1}^{n}d_{jk}\right] \quad i,j\in\mathbb{N}\ ;\ n\geq 3$$

  Where δ_{iu} is the distance (branch length) from the taxa i to the newly created node u and the complementary distance δ_{ju} equals d_{ij} - δ_{iu}.

- Third, the distance matrix is reduced by one entry. To recalculate all the distances to the newly created element u, an averaging approach is taken for all the other k elements.

$$d'_{uk} = \frac{1}{2}(d_{ik} + d_{jk} - \delta_{iu} - \delta_{ju})$$

BIONJ slightly differs from the standard NJ algorithm (Saitou and Nei, 1987) (47) in the fact that the variance among distances is taken into account in the rearranging step of the algorithm. Instead of just averaging the distances to the new matrix element, a pondering constant is added which tries to cancel out the effects of disparity among the measures that are more disperse.

$$d'_{uk} = \lambda d_{ik} + (1-\lambda) d_{jk} - \lambda \delta_{iu} - (1-\lambda)\delta_{ju}$$

Where λ is the pondering constant and it corresponds to the minimization of the variance of the rearranged distance matrix:

$$\lambda = \frac{1}{2} + \frac{\sum_{i=3}^{n} (var(d_{jk}) - var(d_{ik}))}{2(n-2)var(d_{ij})}$$

And var(d_{ij}) can be approximated as $var(d_{ij}) \approx \frac{1}{n} d_{ij}$. Both var(d_{ik}) and var(d_{jk}) are extracted from a variance matrix which is calculated in the standard fashion through the corresponding rows and columns of the distance matrix. The variance matrix v_{ij} is also reduced in one element together with the distance matrix following the formula:

$$v'_{iu} = \lambda var(d_{ik}) + (1-\lambda) var(d_{jk}) - \lambda(1-\lambda) var(d_{ij})$$

In this way, the more consistent measures are given more importance in the construction of the tree and therefore sequences with variable substitution rates are accounted for (they have higher variances which results in their branch lengths being correctly underestimated).

Neighbor joining trees are based on empirical observations and make no assumption about the underlying evolutionary model of the DNA sequences, thus the reliability of the resulting phylogenetic trees depends solely on the quality of the starting data.

3.3.2. Implementation

An initial neighbor joining (NJ) tree was produced using the software BIONJ by Olivier Gascuel (48).

The initial obtained alignment fro MAFFT was transformed into an interleaved relaxed PHYLIP compatible format featuring sequence names of at most 50 characters long.

3.4. Maximum likelihood tree

3.4.1. Background

The maximum likelihood (ML) approach to phylogenetic tree construction (49) is, to date, considered the most reliable method available. The reason is that an exhaustive search over many tree topologies is carried out taking a wide number of models and parameters into consideration.

In short, the ML method applies a likelihood score to all the trees it evaluates based on how good they fit a substitution model given multiple parameters such as the branch lengths, mutation rates or the tree topology itself. Once the whole likelihood space is more or less evaluated, the tree topology which has the best score is considered to be the most reliable.

In order to evaluate a topology, a couple of assumptions have to be made about its underlying evolutionary model:

- Most importantly, the DNA is assumed to evolve following a memoryless, continuous-time Markov chain where each site is independent from every other. That means, each DNA position can have one in four states (A, C, T, G) and these states may vary over continuous time with a given probability at a given rate. The fact that there has/has not been a mutation in a position after a certain time does not influence the appearance of new mutations in either the same site or in any other site.
- The probability with which each position transitions from one state to another is given by a theoretical substitution model that can be lightly (e.g. Jukes-Cantor) or complexly (e.g. GTR+I+G) parametrized depending on the previous knowledge of the data. This model constitutes the basic criterion for the evaluation of tree topologies.
- The rate at which the model transitions from one state to another is often assumed to follow a Poisson distribution, with the mean term being the mutation rate.
- The Markov chain is considered to be time reversible. That means the probability to go from one state to another (e.g. A to G) is the same as the probability to move from the latter state to the former one (e.g. G to A). Because of this property, the branch lengths between two neighboring taxa and their corresponding node can be redistributed

without affecting the overall likelihood of the tree as far as they add up to a certain value. This allows for a readier branch length optimization.

- The Markov chain is considered to be irreducible, which means that all states can be accessed from one another and there is a probability for recursiveness, that is, a state transitioning from itself to itself (e.g C to C).

Standard ML algorithms evaluate a given topology by finding its optimal branch lengths. To do so, an iterative process is performed where for each iteration run, all sites of the DNA sequence are evaluated and the value for one single branch is estimated at a time through the following formula:

$$\left(1-e^{-v_i}\right)^{[r+1]}=\frac{1}{S}\sum_{x=1}^{S}\frac{\left(1-e^{-v_i}\right)^{[r]}\left[\sum_k \pi_k L_k^{(p)}\sum_l \pi_l L_l^{(q)}\right]_x}{\left(e^{-v_i}\right)^{[r]}\left[\sum_m \pi_m L_k^{(p)} L_l^{(q)}\right]_x+\left(1-e^{-v_i}\right)^{[r]}\left[\sum_k \pi_k L_k^{(p)}\sum_l \pi_l L_l^{(q)}\right]_x}$$

Where v_i is the branch length that needs to be estimated, r is the iteration counter (not an exponential operator), Π is the equilibrium frequency of each nucleotide (estimated via the substitution model), L^p and L^q are the likelihoods of the previous states for a given position, k and l are the current states of the system, m is the ancestral state of the system (if unknown, a sum over all possible states has to be done), x is the position which is being evaluated and S is the total number of positions. The value of v_i that is obtained in an iteration round is used in the next one until the algorithm converges to a reasonable number, then the resulting likelihood is used to calculate another branch length. Generally, a leaf-to-root approach is taken and the branch length for a pair of external neighboring taxa is estimated before moving to the internal levels of the tree, for which the nucleotide state is generally unknown. The formula for the likelihood in any node following the previous nomenclature is $L_s^m=\sum_s[(\sum_k P_{mk}(v_t)L_k^{(p)})(\sum_l P_{ml}(v_u)L_l^{(q)})]$

where $P_{ml}(v_i)$ is the probability of transitioning from m to l in v_i units of time and s are the possible states of ancestral node m.

Incidentally, a *brute force* search of all topologies and branch lengths from scratch would require at least O(n!nS) (linear-factorial) computational times depending on the number of taxa (n) and the sequence length (S), and would therefore be inviable for more than about thirty samples. This renders the evaluation of the whole topology space impossible for any reasonable ML software and calls for optimized strategies to reduce the tree space search.

Some of these strategies are the gradual introduction of taxa into the tree, the input of a starting topology or iteratively re-rooting the tree.

The phyML algorithm requires a starting tree topology input (e.g. BIONJ) which is to be optimized via maximum likelihood. In order to explore new topologies, local nearest neighbor interchanges (NNI) are performed, which involve switching the position of two non-equivalent neighboring branches and checking if that increases the overall likelihood of the tree; if there is a foreseeable increase, further local NNI is performed, otherwise NNI is performed in a different part of the tree. Since the tree is not built anew and the starting topology is bound to have good ML scores, a lot of computational time is spared through this method.

3.4.2. Implementation

A maximum likelihood tree was built using the software phyML by Guindon et al (50). The previously generated BIONJ tree was used as a template for the program to work with and all the parameters were set to those yielded by SMS (Section 2.6.). The tree was constructed using the web online tool (http://www.atgc-montpellier.fr/phyml/).

3.5. Model testing with SMS

Phylogenetic programs serve the users with many variety of models to represent the nucleotide variability as well as the substituion process. Smart Model Selection (SMS) (51) is a software designed by Gascuel et al. to select the best model substitution based on likelihood-based criteria (e.g., AIC or BIC). It is implemented in the PhyML webserver and It works by finding the best phylogenetic tree each model can produce and comparing their likelihood scores. The model that produces the highest likelihood score is considered to be the one that best fits the datasets at hand. For a model θ and a tree τθ constructed with that model, SMS finds $\theta_{optimal}$ = argmax -ln(L(τθ| θ)) using the Bayesian Information Criterion (BIC) or Akaike Information Criterion among other. Authors claim that the runtime of the program is half the consumed by JmodelTest2 (52).

3.6. Estimation of additional parameters of interest with BEAST2

3.6.1. Background

Since there is an extensive bibliography surrounding the topics of dog domestication and divergence times in the Canidae family, we sought to evaluate whether some additional parameters such as population sizes or most recent common ancestor times (tMRCA) could be estimated from our dataset given the information that has already been published. Additionally,

we also wanted to infer the population history of dogs based on our data and compare our results with other published results adressing that matter. Bayesian Evolutionary Analysis by Sampling Trees 2 (BEAST2) (53) was the software of choice to fit an initial structured tree (a tree with a known topology and a set of priors for all the missing parameters e.g. divergence times, mutation rates, site models, population models, etc) into a whole evolutionary framework via Markov Chain Monte Carlo (MCMC).

BEAST2 relies on the fact that all missing parameters contribute to the tree prior in one way or another, so for each link of the MCMC, BEAST2 proposes some values for the missing parameters (sampled from their prior distributions) and does some minor changes in the tree topology. If these operations improve the posterior likelihood score of the tree by a significant amount, the new parameters and topology are kept, while if there is no improvement to the likelihood, the former topology and parameters are used. This is, in fact, an application of the Metropolis-Hastings algorithm, where the missing parameters would be the unknown distributions that need to be approximated, the minor topology operations and priors would contribute to the proposal function and the posterior of the tree would be the acceptance criterion. All of that is depicted in the pseudoformula: $P(\tau^*\mu^*\rho^* N^* | \tau\vartheta SN\mu\rho) \propto P(\tau | S\vartheta\mu)\, P(\tau | \rho N)$ in which the distribution of the missing parameters μ* (mutation rate), ρ* (population growth rate), N* (starting population size) and τ* (optimal tree topology) , $P(\tau^*\mu^*\rho^* N^* | \tau SN\mu\rho)$, is related to the probability of a tree likelihood (τ) given the data (S), a substitution model (θ) and a mutation rate prior (μ), $P(\tau | S\vartheta\mu)$; and the tree likelihood can also be explained by the coalescent theory which involves a population growth rate prior (ρ) and an initial population size prior (N), $P(\tau | \rho N)$.

3.6.2. Implementation

Two beast files were generated with the assistance software BEAUti (54):

- The first one contained four partitions (D-loop, non-coding regions, coding regions [first and second nucleotides of codons] and coding regions [third nucleotides of codons] along with many priors according to the different divergence events that were contained in our dataset to calculate the MRCAs (Dog-Wolf, Dog-Wolf-African Golden Wolf, Dog-Wolf-African Golden Wolf-Jackals, Dog-Wolf-African Golden Wolf-Jackals-Ethiopian Wolf, Dog-Wolf-African Golden Wolf-Jackals-Ethiopian Wolf-Coyote, Dog-Wolf-African Golden Wolf-Jackals-Ethiopian Wolf-Coyote–Andean Fox and Dog-Wolf-African Golden Wolf-Jackals-Ethiopian Wolf-Coyote-South American

Fox-Foxes). The thee prior was set to default (Yule model). A uniform wide prior was proposed for each of the divergence events with standard deviation equal to half the proposed mean value. The molecular clock was set to Strict Clock with a initial rate of 0.01 substitutions/site/Mya (average mtDNA substitution rate in dogs). Based on best model selection by SMS (section 2.6), the Model Site was initially set to:

- Substitution model: Generalised Time Reversible (GTR). The relative rate parameters was put to be calculated empirically while initial GTR relative rate parameters were set to:
  - A <-> C 0.42749
  - A <-> G 287.94976
  - A <-> T 0.52998
  - C <-> G 0.67350
  - C <-> T 5.66567
  - G <-> T 1.00000
- Gamma categories: 4
- Gamma shape parameter: 0.277
- A fixed substitution rate.

- The second file contained the same partitions and had a strict molecular clock with value 0.01 substitutions/site/Myr (average mtDNA substitution rate in dogs). The tree prior was set to bayesian skyline analysis (55) to infer the population history. Model and Clock site parameters were set to the values obtained from the first run of the first file.

Every file was run independently with a MCMC of 30 and 40 millions respectively. Results were summarized with Tracer 1.6 (56).

3.7. Evolutionary statistics

3.7.1. Background

3.7.1.1. Non-synonymous/Synonymous ratio

The non-synonymous/synonymous ratio (dn/ds) is the method of choice when studying possible changes in the selective pressure of a given nucleotide sequence, that means, when there is an unexpected amount of substitutions that have a biological repercussion. In other words, it estimates how much non-neutral (non-synonymous) evolution has happened relative to neutral (synonymous evolution). Thus, if the mtDNA is evolving neutrally, the dn/ds will be equal to 1, that is, there is no selection to non-synonymous changes. However, if dn/ds is below

1, then purifying selection is acting on mtDNA avoiding non-synonimous changes while if dn/ds is above 1, then there is a relaxation on selection and multiple non-synonymous changes are happening favored by natural selection. Given the non-recombinant nature of mitochondrial DNA, all mutations have a tendency to accumulate and cannot be reversed or diluted (Muller's ratchet principle), therefore, the mitochondrion is under a basal, strong purifying selection. Non-synonymous mutations imply a change in the amino-acid sequence of the resulting gene product and, as a consequence, are potentially deleterious. Given the assumption that the mitochondrion is under a strong purifying selection (see Section 1.2 for more details), it is expected that dn/ds for mitochondrial DNA is below 1 for each species, as most mutations will be synonymous changes.

There are two main approaches when it comes to the calculation of the dn/ds ratio:

- The ratio can be calculated through maximum likelihood methods by providing the model with a codon substitution matrix and looking for the dn/ds value that best fits the input data.

- The ratio can also be calculated by performing a pairwise comparison of the sequences, keeping count of the differences in synonymous and non synonymous positions and correcting for the number of synonymous and non-synonymous sites respectively.

Given a big enough sample size, both methods should yield similar results, although straight counting of the differences might tend to underestimate the number of synonymous sites and buff the overall dn/ds ratio.

3.7.1.2. McDonald-Kreitman Test

As we have discussed in the section 2.7.1.1, the Non-synonymous/Synonymous ratio can be used to search for natural selection in the mitocondria but has a big limitation as it can mix up constrained regions with adapted evolutive regions, understerminating, overstimating, or even faking the ratio to 1 by producing too many false negatives. As a consequence, dn/ds tends to be a very conservative test. The McDonald-Kreitman test (57) is based on the assumption that neutral theory predicts that the ratio of non-synonymous to synonymous changes should be constant through time, as most non-deleterious, non-synonymous changes tend to be neutral because they change aminoacids with the same properties as the older ones as they fix with the same probability than synonymous changes. This theory arrives at the conclusion that the dn/ds

ratio among individuals within species should be equal the to ratio observed between species if genes or regions are evolving neutrally. The McDonald-Kreitman test evaluates if the non-synonymous/synonymoys ratio is constant within and between species contransting “present” (within) with “historical” (between) changes. Thus, two sets of sequences are needed to identify if a variable nucleotide sites have:

- Non-synonymous differences within species.
- Synonymous differences within species.
- Non-synonymous differences between species.
- Synonymous differences between species.

The test computes a 2x2 contingency table (Table 3.3) where, in the rows, we have the non-synonymous changes and synonymous changes. Synonymous changes and non-synonymous changes that happen within species are separated from those that happen only between species in columns.

Table 3.3. McDonald-Kreitman Test contingency table scheme.

Type of mutation	Between species	within species
Non-synonymous	A	B
Synonymous	C	D

- If all non-synonymous changes are neutral, expect A/C = B/D. We can not reject neutrality.
- If some non-synonymous changes between species are advantageous and selected, expect A/C > B/D
- If non-synonymous changes does not fix because of their deleterious potential, expect A/C < B/D.

3.7.1.3. D-Statistics

D-statics are methods to intuitively know whether a sample is evolving under neutral conditions or there is some kind of selection based on the distribution of differences between DNA sequences. Tajima's D equation (58) measures the deviation between the expected variation of a sample set and its observed variation. This statistic has been proven to be beta distributed, so a test for significance can be performed:

$$D = \frac{\Pi - \frac{S}{\sum_{i=1}^{n}(\frac{1}{i})}}{\sqrt{var(\Pi - \frac{S}{\sum_{i=1}^{n}(\frac{1}{i})})}}$$

Where π is the average pairwise number of differences, $\frac{S}{\sum_{i=1}^{n}(\frac{1}{i})}$ is the harmonic mean of the number of segregating sites (n being the number of samples), and everything is normalized by the square root of the variance of the deviation measures. Negative values of Tajima's D mean that the pairwise number of differences is smaller than the expected variation. That is usually the case of populations where the minor alleles are found at very low frequencies and most of the variation is private.

Negative values of Tajima's D tend to be a consequence of selective sweeps where most of the variation has been washed away and the individuals are highly invariant for the region of interest. If a selective sweep is detected, that means that the resulting population is probably expanding.

Positive values of Tajima's D involve a somehow structured kind of variation: in short, the population is divided into groups, where an individual belonging to a given subset is very similar to the ones in the same subset but at the same time very different to all the other groups. Biologically, that can either mean that there is an excess of heterozygous individuals, which could indicate balanced selection, or that some individuals are disappearing and the population is progressively becoming structured (population contraction).

Incidentally, there are some other complementary ways to assess the presence of selection in a given sample. That is the case of Fu and Li's D* (59), which compares the expected number of singletons in a sample to the observed one:

$$D^* = \frac{S - \frac{\eta}{\sum_{i=1}^{n}(\frac{1}{i})}}{\sqrt{var(S - \frac{\eta}{\sum_{i=1}^{n}(\frac{1}{i})})}}$$

Where S is the average number of segregating sites and $\frac{\eta}{\sum_{i=1}^{n}(\frac{1}{i})}$ is the harmonic mean of the number of singletons (*η)*. A big number of singletons will cause *D** to be negative and, in that case, the meaning of Tajima's D will hold; however, a positive value of Fu and Li's D* does not necessarily relate to that of Tajima's D. Fu and Li's statistic can give a little more insight into the scenarios where Tajima's D is negative.

3.7.2. Implementation

The dn/ds ratio, Tajima's D and Fu and Li's D* and McDonald Kreitman test were determined using the software for evolutionary research DNAsp versión 6 by Rozas et al(4) . All individuals were separated into species and all settings were adjusted according to the dataset before performing any operations. Coding regions were annotated in the program to differentiate between coding and non-coding regions. The dn/ds ratio was calculated via mere counting while both Tajima's D and Fu and Li's D* were calculated in the standard way. No maximum likelihood estimates of these parameters were used for this project. McDonald Kreitman was calculated comparing two sets of species not too much similar nor divergent to avoid loss of information.

4. Results

4.1. MtDNA reconstruction

After the execution of our pipeline, 515 samples out of 519 were succefully reconstructed. All the failed reconstructions belonged to dog tumor samples whose quality reads were not good enough. In addition, their reads were so different to the reference sequence that they seemed to be rather spurious reads (nuclear reads from mutated sequences) than mitocondrial reads, which made variant calling imposible to work as well as assembly stage.

The coverage measures after running the first stage of the mtDNA reconstruction pipeline were expected to correlate with the coverages of subsequent assembly stage and showed a very heterogeneous distribution. This heterogeneity was intrinsic of the initial data and could not be easily dealt with, however, the median per-sample coverage was considered to be more than enough to ensure the success of the reference upgrading process.

No coverage data from the assembly stage was retrieved, but the validity of the sequence reconstructions was assessed during all the subsequent tests and analyses.

However, as introduced in Section 1.2, the mtDNA from canids has a large repetitive region in the control region (D-loop) formed by short tandem repeats (GTACACGT(A/G)C) (ref. bp 16130-161430). It is important to highlight that this repetitive region generates two main problems for short paired-end reads.

First, most of the paired-end reads of the sequenced samples have a maximum length of 100 base pairs which is a very important constraint, since reads belonging to the repetitive region map multiple times along the repetitive region, specially in the center of the repetitive region (the reads that have a portion of their sequence outside the repetitive region either map to the start or the end of the repetitive region and do not map multiple times) as those reads are shorter than the repetitive region making imposible to call variants and making assembly harder to Velvet Assembler (that is why the updated reference sequence is also needed during assembly). Because of this limitation, variant calling at coordinates 16230-16330 (the center of the repetitive region) could not be performed correctly as reads with a map quality less than 30 were ignored and most of the reads inside that region mapped multiple times having as consequence a map quality less than 30 (Figure 4.1).

Second, another disadvantage of those short reads is that indels in the repetitive regions will encounter many problems to be annotated properly since, as we have remarked before,

that repetitive region is formed of small short tandem repeats (Figure 4.1). We would have many scenarios of deletion events that cannot be detected in a proper manner:

- Deletions and insertions of whole short tandem repeats (GTACACGT(A/G)C) would not be detected even in the 5' and 3' ends of the repetitive region since reads will have more of the same short tandem repeats and thus would map to the reference sequence without detecting it.
- Furthermore, depending on the penalty that we would have given to insertions and deletions during mapping with bwa mem, a true deletion shorter than the short tandem repeat (i.e. GTACACG or CG deletion) would be treated as a deletion (if the insertion penalty is greater than the deletion penalty) or as a false insertion (if the deletion penalty and gap extension penalty is greater than insertion penalty). The same happens with small insertions as they would be treated like an insertion or deletion depending on the given penalties. In addition, those false deletions/insertions would be added to the updated sequence in the iterative stage because of the multiple mapping of the reads (if those reads would count to variant calling).
- Lastly, deletions or insertions longer than the short tandem repeat (i.e. GTACACGT(A/G)CACA) would not be detected properly. The short tandem deletion/insertion will not be detected (look at first sceneario we have stated just before), while the other part (last ACA) will be treated as a deletion or insertion depending on the penalties.

The only reported cases in our samples are the Dhole samples (*Cuon alpinus)*. According to the reference genome of *Cuon alpinus* (NC_013445.1), which has 16672 base pairs, having used a dog reference genome instead of the Dhole's one resulted in a reconstructed sequence with more than 70 insertions in the repetitive region compared with the reference sequence of dholes. As explained before, what should be called as a deletion variant was called as an insertion variant, probably due to the harsher penalties to deletions during the mapping and multiple read mapping into different locations of the repetitive region also included false insertions (Figure 4.1). However, there is another sequence from *Cuon alpinus lepturus* (KF646248.1), a subspecies of dhole which has 16767 base pairs (95 base pairs more than Dhole's reference sequence being the majority of insertions in the repetitive region) suggesting that Dhole's reference sequence misses some nucleotides in the repetitive region. If we compare the reconstructed dhole's sequences with the last cited one, they have 20 base pairs less in the

repetitive region, which could indicate that the insertions were true except for the ones that were not detected (for example, short tandem insertions) or that were treated as deletions.

Using larger reads (i.e. of 250 base pairs) would fix some of these issues. First, although some reads would fall inside repetitive regions, many of them will have a big or small portion of their sequence outside the repetitive region which would act as an anchor that would allow reads to map correctly in their correct coordinates. In that case, mapping reads on the left and on the right ends of the repetitive region would cover all the region correctly and mismatches would be called correctly. For example, we reconstructed two Hunting dog samples whose reads had a 250 base pairs length and covered the repetitive region properly, so variant calling was done correctly in that region. Indeed, correct variants (mismatches) would be called, but deletions and insertions would still be hidden or would be called badly.

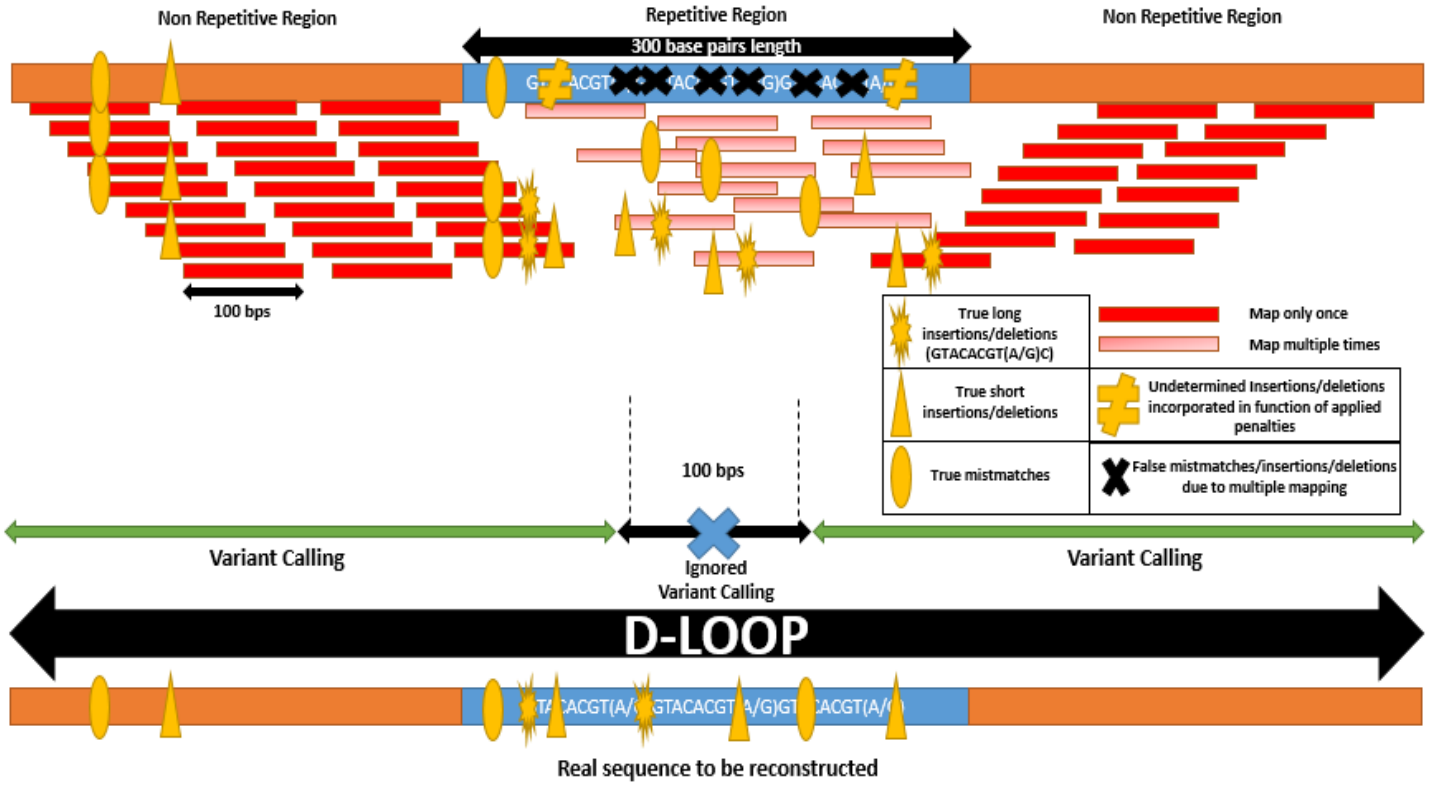


Figure 4.1. Mapping reads to canid mtDNA in the D-loop region. Reads of 100 base pair length that map entirely in non-repetitive regions or have a fragment that map there and another in the D-loop map only once (high quality mapping). However, reads that map inside the repetitive region without having any portion that map to the non-repetitive region will map multiple times as they are smaller than the repetitive region which have a length of 300 base pairs. Variant calling is imposible at the center of the repetitive region but not in its 5' and 3' end since there are reads (red) which map accurately thanks to the non-repetitive flanking regions. However, indels cannot be called accurately in the repetitive region and long indels wouldn't be detected as well.

We would need to use even longer reads which would be able to cover the entire length repetitive region (400 base pairs would be suitable) so that true insertions and deletions would be detected as well as mismatches, although there are few accesible technologies allowing these lengths. Another possibility would be the testing of a variant caller that targets the repetitive regions specifically. An interesting variant caller to be tested in the future is Sniper (60) whose authors claim to treat SNP discovery through a Bayesian probabilistic model enabling better variant discoveries.

4.2. Alignment

The alignment produced by MAFFT was analysed with MEGA 7 (61). The analysed alignment contained 165 gaps which the majority of them corresponded to gaps found in the D-loop. This observation is very consistent since our dataset was very large, and contained different species and the D-loop is a very heterogenous which varies even more between species (with many mismatches, insertions and deletions). Our alignment also makes sense since most of the gaps and mismatches found in non-coding and coding regions were shared within species (Figure 4.2) even in the D-loop (although more heterogeneity if found there). Mismatches were more variable inside dog breeds, which is also reasonable given the large number of breeds in the dataset but most of them were found on third codon position resulting in synonymous changes.

Surprisingly, a difference in the starting codon of gene *ND4L* was found between all of our factual sequences and the reference sequence (Figure 4.3). Determination on how this difference could affect translation remains unknown to us, but no report of it has ever been made. The insertion in the starting codon consist on a *TG* in dogs which creates a codon codifying for another methionine but displaces all the reading frame producing a stop codon after 3 codified aminoacids. However, the official codified sequence for *NDL4* starts with two methionines:

- (***M***S***M***VYINIFLAFILSLMGMLVYRSHLMSSLLCLEGMMLSLFVMMSVTILNNHLTLASMMPI VLLVFAACEAALGLSLLVMVSNTYGTDYVQNLNLLQC).

It could suggest that if the insertion is real (as it is present in all the sequences), the two first codons (including the insertion) would not be codons instead but would belong to non-coding sequences.

The coding region would start at the second Methionine so the sequence would be the next:

- (*~~MS~~**M**VYINIFLAFILSLMGMLVYRSHLMSSLLCLEGMMLSLFVMMSVTILNNHLTLASMMPIVL LVFAACEAALGLSLLVMVSNTYGTDYVQNLNLLQC*).

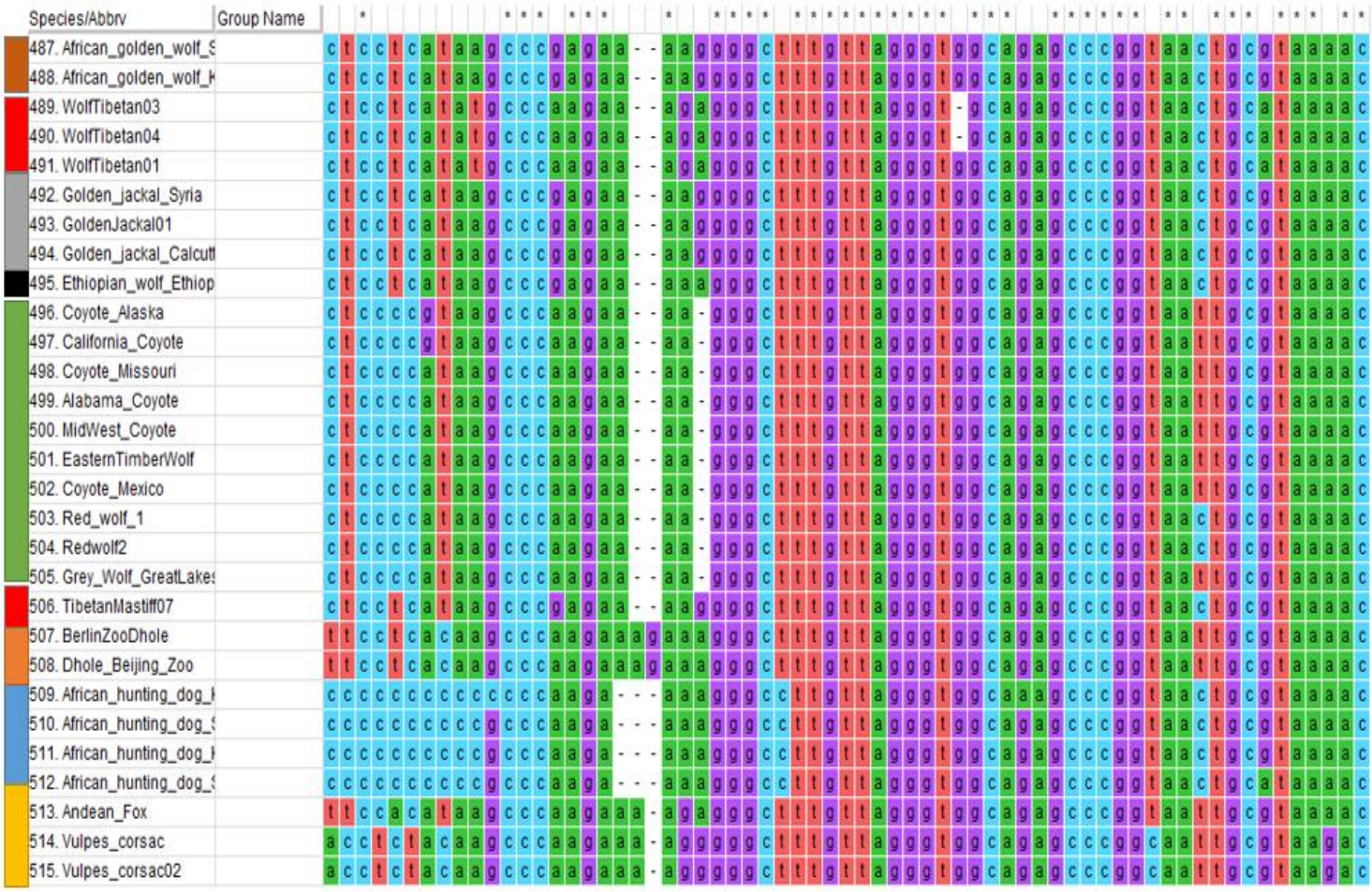


Figure 1.2. Multiple Sequence Alignment of some canid samples. Example of a 80 base pairs window fragment corresponding to a non-coding region whose product is the 16S ribosomal RNA. We can observe that most gaps and mismatches are shared within species. Yellow (south american fox (Lycalopex culpaeus) and two corsac foxes (Vulpes Corsac)), blue (african hunting dogs (Lycaon pictus)), orange (dholes (Cuon alpinus)), red (grey wolves (Canis lupus)), green (coyotes (Canis latrans), red wolves* (Canis rufus), eastern grey wolf* and Greak Lakes grey wolf*), black (ethiopian wolf (Canis simensis)), grey (eurasian golden jackals (Canis aureus)), Brown (african golden wolves (Canis anthus)). *Hybrids species.

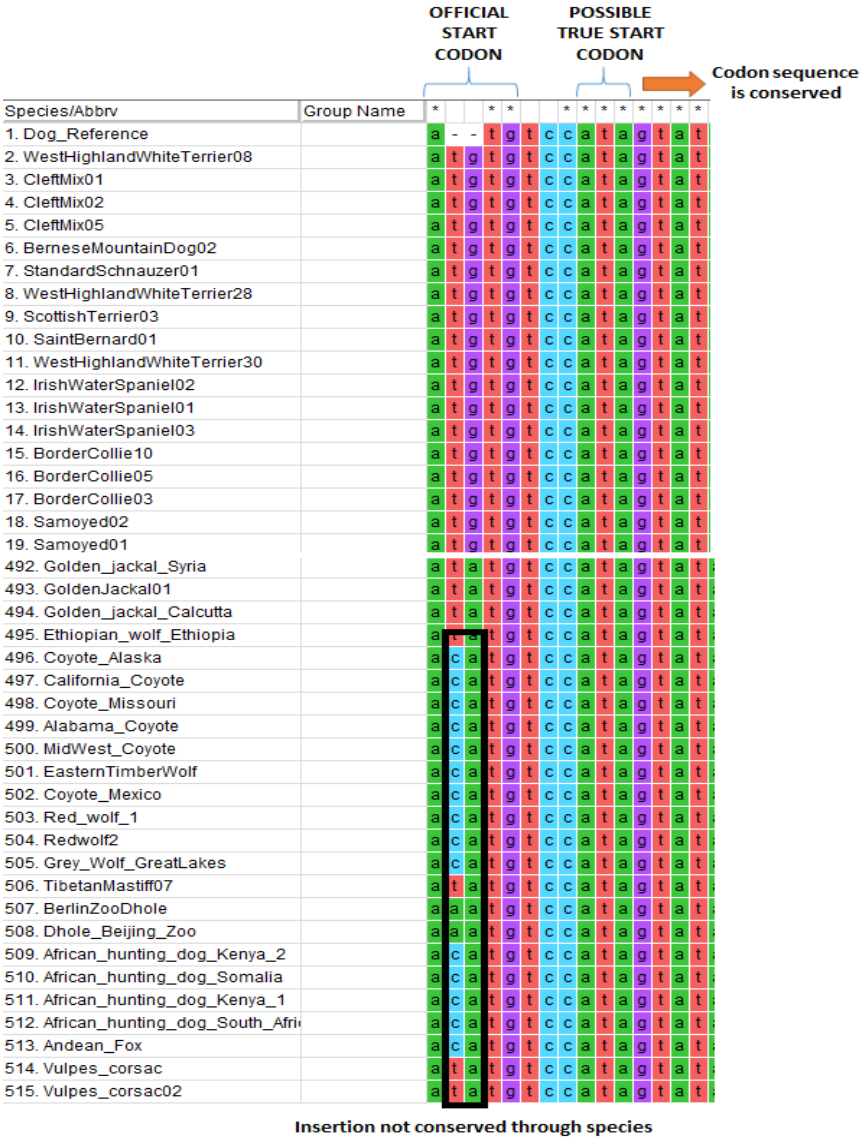


Figure 4.3. Multiple Sequence Alignment of some canid samples. We can see in the alignment that an insertion is present in all the samples compared to the dog reference sequence in the starting codon of NDL4, displacing the Reading frame and creating a stop codon (ATG-TGT-CCA-ATA: Met-Cys-Pro-Stop). Furthermore, we can see that some species incorpore another insertion that does not code for a start codon. We can see another start codon (ATA) which codes for a methionine after the insertion which could be the real start codon, as it is conserved in all the species and does not displaces the Reading frame.

4.3. Model test

As shown in Table 4.1, the complexity of the data was generally very high and there was a tendency to accept heavily parametrized models (columns omitted). The big sample size of our dataset resulted in some very large values of the negative log likelihood and Akaike's Information Criterion (AIC), but that did not seem to interfere with the decision making process implemented in the software.

The model substitution with the best likelihood according to AIC was Generalised Time Reversible (GTR) model with Gamma shape parameter and Invariant sites as decorations.

Table 4.1. Best model substitutions tested by SMS and based primarily in Akaike Information Criterion (AIC). Model denotes substitution model (Generalised Time Reversible (GTR) and Tamura Nei's 93 (TN93), Decoration refers to distinct parameters (Gamma shape parameter (G) and Invariant sites (I)).

Model	Decoration	K	Llk	AIC	BIC
GTR	+ G + I	1037	-97795,5452	197665,09	205688,012
TN93	+ G + I	1034	-97882,7794	197833,559	205833,271
GTR	+ G	1036	-98382,6601	198837,32	206852,505
GTR	+ I	1036	-104523,942	211119,884	219135,069
GTR		1035	-112921,361	227912,722	235920,17

4.4. Phylogenetic tree and clustering

A thorough analysis of the ML tree and its NJ counterpart found no big differences between the two.

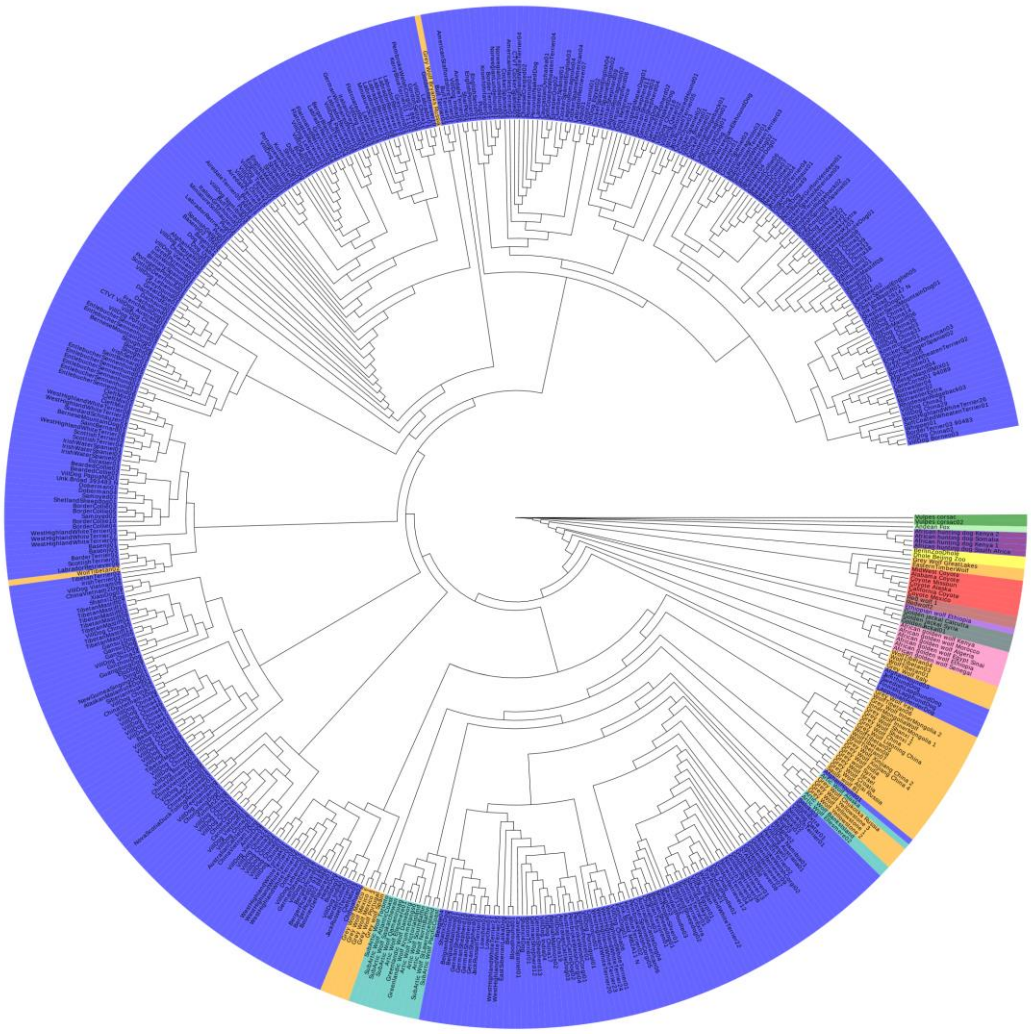


Figure 4.4. Scaled phylogenetic tree plots designed with EvolView (62)**.** Darkgreen: Foxes (Vulpes corsac), Lightgreen: Andean Fox (South American Fox, Lycalopex culpaeus), Violet: African Hunting Dogs (Lycaon pictus), Yellow: Dholes (Cuon alpinus), Red: Coyotes (Canis latrans), Brown: Red Wolves (Canis rufus), Lightviolet: Ethiopian Wolf (Canis simensis), Grey: Golden Jackals (Canis aureus), Pink: African Golden Wolves (Canis anthus), Orange: Grey Wolves (and subspecies, Canis lupus), Lightblue: Alaskan wolf (Canis lupus occidentalis, labelled as Arctic wolf Alaskan), Arctic (Canis lupus arctos), Subarctic (or Great Plains wolf, Canis lupus nubilus), and Greenland Wolves (Canis lupus orion), Blue: Dog breeds (Canis lupus familiaris).

As seen in Figure 4.4, foxes, hunting dogs, dholes, coyotes, the ethiopian wolf, golden jackals and african golden wolves cluster independently which is to be expected given that they both diverged from dogs more than 1 Mya. However, some appreciations must be explained:

Red wolves are clustered along with Coyotes following the line of the last publications which suggests that red wolves may be a hybrid species product of the hybridization between coyotes and grey wolves (63,64), although other scientist claim that red wolves are a distinct species which diverged from coyotes 150-300 kyr (65). On the first scenario, a coyote maternal inheritance would explain the closeness whereas the second scenario would reflect only the divergence due to the mtDNA evolution during time. The same happens with the eastern wolf and Great Lakes boreal wolf which both cluster together with coyotes and red wolves. Eastern Wolf (*Canis lupus lycaon* or *Canis lycaon*) has also been suggested to be product of introgression between grey wolves and coyotes (64,66), whereas like red wolves, other claims that they are distinct species from grey wolves, consistent with the idea that grey wolves and coyotes did not extend into the eastern United States (13,67). About Great Lakes boreal wolf it has even been hypothesized that they have emerged from introgession between eastern wolves, grey wolves, and coyotes (13).

Jamthund, Swedish Lapphund, Finnish Lapphund and Lapponian Herder dogs integrate with wolf branches, and form an independent clade with a Italian grey wolf. It has been suggested and published that those breeds emerged in post-domestication event from the hybridization of a female wolf with a male dog as they have a unique haplogroup (subclade d1) which is only present in Scandinavia (68,69). Last but not least, it is really curious that they cluster with Italian grey wolf. This wolf is a grey wolf subspecies (*Canis lupus italicus*) that does not share haplotypes with any other european grey wolf having a unique mitochondrial haplogroup (70,71). This observation would correspond to the last remaining european wolf conserving this haplogroup that was shared among ancient western european wolves, as Italian wolves shares this so called haplogroup 2 and cluster closer to ancient wolves from the Late Pleiostecene (72). It would be possible that *Sami* dogs (Jamthund, etc.) originated from a cross hybridization with a female ancient wolf of the haplogroup 2 which was later replaced by wolves with the haplogroup 1 some thousands of years ago. In addition, that ancient wolf has not been matched across Eurasia yet (73). An Afghan Hound is also found in this clade, although it may be due to a bad reconstruction or mislabelling since a secondone clusters with other wolves but the rest are clustered with other dog breeds and no reports have been made about a possible hybridization.

Interesting, we can encounter 3 clades cluster asiatic grey wolves: On the first hand, we can encounter Tibetan wolves (*Canis lupus filchneri*) that are grouped with other chinese wolves and Mongolian wolves (*Canis lupus chanco*). In addition, in an inner clade we find that the Indian (*Canis lupus pallipes*), Israeli and Syrian wolves (*Canis lupus pallipes* also) group together. They all are in close contact with the Croatian and Russian wolves from Altai Republic (both *Canis lupus lupus* or also called Eurasian wolf). Starting with tibetan and chinese grey wolves, looking at the tree they seem to be the most basal wolves in the tree (that is, the oldest wolves) which is supported by some publications that claim Tibetan wolves form two basal clades along with himalayan, indian, chinese and mongolian wolves indicating a common ancestor for all of them (15,74) and that dogs descended from grey wolves but not from tibetan wolves (75). Three tibetan wolves cluster independently way before the other wolves, which may indicate that those tibetan wolves live in isolated zones and that there may have been a closer contact between the tibetan wolves and the chinese wolves that cluster together with mongolian wolves. It is not surprising that Indian wolves from India, Israel and Syria are more divergent with tibetan, chinese and mongolian wolves, but belong to the same clade because as we have stated before, it is suggested that they are basal wolves and share a common ancestor with them. What is more surprising is the proximity between Croatian and Russian wolves with *Canis lupus pallipes* samples because they do not cluster with other eurasian wolves. A possible reason for this clustering may be an interaction between these wolves allowing gene flow between them due to their geographical proximity, between India, Israel, Syria, Croatia and Altai Republic (Russia).

On the other hand, we have a Saudi wolf (*Canis lupus arabs*) which clusters solely with an Afghan Hound (it may be a bad reconstruction). The observation that *Canis lupus arabs* does not cluster with *Canis lupus pallipes* may give some insight whether they are really the same species or not since some scientists suggest that there is no distinction between both (76) and others say that there is a physical and genetic distinction (77).

Furthermore, there are two major clades containing all “arctic” wolves (blue OTUs) with Grey wolves from Yellowstone and Chukotka Peninsula, which is the eastern end of Russia and is the closest point to Alaska from Eurasia. It makes sense that grey wolf from Chukotka and the alaskan wolf (*Canis lupus occidentalis*) form a inner clade because that strongly supports the theory that wolves crossed the Chukotka Peninsula to the Alaska Peninsula during different glacial eras during the Bering land bridge (78). Then, the ancestors of *Canis lupus occidentalis* crossed the Bering land bridge during the last glacial period. Before its ancestors crossed the

bridge, the ancestors of *Canis lupus nubilus* (subarctic wolves) did the same (13) invading North American (USA, Canada and Greenland) and, as the tree shows, it could be that arctic and greenlandic wolves (*Canis lupus arctos* and *Canis lupus orion)* descend from the same ancestor than *Canis lupus nubilus*. Last but not least, grey wolves from Yellowstone seem to be related to the same origin of *Canis lupus nubilus* and that *Canis lupus occidentalis* is closer to eurasian wolves.

The ancestors of Mexican wolves (*Canis lupus baileyi*) were likely to be basal to ancestors of *Canis lupus occidentalis* and *Canis lupus nubilus* and the first to cross the Bering land bridge and they form and independent clade separated from the rest of wolves which is also supported by a peer review publication (13).

Like the italian wolf, the spanish and portuguese wolf (Iberian wolves or *Canis lupus signatus*) form an independent clade diverging with the rest of wolves. A study based on the mitochondrial control region claims as well that Italian and Iberian wolves have the two most distinct haplotypes compared to the rest of european samples, clustering both independently and apart from the rest of samples (74).

Especially, there is an important discrepancy between our results and those of others. According to the phylogenetic tree, ethiopian wolves (*Canis simensis*), eurasian golden jackals (*Canis aureus*) and african golden wolves (*Canis anthus*), would be more evolutionarily close to dogs than coyotes, while reports based on nuclear dna suggest that they are basal to coyotes and thus more divergent (5,56)

Regarding dogs, just a few logical clustering patterns can be assigned. At first sight, with the exception of some local clusters, there does not seem to be a clear tendency for the dogs to group together by their breed or country of origin. At a very global level, some distinctions can be made regarding the overall provenance of some individuals (Asian dogs slightly tend to cluster apart from European dogs), and their divergence time (more ancient breeds tend to cluster apart from the modern ones); but even these findings are not without exceptions. Although shocking, these results were mostly expected given the undocumented provenience of almost all of our samples. In addition, two wolf samples (Wolf Tibetan 02 and Grey Wolf Bryanks Russia) clustered with dogs in inner branches. The most likely causes for this could be an unssuccesful reconstruction, bad qualities of the samples, mislabelling or undocumented introgression, which sounds the least possible theory.

It should be stated that the phenomena of partial integration and complete integration of dogs and wolves have both been described before (Elaine A. Ostrander and Robert K. Wayne, 2005; Savolainen et al, 1997; Larson et al, 2012) (79–81), and can mostly be attributed to the very recent divergence times of the two species.

We found no big indication of reconstruction bias in our samples due to the strict data filtering step during annotation. A bunch of samples were monitored with the IGV software (82) to check the quality of the alignment and posible bad mappings. The phylogenetic tree can be downloaded at Dropbox link and can be visualized on-line at http://etetoolkit.org/treeview/ or with any other downloadable tree viewer.

4.5. Parameter estimation

We observed that our Markov chains converged except some parameters that had very low ESS. Some interesting results regarding divergence times and population sizes in dogs could be drawn from them.

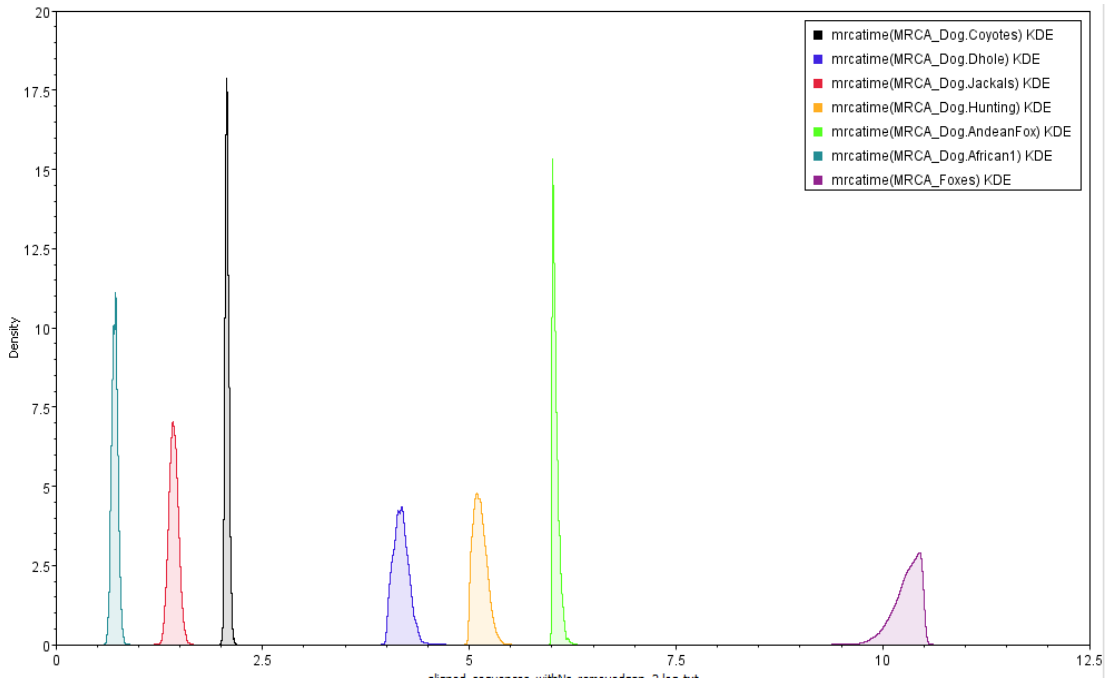


Figure 4.5. Distribution densities of the Time to the Most Recent Common Ancestor (TMRCA) for the partitions from the left to the right: Dog..African Golden Wolf (lightblue), Dog..Golden Jackal (red), Dog..Coyote (black), Dog..Dhole (dark blue), Dog..Hunting Dog (yellow), Dog..Andean Fox (green) and Dog..Fox (violet). Y-axis represents density values while X-axis represents time in million years.

The Figure 4.5 shows the TMRCAs. The divergence time between dogs and wolves and between dogs and the ethiopian wolf were not included. The height of their peak densities could

indicate a big restriction on the proposed range of time distribution since the MCMC chain has not varied during the iterations. On the other hand, the rest of partition seems to have converged during sampling with good ESSs (Table 4.2). The mean divergence times with respect to dogs are represented in Table 4.2.

Our results agrees with some of the published from those of Lindblad-Toh et al, 2005 (29). They published a maximum parsimony phylogenetic tree based on SNP data, where they suggest divergence times between dogs and andean fox (6-7.4 Myr versus 6-6.1 Myr), and foxes (9-10 Myr versus 9.98-10.5 Myr). On the other hand, Koepfli et al, 2015 (20) have differences and coincidences with our results. Their results based on nuclear data show a different divergence time between dogs and coyotes (0.81 – 1.42 Myr versus 2.02 – 2.12 Myr), dogs and dholes (2.15-3.38 Myr versus 4-4.33 Myr), dogs and hunting dogs (2.43-3.78 Myr versus 5-5.27 Myr) but coincide with dogs and eurasian golden jackals (1.50-2.38 Myr versus 1.31 – 1.51 Myr), and with foxes (7.20 – 10.28 versus 9.98-10.5). Further studies would be needed to assure the reliability of our results as well as from the others as results are quite different with some divergent times.

Table 4.2. Time to Most Recent Common Ancestor (units in million years). MRCA column shows values per prior sample, as well as its interval (95% HPD Interval), and the effective sample. ESSs under 200 are innadecuate.

Sample	MRCA (in Mya)	95% HPD Interval	ESS
Dog..African golden wolf	0.714	[0.6469-0.7864]	851
Dog..Eurasian golden jackal	1.426	[1.3158-1.5322]	1829
Dog..Coyote	2.072	[2.0285-2.1193]	1029
Dog..Dhole	4.716	[4.00175-4.3328]	1226
Dog..African hunting dog	5.132	[5-5.2782]	1281
Dog..Andean fox	6.045	[6-6.1225]	724
Dog..Foxes	10.288	[9.9824-10.5]	10212

As regards to population sizes, we were able to correctly date and identify the selective sweep in dogs due to domestication. As seen Figure 4.6, the population was estimated to be constant during all the time before the divergence of dogs and wolves. Also seen in the figure is the notorious population contraction around 30 kya, when the effective population size of the dataset was reduced more than 6-fold. That contraction can be easily attributable to a selective sweep related to the domestication of dogs. Our results match those in the bibliography

Thalmann O et al., (10), which believe the effective population size after domestication decreased dramatically. Interestingly, even the population expansion following the selective sweep can be observed in the graph.

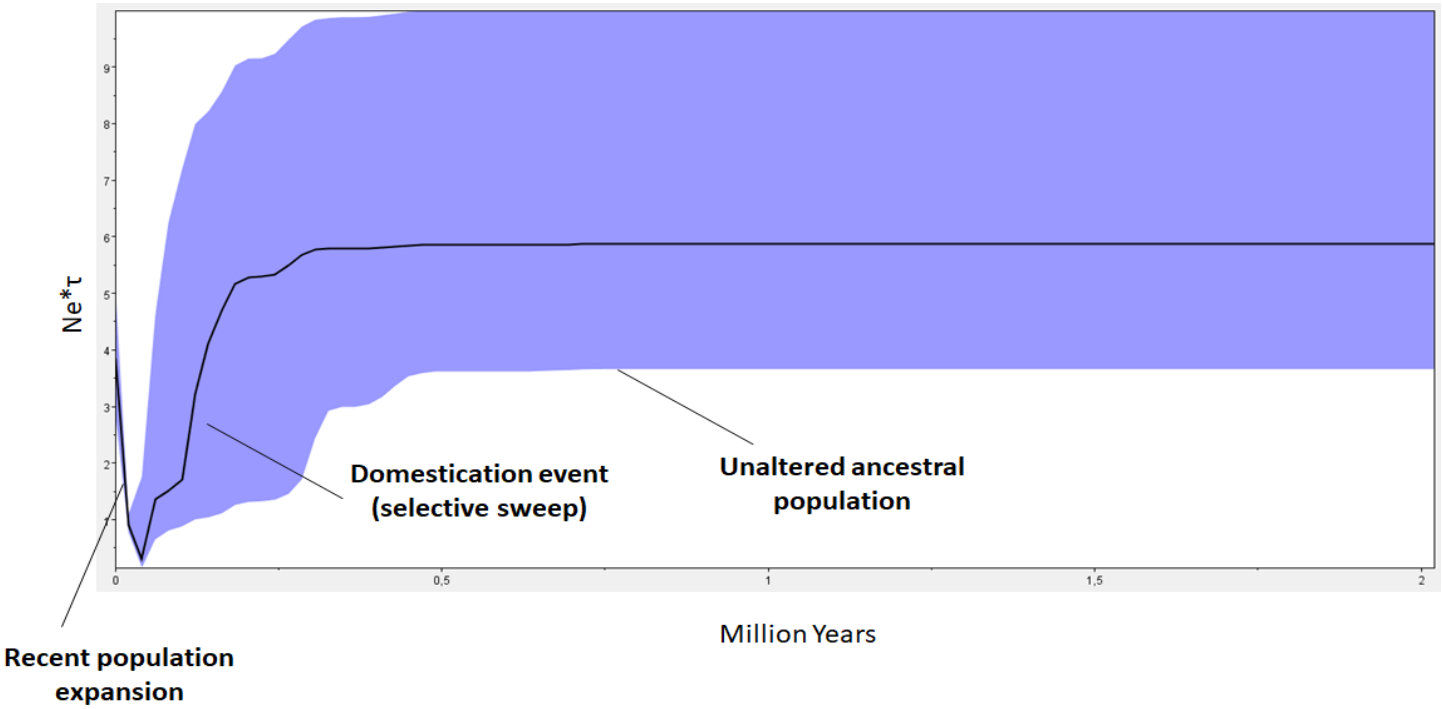


Figure 4.6. Bayesian skyline plot of the dataset population sizes against time. X-axis represent time in million years (Mya). Y-axis represent the results in $N_e * \tau$ units (where N_e equals population size and τ equals the number of generations in Myr units) (83). Solid intervals correspond to 95% HPD.

4.6. Relaxation of the selective constraint

Jackals, dholes and foxes had to be discarded during this analysis due to the lack of samples to analyse. A minimum number of four samples per species was set as threshold to perform the analysis. As previously described by Björnerfeldt et al, 2006 (84), dogs show a bigger dn/ds ratio than any other closely related species, showing a ratio of 0.088 (Figure 4.7). This could be explained by a possible relaxation in the selective pressure in the canine mitochondrion, as results show that dogs have been found to significantly outmatch both wolves, african golden wolves and coyotes. The dn/ds values in wolves seems to be relatively low despite the mix of the different subspecies inside the groups analysed. As wolves, african golden wolves, coyotes and hunting dogs have low dn/ds although their sample size was too small so the results may not be sufficiently reliable.

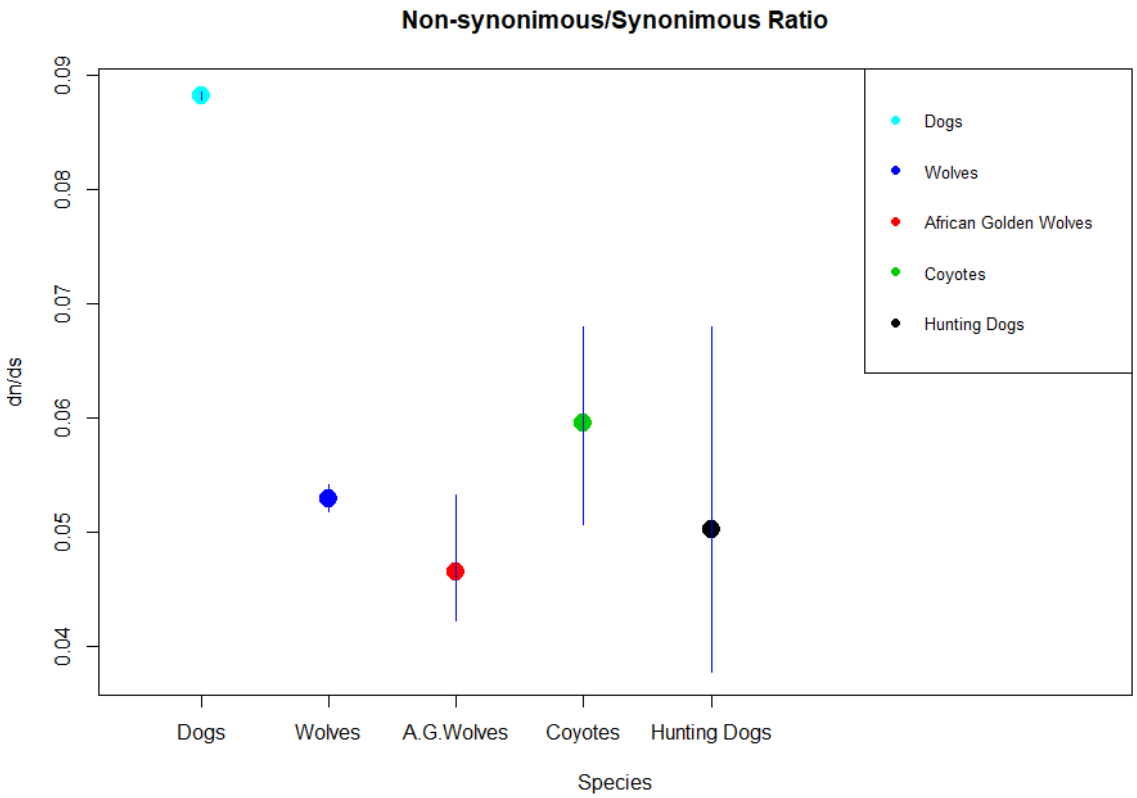


Figure 4.7. Plot of the bootstraped (sampling=1000) dn/ds means for each species. dogs (lightblue), wolves (blue), african golden wolves (red), coyotes (green), hunting dogs (black). Point indicates dn/ds results. Lines indicate a Interval Confidence of 95% (IC 95%).

With the samples analysed, coyotes seems to have a higher dn/ds than wolves dogs but that is mainly due to the structure of the working dataset which contains very few coyote samples with different subspecies and different geographical locations (*Canis latrans cagottis, Canis latrans clepticus*, *Canis latrans incolatus*, etc.). It is our guess that this data setup might have tilted the ratio towards higher numbers, although we would need to compare more samples to have solid conclusions (between subspecies and within subspecies). Nevertheless, all species except dogs seem to have a similar dn/ds which would reflect their feral way of life.

The ratio between non synonymous and synonimous substitutions has highlighted that all species, having a dn/ds near to 0 and thus reflecting a negative selective pressure agaisnt aminoacid changes in the mitochondrial genes. However, dogs show a higher dn/ds compared to the rest of species which could indicate a relaxation on purifying selection.

4.7. Selection

One of the main aims of our experiments was to confirm the presence of selection in dogs with respect to the other species. As for dn/ds test, some species were excluded due to the insufficient sample size (golden jackals, ethiopian wolf, dholes and foxes). As observed in

Table 4.3, dogs show a significant, negative value of Tajima's D, which is also maintained throughout the non-synonymous and even the synonymous parts of the exome. That means, supported by dn/ds results, that dogs have an excess of low frequency polymorphisms but show many stretches of low variability. It is important to remark that the level of significance holds when applying Fu and Li's calculation, which reliably confirms that dog mtDNA has not evolved under the neutral model of evolution.

Remarkably, wolves, african golden wolves, coyotes and hunting dogs also show negative values of Tajima's D, in both non-synonymous and synonymous polymorphism positions. Contrary to dogs, the statistical significance of the values is not maintained in Tajima's D nor Fu and Li's formula, which makes it difficult to determine whether there has been a real selection.

Table 4.3. Tajima's D, and Fu and Li's D tests. throughout the exome and in non-synonymous (Nsyn) and synonymous (Syn) mutations as well as Silent Sites (synonymous mutations in non-coding regions). Ratio equals Nsyn/syn. Fu and Li's D was calculated over the whole exome. Significance: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Species	Tajima's D	Nsyn Tajima's D	Syn Tajima's D	Silent Sites Tajima's D	Ratio	Fu and Li's D
Dogs	-2,36298 **	-2,53737***	-2,30487**	-2,31600**	1,10087	-13,25092**
Wolves	-1,66728	-1,75071	-1,60193	-1,65010	1,09287	-1,05266
African golden wolves	-0,55028	-0,80946	-0,70572	-0,51480	1,14700	-0,64288
Coyotes	-0,66856	-1,09716	-1,14228	-0,59819	0,96050	-0,68775
African hunting dogs	-0,65911	-0,59994	-0,72039	-0,66747	0,83280	-0,68446

McDonald Kreitman test showed also interesting results (Table 4.4). The choice of taxas to be compared was critical, as neither very similar nor very divergent sequences contain much information and thus the test would not be applicable. As we can see, all comparisons showed a neutrality index bigger than 1, which indicates that negative selection is acting over all the mitochondrial sequence. The results are statistically significant except for dholes, african hunting dogs and foxes. Those results can confirm, supported by Tajima's D, Fu and Li's D and dn/ds ratios that purifying selection is acting over mitochondrial DNA sequences of at least dogs, grey wolves, african hunting wolves, jackals and coyotes. Even if we do not have significant

results for african hunting dogs, results of dn/ds, McDonald-Kreitman and D statistics, supports negative selection acting on them, although we can not reject neutrality.

Table 4.4. McDonald Kreitman test between two pair of species. Comparisons are listed in the first two columns. Neutrality Index (NI) indicates whether sequence is subjected by neutrality (NI=1), purifying selection (NI > 1) or positive selection (NI<1). Results are supported by Fisher's exact test and G test. Significance: * p < 0.05; ** p < 0.01; *** p < 0.001.

Species 1	Species 2	Neutrality Index	Fisher's test. P value	G test. P value
Dogs	**African golden wolves**	3,289	0,000003***	0,00000***<
Wolves	**African golden wolves**	2,093	0,049346*	0,03206*
Coyotes	**African golden wolves**	1,864	0,002570**	0,00214**
Golden Jackals	**African golden wolves**	1,993	0,003059**	0,00212**
Dholes	**Coyotes**	1,606	0,164588	0,13371
African hunting dogs	**Foxes**	1,359	0,133047	0,13140

5. Discussion

5.1. MtArchitect redesign

We have redesigned our in-house mitochondrial DNA reconstruction pipeline, called mtArchitect, to ensure a better efficiency in the reconstruction of mtDNA. It was designed and used to reconstruct mtDNA sequences from closely related and divergent species lacking of a reference genome. Initially, we started working with a version that was tested on chimps and humans (32) where its performance was quite good. However, we found on that version some important constraints:

- First, the developed tool did not have a good performance with samples whose coverage was very irregular along the sequence causing false deletions on the reconstructed sequences.
- Second, the previous de novo assembly software that was used during the assembly stage, Hapsembler (85), was not creating long contigs but short ones instead causing problems in the consensus sequence.
- The used variant caller (samtools mpileup) did not call indel variants accurately through the sequence as well as some inclusions of false mismatches.
- The parameters during the different stages (lax mapping, iterative and assembly stage) were too stringent for the most divergent species related to dogs which impeded their reconstruction.
- As we stated in Section 1.2. and 4.1., the mtDNA of the family *Canidae* is much more complex than the family *Hominidae'*s mtDNA as those from the canine family have a control region that is quite difficult to resolve properly due to the presence of a repetitive region of approximately 300 base pairs formed by small short tandem repeats.

Consequently, we decided to redesign the pipeline to address these issues. We dealt with all these problems changing and redesigning all the steps with new strategies, programs and parameters (all changes as well as the whole steps of the pipeline are explained in the section 3.1.). However, as discussed in Section 4.1., the repetitive region could not be updated properly during the iterative step which resulted in a bias during the assembly step. This issue is more of a constraint of the paired-end read's length rather than of the designed pipeline since higher lengths of the reads covering the whole repetitive region would fix the issue allowing us to discover the variants correctly in the repetitive region. Last but not least, the previous version

of mtArchitect was compared with MITObim (86), the first developed method to reconstruct mitochondrial sequences which uses the MIRA assembler (87). Although MITObim has a good accuracy in their reconstructions, it can potentially introduce NUMTs and has a low efficiency on the ends of the linearized mitochondrial sequence, which encouraged our lab to develop mtArchitect. In Lobon, I et al, 2016, 10 reconstructed samples with mtArchitect and MITObim were compared with the sequences obtained from long-range PCR showing a mean identity of 99.96% for mtArchitect versus a 99.41% for MITObim showing the better accuracy of mtArchitect.

MtArchitect has been designed to be executed in the command line of UNIX operating systems.

5.2. Phylogenetic tree lineages

As previously shown by Parker et al., (88), there is no necessary relationship between the clustering of dogs and their origin or breed. In our tree, we found a very weak link between the three parameters. The reason for that can be attributed to evolutionary causes, but also to some intrinsic flaws in using mtDNA as a source for comparison.

The evolutionary component of this scattering phenomenon is possibly derived from the impact of human activity in dog breeding and migration over the last centuries: although there has been a great amount of artificial selection in order to create new breeds, the current mating pattern of dogs is mostly spurious and undocumented, especially in countries where dogs live in a free-ranging state i.e. in a state of partial wilderness or as strays. As regards to migration, the close proximity between men and dogs has caused the latter to indirectly undergo the effects of globalization, in other words, the demographic distribution of dogs has turned less predictable and more dependent on humans. Both these factors pose an added difficulty on top of the already complicated task of tracking dogs genetically, and make it very complicated that two dogs with the same breed or geographical status actually belong to a common lineage.

In addition to what was stated above, the use of mtDNA also contributes to these dispersion patterns: first the D-loop, the main source of variability of the mtDNA, has a repetitive region formed by small tandem repeats whose entire extension could not be updated due to the multiple mapping of paired-end reads whose length is smaller than the extension of the repetitive region, consequently lowering the resolution power of the phylogeny analyses. Secondly, as useful as mtDNA might be for evolutionary or phylogenetical purposes, it does not perform as well when trying to explain sheer phenotypical differences (i.e. breed types) from

recently diverging individuals; the reason for that is that most of the genes encoded in the mitochondria are related to energy production rather than physical morphology. Lastly, the maternal inheritance of mtDNA makes it impossible to trace back the paternal breed of a crossbred individual, which introduces an even bigger bias to the phenomenon of clustering by breed.

It can be concluded that mtDNA does not always fulfill the requirements to correctly classify dogs by their breed. Alternative techniques based in autosomal genotyping should be inspected to try and better fill in this need, from which many fields such as forensics or medicine could greatly benefit themselves.

On different note, we have also seen interesting aspects about the classifications of grey wolves subspecies as well as coyotes. We have observed the clustering of some grey wolves (eatern wolf and Great Lakes boreal wolf along with coyotes and red wolves. This observation supports the theories arguing that those wolves may be either a hybrid species product of introgession of grey wolves with coyotes or a distinct species, and that they share a different evolutionary history from that of the grey wolves. We have formuled some theories about of what seems to be ancient wolves (tibetan, chinese and indian wolves) which are basal to the newly grey such as eurasian grey wolves among others. The theories of which we have spoken previously are found in section 4.4.

In contrast to the results obtained with dog breeds, it seems that using mtDNA for a phylogenetic approach is suitable for another species from the family *Canidae* like grey wolves, african golden wolves, and other species that are basal to dogs. However, we found discrepancies about the divergence between dogs, eurasian golden jackals, the ethiopian wolf and coyotes. Most publications suggest that coyotes are located closer to dogs and wolves than eurasian golden jackals and ethiopian wolves although they have used nuclear DNA in their studies. However, as we see in the phylogenetic tree (Figure 4.4), eurasian golden jackals and ethiopian wolves appeared to be closer to dogs and wolves whereas coyotes seemed to be more divergent and thus more ancient species than eurasian golden jackals and ethiopian wolves.

5.3. Evolutionary statistics

Our dn/ds ratio results strongly support the hypothesis that there has been a relaxation in the selective pressure of dog mitochondrial DNA. The enrichment of non-synonymous mutations was found to be highly notorious and, on average, one in every ten mutations was thought to be biologically active and potentially deleterious in contrast with wolves where one

in every sixteen mutations are non-synonymous. We think that these results could be a byproduct of dog domestication and artificial selection: most evidently, there was a radical change in the way of life of ancestral dogs, which moved from being a hunter, nomad species to living a sedentary life under human assistance. This change might have permitted the accumulation of slightly deleterious mutations which would have rendered individuals unfit in a wild environment but passed unnoticed under domestication. Furthermore, through the process of new breed creation and artificial inbreeding, humans might have obliviously introduced a huge source of non-synonymous variation not only within the mtDNA but also at the whole genome level. The implication that these findings might have in the field of dog health are paramount, as they could result in an increase of metabolic diseases, in an increase in the propensity to develop cancer related pathologies and in an overall loss of quality of life.

Wolves were found to have an expected dn/ds ratio, possible due to their feral way of life. However, they have a slightly higher dn/ds than african golden wolves. A plausible reason for that higher ratio is yet to be determined, but it might have something to do with the shrinking status of some current wild wolf populations and with the increasing interbreeding events between dogs and wolves due to that very same fact. That might also be the case of coyotes, which show some high and variable values of the dn/ds ratio. In coyotes however, the effect of introgression should not be that noticeable because, although fertile, the offspring of coyotes and dogs is rarely viable due to ethological reasons e.g. parental rejection of the pups, lack of father-mother pair bonds, unwillingness to mate, etc., and we had too few samples of different subspecies of coyotes to analyse. It would be advisable to analyse more samples of all these species to establish some theories.

The results of Tajima's D test complement those of the dn/ds ratio and also denote some strong signatures of selection in the mitochondrial DNA of dogs. As mentioned in the methods, negative values of Tajima's D normally signal that there has been a recent selective sweep. That most surely points out to the domestication event of wolves, where a small subset of individuals was picked from a larger population, consequently reducing variation over the domesticated animals and thus creating the aforementioned selective sweep. It is also known that the dog population has expanded when compared to the previous centuries (official reports are only produced every 5 to 10 years), which would also fit the pattern of a selective sweep if it was not for the parallel human action involved.

McDonald-Kreitman test confirmed that dogs, wolves, african hunting wolves, jackals, coyotes are subjected by purifying selection with statistical significance, a result supported by

dn/ds ratios and D statistics. Dholes and african hunting wolves also showed this tendency although they did not show statistical results in the test, which unable us to confirm any conclusion about their selection.

5.4. Time to Most Recent Common Ancestor (TMRCA) and inferring of the demographic history of dogs.

As regards to parameter estimation, our results show seem have failed in the calculation of the TMRCA between dogs and wolves, and dogs and ethiopian wolf. Too much stringent parameters for these two priors could be explained. It would be expected that by running BEAST 2 with looser parameters, the two priors could obtain stimated MRCAs. In future experiments we expect results to be similar with those obtained in Thalmann et al, 2013 and other publications whose observations contributes more evidence to the growing belief that dog domestication ranged from 15000 to 40000 years ago (48). In addition,the possibility for more than two domestication events is open to debate: it is known that all the dog fossils that date back to pre-Columbine America have an Eurasian origin (89), but it is still uncertain whether there might have been another main domestication event in the middle East or in Africa (90). Some other authors advocate for a model with multiple minor domestication events along time, but that adds an extra level of complexity to the already complicated task of formulating priors for the model.

The topic of ancient dog demography is still being unraveled, although presently there is a rough estimate of the dog effective population size before and during domestication. We have performed a Bayesian Coalescent Skyline analysis in which we could see a selective sweep of the population after dog domestication at around 30 kya and a posterior population expansion. The decrement in the population size after domestication agree with Thalmann O et al., (10). However, to achieve a better sensitive results, it would be appropriate to repeat the analysis changing parameters as well as incrementing MCMC chains to allow a better mixing of the parameters and be able to see any changes through the time.

6. Conclusions

By looking at the results provided by mtArchitect, it is clear that mtArchitect is a trustable tool to reconstruct mtDNA sequences from closely and divergent species using only a reference genome like CanFam 3.1. However, it has been only designed for Illumina technology of short paired-end and single end reads which nowadays is widely used worldwide.

Our phylogenetic and evolutionary analyses have provided a broader understanding of the history and origin of dogs, as well as for other species like grey wolves, red wolves, jackals and coyotes. Our study of the whole mitochondrial genome has allowed us to reassure most of the existing hypotheses about dog genealogy and demography, as well as to produce some new and more accurate data about some of the key points in the evolution of the *Canidae* family.

Our current findings reveal that, as previously suggested, the variation among dogs is scarce and that such small variation tends to be non-synonymous and potentially functional. However, the question of how the huge amount of phenotypical diversity of dogs has arisen from such little variation in such a small time remains unanswered. We theorize that the variation distribution of the mitochondrial DNA could hold throughout the whole genome and that even if there is a small overall variation, if a great part of it is non-synonymous, there is a huge potential for phenotypical differentiation under assisted selection.

On a different note, our large dataset clustering analyses yielded no definite results on the topic of phenotypical labeling of dogs. We conclude that genomic information is needed for breed identification or provenance tracking and that a better pedigree documentation is paramount for the analyses to come. That opens the gates for more challenging identification methods such as copy number variance or identity by descent.

Whatever the outcome of future investigations is, it is our hope that the present results might have helped to shed a little light into the field of dog genomics. Further research in the topic might lead to a better understanding of the history and domestication process of the *Canidae* family with a direct repercussion on the history, evolution and health of our species.

7. References

1. Tedford RH, Wang X, Taylor BE. Phylogenetic Systematics of the North American Fossil Caninae (Carnivora: Canidae). http://dx.doi.org/101206/5741. American Museum of Natural History, Library-Scientific Publications Central Park West at 79th St., New York, NY 10024 ; 2009 Sep 9;

2. Tedford RH, Wang X, Taylor BE. Phylogenetic Systematics of the North American Fossil Caninae (Carnivora: Canidae). Bull Am Museum Nat Hist. American Museum of Natural History, Library-Scientific Publications Central Park West at 79th St., New York, NY 10024 ; 2009 Sep 3;325:1–218.

3. Molecular evolution of the dog family. Trends Genet. Elsevier Current Trends; 1993 Jun 1;9(6):218–24.

4. Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J.C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S.E., Sánchez-Gracia, A. (2017). DnaSP 6: DNA Sequence Polymorphism Analysis of Large Datasets. Mol. Biol. Evol. XX: (in press). DOI: 10.1093/molbev/msx248.

5. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

6. Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, Silva PM, et al. Genome Sequencing Highlights the Dynamic Early History of Dogs. Andersson L, editor. PLoS Genet. Public Library of Science; 2014 Jan 16;10(1):e1004016.

7. Parker HG, Shearin AL, Ostrander EA. Man's Best Friend Becomes Biology's Best in Show: Genome Analyses in the Domestic Dog. Annu Rev Genet. 2010 Dec;44(1):309–36.

8. Archaeological dogs from the Early Holocene Zhokhov site in the Eastern Siberian Arctic. J Archaeol Sci Reports. Elsevier; 2017 Jun 1;13:491–515.

9. Fossil dogs and wolves from Palaeolithic sites in Belgium, the Ukraine and Russia: osteometry, ancient DNA and stable isotopes. J Archaeol Sci. Academic Press; 2009 Feb 1;36(2):473–90.

10. Thalmann O, Shapiro B, Cui P, Schuenemann VJ, Sawyer SK, Greenfield DL, et al. Complete Mitochondrial Genomes of Ancient Canids Suggest a European Origin of Domestic Dogs. Science (80-). 2013;342(6160).

11. Parker HG, Dreger DL, Rimbault M, Davis BW, Mullen AB, Carpintero-Ramirez G, et al. Genomic Analyses Reveal the Influence of Geographic Origin, Migration, and Hybridization on Modern Dog Breed Development. Cell Rep. 2017;19(4):697–708.

12. Sutter NB, Ostrander EA. Dog star rising: the canine genetic system. Nat Rev Genet. 2004 Dec;5(12):900–10.

13. Chambers SM, Fain SR, Fazio B, Amaral M. An Account of the Taxonomy of North American Wolves From Morphological and Genetic Analyses. North Am Fauna. US Fish & Wildlife Service ; 2012 Oct 1;77:1–67.

14. Mech LD, Christensen BW, Asa CS, Callahan M, Young JK. Production of Hybrids between Western Gray Wolves and Western Coyotes. Michalak P, editor. PLoS One.

Public Library of Science; 2014 Feb 25;9(2):e88861.

15. Zhang H, Chen L. The complete mitochondrial genome of dhole Cuon alpinus: phylogenetic analysis and dating evolutionary divergence within canidae. Mol Biol Rep. Springer Netherlands; 2011 Mar 22;38(3):1651–60.

16. Werhahn G, Senn H, Kaden J, Joshi J, Bhattarai S, Kusi N, et al. Phylogenetic evidence for the ancient Himalayan wolf: towards a clarification of its taxonomic status based on genetic sampling from western Nepal. R Soc open Sci. 2017 Jun;4(6):170186.

17. Fan Z, Silva P, Gronau I, Wang S, Armero AS, Schweizer RM, et al. Worldwide patterns of genomic variation and admixture in gray wolves. Genome Res. Cold Spring Harbor Laboratory Press; 2016 Feb 1;26(2):163–73.

18. Aggarwal RK, Kivisild T, Ramadevi J, Singh L. Mitochondrial DNA coding region sequences support the phylogenetic distinction of two Indian wolf species. J Zool Syst Evol Res. Blackwell Publishing Ltd; 2007 May 1;45(2):163–72.

19. Sharma DK, Maldonado JE, Jhala Y V, Fleischer RC. Ancient wolf lineages in India. Proceedings Biol Sci. The Royal Society; 2004 Feb 7;271 Suppl 3(Suppl 3):S1-4.

20. Koepfli K-P, Pollinger J, Godinho R, Robinson J, Lea A, Hendricks S, et al. Genome-wide Evidence Reveals that African and Eurasian Golden Jackals Are Distinct Species. Curr Biol. 2015 Aug;25(16):2158–65.

21. Sato M, Sato K. Maternal inheritance of mitochondrial DNA by diverse mechanisms to eliminate paternal mitochondrial DNA. Biochim Biophys Acta - Mol Cell Res. 2013 Aug;1833(8):1979–84.

22. Travis J. Mom's Eggs Execute Dad's Mitochondria. Sci News. Society for Science & the Public; 2000 Jan 1;157(1):5.

23. Chinnery PF, Hudson G. Mitochondrial genetics. Br Med Bull. Oxford University Press; 2013;106(1):135–59.

24. Fernández-Silva P, Enriquez JA, Montoya J. Replication and transcription of mammalian mitochondrial DNA. Exp Physiol. 2003 Jan;88(1):41–56.

25. Scarpulla RC. Molecular Biology of the OXPHOS System. Landes Bioscience; 2013;

26. Kemp BM, Judd K, Monroe C, Eerkens JW, Hilldorfer L, Cordray C, et al. Prehistoric mitochondrial DNA of domesticate animals supports a 13th century exodus from the northern US southwest. PLoS One. 2017;12(7):e0178882.

27. Gaubert P, Bloch C, Benyacoub S, Abdelhamid A, Pagani P, Djagoun CAMS, et al. Reviving the African wolf Canis lupus lupaster in North and West Africa: a mitochondrial lineage ranging more than 6,000 km wide. PLoS One. 2012;7(8):e42740.

28. Kim KS, Lee SE, Jeong HW, Ha JH. The Complete Nucleotide Sequence of the Domestic Dog (Canis familiaris) Mitochondrial Genome. Mol Phylogenet Evol. 1998 Oct;10(2):210–20.

29. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, et al. Genome

sequence, comparative analysis and haplotype structure of the domestic dog. Nature. Nature Publishing Group; 2005 Dec 8;438(7069):803–19.

30. Wallace DC. Why Do We Still Have a Maternally Inherited Mitochondrial DNA? Insights from Evolutionary Medicine. Annu Rev Biochem. 2007 Jun 7;76(1):781–821.

31. Galtier N, Nabholz B, Glémin S, Hurst GDD. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. Mol Ecol. 2009 Nov;18(22):4541–50.

32. Lobon I, Tucci S, De Manuel M, Ghirotto S, Benazzo A, Prado-Martinez J, et al. Demographic history of the genus Pan inferred from whole mitochondrial genome reconstructions. Genome Biol Evol. Lincoln Park Zoo, Chicago, IL, USA; 2016 Jun 1;8(6):2020–30.

33. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.

34. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Internet]. 2013 [cited 2017 Sep 9]. Available from: http://arxiv.org/abs/1303.3997

35. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. Oxford University Press; 2009 Aug 15;25(16):2078–9.

36. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156–8.

37. Laurie S, Fernandez-Callejo M, Marco-Sola S, Trotta J-R, Camps J, Chacón A, et al. From Wet-Lab to Variations: Concordance and Speed of Bioinformatics Pipelines for Whole Genome and Whole Exome Sequencing. Hum Mutat. 2016 Dec 1;37(12):1263–71.

38. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. 2012 Jul 17;

39. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. Oxford University Press; 2010 Mar 15;26(6):841–2.

40. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008 Feb 21;18(5):821–9.

41. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. Oxford University Press; 2013 Apr 1;30(4):772–80.

42. Millennial-scale Asian summer monsoon variations in South China since the last deglaciation. Earth Planet Sci Lett. Elsevier; 2016 Oct 1;451:22–30.

43. Wang G, Zhai W, Yang H, Fan R, Cao X, Zhong L, et al. The genomics of selection in dogs and the parallel evolution between dogs and humans. Nat Commun. 2013 May 14;4:1860.

44. Auton A, Rui Li Y, Kidd J, Oliveira K, Nadel J, Holloway JK, et al. Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. Henderson I, editor. PLoS Genet.

Public Library of Science; 2013 Dec 12;9(12):e1003984.

45. Fan Z, Silva P, Gronau I, Wang S, Armero AS, Schweizer RM, et al. Worldwide patterns of genomic variation and admixture in gray wolves. Genome Res. 2016 Feb;26(2):163–73.

46. vonHoldt BM, Kays R, Pollinger JP, Wayne RK. Admixture mapping identifies introgressed genomic regions in North American canids. Mol Ecol. 2016 Jun;25(11):2443–53.

47. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. Oxford University Press; 1987 Jul 1;4(4):406–25.

48. Gascuel O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol. 1997 Jul;14(7):685–95.

49. Cho A. Constructing Phylogenetic Trees Using Maximum Likelihood. Scripps Senior Theses. 2012.

50. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol. 2010 Mar 29;59(3):307–21.

51. Lefort V, Longueville J-E, Gascuel O. SMS: Smart Model Selection in PhyML. Mol Biol Evol. 2017 Sep 1;34(9):2422–4.

52. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 2012 Jul 30;9(8):772–772.

53. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, et al. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. Prlic A, editor. PLoS Comput Biol. Public Library of Science; 2014 Apr 10;10(4):e1003537.

54. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012 Aug;29(8):1969–73.

55. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. Mol Biol Evol. 2005 Feb 9;22(5):1185–92.

56. Rambaut A, Suchard MA, Xie D & Drummond AJ (2014) Tracer v1.6, Available from http://tree.bio.ed.ac.uk/software/tracer/.

57. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. Nature Publishing Group; 1991 Jun 20;351(6328):652–4.

58. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. Genetics Society of America; 1989 Nov;123(3):585–95.

59. Mcvean G. Natural selection. 2002;

60. Langmead B, Trapnell C, Pop M, Salzberg SL, Ruan J, Homer N, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. BioMed Central; 2009 Jun 20;10(3):R25.

61. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016 Jul;33(7):1870–4.

62. He Z, Zhang H, Gao S, Lercher MJ, Chen W-H, Hu S. Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. Nucleic Acids Res. Oxford University Press; 2016 Jul 8;44(W1):W236-41.

63. vonHoldt BM, Pollinger JP, Earl DA, Knowles JC, Boyko AR, Parker H, et al. A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. Genome Res. Cold Spring Harbor Laboratory Press; 2011 Aug;21(8):1294–305.

64. vonHoldt BM, Cahill JA, Fan Z, Gronau I, Robinson J, Pollinger JP, et al. Whole-genome sequence analysis shows that two endemic species of North American wolf are admixtures of the coyote and gray wolf. Sci Adv. 2016;2(7).

65. Rutledge LY, Wilson PJ, Klütsch CFC, Patterson BR, White BN. Conservation genomics in perspective: A holistic approach to understanding Canis evolution in North America. Biol Conserv. 2012 Oct;155:186–92.

66. KOBLMÜLLER S, NORD M, WAYNE RK, LEONARD JA. Origin and status of the Great Lakes wolf. Mol Ecol. 2009 Jun;18(11):2313–26.

67. Wilson PJ, Grewal S, Lawford ID, Heal JN, Granacki AG, Pennock D, et al. DNA profiles of the eastern Canadian wolf and the red wolf provide evidence for a common evolutionary history independent of the gray wolf. Can J Zool. 2000 Dec;78(12):2156–66.

68. Duleba A, Skonieczna K, Bogdanowicz W, Malyarchuk B, Grzybowski T. Complete mitochondrial genome database and standardized classification system for Canis lupus familiaris. Forensic Sci Int Genet. 2015 Nov;19:123–9.

69. Pang J-F, Kluetsch C, Zou X-J, Zhang A, Luo L-Y, Angleby H, et al. mtDNA data indicate a single origin for dogs south of Yangtze River, less than 16,300 years ago, from numerous wolves. Mol Biol Evol. Oxford University Press; 2009 Dec;26(12):2849–64.

70. Wayne RK, Lehman N, Allard MW, Honeycutt RL. Mitochondrial DNA Variability of the Gray Wolf: Genetic Consequences of Population Decline and Habitat Fragmentation. Conserv Biol. Blackwell Science Inc; 1992 Dec 1;6(4):559–69.

71. Randi E, Lucchini V, Christensen MF, Mucci N, Funk SM, Dolf G, et al. Mitochondrial DNA Variability in Italian and East European Wolves: Detecting the Consequences of Small Population Size and Hybridization. Conserv Biol. Blackwell Science Inc; 2000 Apr 1;14(2):464–73.

72. Pilot M, Branicki W, Jedrzejewski W, Goszczyński J, Jedrzejewska B, Dykyy I, et al. Phylogeographic history of grey wolves in Europe. BMC Evol Biol. BioMed Central; 2010 Apr 21;10:104.

73. Klütsch CFC, Savolainen P, Lohi H, Fall T, Hedhammar Å, Uhlén M, et al. Regional occurrence, high frequency, but low diversity of mitochondrial dna haplogroup d1 suggests a recent dog-wolf hybridization in scandinavia. J Vet Behav Clin Appl Res. Elsevier; 2011 Jan 1;6(1):85.

74. Ersmark E, Klütsch CFC, Chan YL, Sinding M-HS, Fain SR, Illarionova NA, et al. From the Past to the Present: Wolf Phylogeography and Demographic History Based on the Mitochondrial Control Region. Front Ecol Evol. Frontiers; 2016 Dec 2;4:134.

75. Li Y, Li Q, Zhao X, Xie Z, Xu Y. Complete sequence of the Tibetan Mastiff mitochondrial genome and its phylogenetic relationship with other Canids (Canis, Canidae). animal. Elsevier; 2011 Jan 16;5(1):18–25.

76. Hefner R, Geffen E. Group Size and Home Range of the Arabian Wolf (Canis lupus) in Southern Israel. J Mammal. Oxford University Press; 1999 May 20;80(2):611–9.

77. Genetic variation and subspecific status of the grey wolf (Canis lupus) in Saudi Arabia. Mamm Biol - Zeitschrift für Säugetierkd. Urban & Fischer; 2014 Nov 1;79(6):409–13.

78. Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, et al. Genetic Variation and Population Structure in Native Americans. PLoS Genet. Addison Wesley; 2007;3(11):e185.

79. Larson G, Karlsson EK, Perri A, Webster MT, Ho SYW, Peters J, et al. Rethinking dog domestication by integrating genetics, archeology, and biogeography. Proc Natl Acad Sci U S A. National Academy of Sciences; 2012 Jun 5;109(23):8878–83.

80. Savolainen P, Rosén B, Holmberg A, Leitner T, Uhlén M, Lundeberg J. Sequence analysis of domestic dog mitochondrial DNA for forensic use. J Forensic Sci. 1997 Jul;42(4):593–600.

81. Ostrander EA, Wayne RK. The canine genome. Genome Res. 2005 Dec 1;15(12):1706–16.

82. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. NIH Public Access; 2011 Jan;29(1):24–6.

83. HO SYW, SHAPIRO B. Skyline-plot methods for estimating demographic history from nucleotide sequences. Mol Ecol Resour. Blackwell Publishing Ltd; 2011 May 1;11(3):423–34.

84. Björnerfeldt S, Webster MT, Vilà C. Relaxation of selective constraint on dog mitochondrial DNA following domestication. Genome Res. Cold Spring Harbor Laboratory Press; 2006 Aug;16(8):990–4.

85. Donmez N, Brudno M. Hapsembler: An Assembler for Highly Polymorphic Genomes. In 2011. p. 38–52.

86. Hahn C, Bachmann L, Chevreux B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads--a baiting and iterative mapping approach. Nucleic Acids Res. Oxford University Press; 2013 Jul;41(13):e129.

87. Chevreux, Wetter T, Suhai S. Genome Sequence Assembly Using Trace Signals and AdditionalSequence Information. 1999;99.

88. Parker HG. Genomic analyses of modern dog breeds. Mamm Genome. NIH Public Access; 2012 Feb;23(1–2):19–27.

89. Sablin MV, Khlopachev GA. The Earliest Ice Age Dogs: Evidence from Eliseevichi 1. Curr Anthropol. The University of Chicago Press ; 2002 Dec 17;43(5):795–9.

90. Ardalan A, Oskarsson MCR, van Asch B, Rabakonandriania E, Savolainen P. African origin for Madagascan dogs revealed by mtDNA analysis. R Soc Open Sci. 2015;2(5).