

Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN
Biodiversità ed Evoluzione

Ciclo XXIV

Settore Concorsuale di afferenza:
05/B1 – Zoologia e Antropologia

TITOLO TESI

Ecological genetics and conservation genomics of wolf (*Canis lupus*)

Presentata da: Marco Galaverni

Coordinatore Dottorato

Barbara Mantovani

Relatore

Ettore Randi

Esame finale anno 2012

Ecological genetics and conservation genomics of wolf (*Canis lupus*)

*From the Major Histocompatibility Complex to the first wolf genome draft
(and some secrets of dog domestication)*

“The greatest danger for most of us
is not that our aim is too high and we miss it,
but that it is too low and we reach it”

Michelangelo

Summary

1. Introduction
 - 1.1. Overview on a fast-changing scientific era
 - 1.2. From Population Genetics to Ecological Genetics and Conservation Genomics
 - 1.2.1. New perspectives
 - 1.2.2. New technologies: Next Generation Sequencing
 - 1.3. The species: *Canis lupus*, Linnaeus 1758
 - 1.3.1. Origin and distribution
 - 1.3.2. Morphology and biology
 - 1.3.3. Threats and legal status

2. The Major Histocompatibility Complex: its variability in the Italian wolf population and its influence on mating choice and fitness traits
 - 2.1. Background
 - 2.1.1. Structure and functions
 - 2.1.2. Genetic features and evolution
 - 2.1.3. Methods
 - 2.1.4. Studies
 - 2.1.5. MHC in canids
 - 2.2. Aims
 - 2.3. Methods
 - 2.4. Results
 - 2.5. Discussion and implications

3. The first wolf genome project
 - 3.1. Background
 - 3.1.1. Overview on whole-genome studies
 - 3.1.2. The dog genome
 - 3.1.3. Dog domestication and breeding
 - 3.2. Aims
 - 3.3. Methods
 - 3.3.1. Sequencing methods
 - 3.3.2. Analyses workflow
 - 3.4. Preliminary results
 - 3.5. Discussion and implications

4. Conclusions

5. Bibliography

6. Acknowledgements

1. Introduction

1.1. Overview on a fast-changing scientific era

A very few fields in research have seen such a fast evolving phase in the recent years as genetics.

As a beginning researcher, I am astonished in seeing the rhythm of new techniques, approaches and questions arising during just the few months of my doctoral project.

But on the other side, this cultural blast offers great opportunities for those who are interested (and so am I) in following the wave, or riding it, and looking for the thousand possibilities of applying the results of research to the widest range of applications.

In this dissertation, we will try to have a closer look at the fields of Ecological Genetics and of the brand-new Conservation Genomics, using as a study species one of the most fascinating carnivores ever appeared on Earth: the wolf.

1.2. From Population Genetics to Ecological Genetics and Conservation Genomics

The past decade has seen a large usage of neutral-behaving genetic markers such as microsatellites (or single tandem repeats, STRs) and mitochondrial DNA (mtDNA) control region, in order to assess the basic genetic variables in animal and plant populations, with particular attention to the ones presenting conservation concerns (Ouborg et al. 2010).

Those estimates have allowed to identify cases of reduced effective population size, restricted gene flow, limited heterozygosity, but also inbreeding, past bottlenecks and hybridization or gene introgression, all factors that could seriously affect the population viability and long-term survival, especially in times of fast climate changes and strong human-driven environment modifications.

The same genetic markers allowed the researchers to reconstruct the phylogenetic relationship, social structure, kin affiliations and individual fitness estimates in many social species, particularly among mammals and birds.

Nonetheless, new questions are arising on the shoulders of the previous ones, and their answers are now much closer also thanks to great technological improvements occurred in the last years, and to the analytical tools related to them.

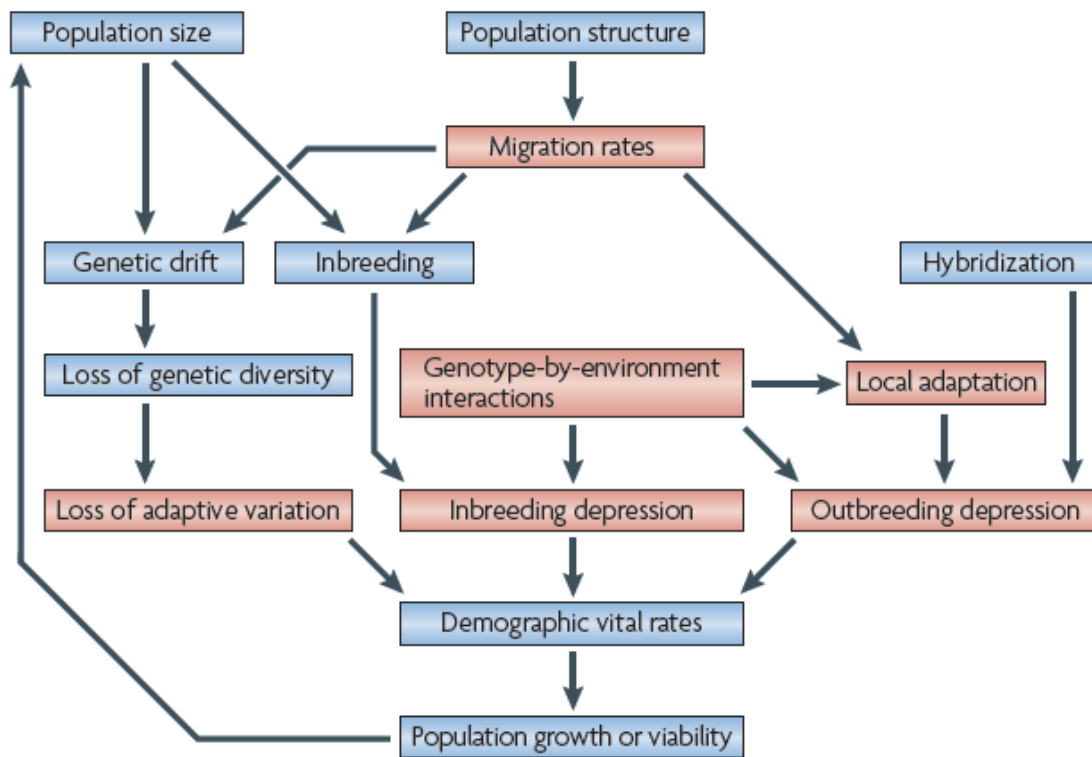


Figure 1: network of interacting factors that can be addressed by conservation genetics (blue) and genomics (red). (From Allendorf et al. 2010)

1.2.1. New perspectives

Ecological Genetics, or Molecular Ecology, is not a new research field.

The study of the relationships of individuals with the environment (represented both by the habitat, the social structure, the food networks -especially the prey-predator relations and co-evolution- the climate and the pathogens), based on their genetic background, and the returning effects of the environment in driving and shaping the genetic features of the individuals through natural and sexual selection, has seen a never-dropping interest.

However, the limited resources usually available to researchers did not allow for the investigation of a large number of genetic markers, therefore often limited to a few genes or non-coding regions of interest.

Nowadays, on the contrary, revolutionary technologies allow for the screening of thousands of genome-wide genetic markers, e.g. single nucleotide polymorphisms (SNPs), or, moreover, whole genome sequences in a very short time and a relatively limited economic effort.

This huge upgrade can make it easier to deepen existing disciplines (as for Ecological Genetics -or Molecular Ecology- and Genome-Wide Association Studies -GWAS), thanks to the enlargement of the number and size of the markers, or even to open the way to the development of new branches, such as Conservation Genomics (Ouborg et al. 2010).

This emerging discipline can be simply defined as the application “of new genetic techniques to solve problems in conservation biology” (Allendorf et al. 2010), such as genetic drift, hybridization, inbreeding or outbreeding depression, natural selection, loss of adaptive variation and fitness (Fig. 1).

The whole genomes of some endangered species have been recently completed, starting from the Great Apes: chimpanzee, gorilla and orang-utan (Locke et al. 2011); however, these data will not automatically provide useful data for their conservation (Frankham 2010), especially given the limited information about population variation that we are able to deduce from single individual sequencing. Nonetheless, this will provide a great aid in identifying genetic markers that can be applied to the study of entire populations (Frankham 2010).

Genomic information will turn out to be useful also to try and recover populations from strong inbreeding depressions, by identifying the genes exposing deleterious alleles (Allendorf et al. 2010) and augmenting the population variability with crosses of the most appropriate individuals (Frankham 2010). On the other side, the same techniques will allow to identify the *loci* most responsible for speciation or cryptic local adaptation, or for exposing populations to severe diseases (Allendorf et al. 2010), such as the facial tumor affecting the Tasmanian devil or the fungus threatening several bat populations.

Having a minor focus on conservation issues, other disciplines (whose limits are often difficult to distinguish) raised, such as evolutionary and ecological functional genomics (EEFG; Feder and Mitchell-Olds 2003).

Meanwhile, international *consortia* such as the 1000 Genomes Project are already aiming at sequencing a population-wide sample of genomes “to provide a deep characterization of [human] genome sequence variation as a foundation for investigating the relationship between genotype and phenotype” (The 1000 Genomes Project Consortium 2010). Moreover, an even more ambitious project, the Genome 10Kproject (<http://www.genome10k.org/>), aims at collecting a whole genomic zoo, including thousands vertebrate species from different genera (Genome 10K Community of Scientists 2009).

However, beside this unique kind of projects, the full sequencing of a *de novo* transcriptome or genome, even in non-model eukaryotic species, is becoming feasible for most of the research institutions.

In the next paragraph, we will see some details that are common to most of the state-of-the-art sequencing platforms, usually indicated as Next Generation Sequencing, that make all the mentioned projects and applications possible.

1.2.2. New technologies: Next Generation Sequencing

Big revolutions often start from very simple ideas that quickly constitute new paradigms.

In our case, the simple idea behind a new generation of sequencing techniques is the passage from a single sequencing reaction and reading (usually capillary-based, ‘Sanger’ method) to a multiple, parallel process involving thousands of fragments.

And as in many other fields, often a *plethora* of similar developments takes place in a very short time and along independent paths, giving place to a number of platforms addressing the same targets in slightly different ways.

However, although in order to highlight the terrific improvements in the sequencing capacity they have been defined as Next Generation Sequencing (NGS) techniques, their immediate diffusion and application should better let us define them as This Generation Sequencing. And their costs, that on a per-base scale are dropping constantly, will soon make them available to an increasingly larger community of investigators in many fields other than human medical research.

The methods

As we just saw, a number of different technologies have been developed in order to achieve similar goals. Nonetheless, the sample preparation and the sequencing methods show a common framework, namely the ability of processing millions of short fragments (also defined as ‘reads’) at the same time, on the same instrument, in the same run (the so-called ‘in parallel’ sequencing).

The starting point (Mardis 2008) is to build a set of fragments (usually named ‘library’) that does not require the cloning by any bacterial vector. These libraries are produced by a mechanical or enzymatic fragmentation of the whole DNA of the target organism or cells.

The fragments with the desired length are selected, ligated to oligonucleotide probes (‘adaptors’) and amplified. The nucleotide sequence of every fragment clone is then fixed on a support, read in parallel through a chemical -usually base-by-base- process, and digitalized. However, every company has developed a unique system based on different techniques.

In the next paragraph, we will see the protocols and instruments adopted by the three major companies that nowadays compete for the largest part of the Next Gen Sequencing market.

The major platforms: Roche-454 GS, Illumina HiSeq, ABI SOLiD.

When comparing sequencing systems, there are a number of factors to be considered (Cokus 2011). First of all, two quantitative measures, namely the total quantity of sequence ('throughput') and the average size of each fragment ('read length'). Secondly, but with a comparable importance, the quality of the output, represented by the type and frequency ('error rate') of the errors, the reproducibility of the output, and the random sampling of the molecules. Beside these, other important factors are the running time (it matters when it comes from hours to weeks!) and the number of runs that can be performed at the same time, but also the simplicity of the sample preparation (in terms of flexibility, quantity of DNA required, and number of libraries that can be pooled in the same run) and the analysis pipeline (comprehensive of the hardware needed, software power and availability, people and expertise needed for managing it). Eventually, but probably the most important, the costs, which should be carefully considered, including the expenses for buying the instrument, its maintenance and depreciation, the reagents and the ordering quantities.

When considering all of this, even for large institutions, outsourcing to a service company can sometimes turn out to be the most affordable solution: in this case, only the costs per sequencing run and per Mb produced have to be weighted to the scale of the project we are interested in performing.

One of the very first platforms to be developed and successfully applied to a number of studies has been produced by Roche-454 and commercialized in 2004. Named *GS Sequencer* (with different editions that appeared through years, such as FLX), is based on a pyrosequencing reaction. The amplification step occurs in an emulsion PCR, where the fragments are ligated to agarose beads thanks to universal adaptors and amplified in an oil-water mixture, in which every water-phase drop contains a single DNA fragment and all the amplification reagents. Every bead-linked clone is then located in a unique, picometer-sized well on the surface of a titer plate (PTM). Nucleotides and pyrosequencing reagents are then added in cycles, and the light emission caused by the luciferase during the incorporation of a given nucleotide allows the imaging of the sequence as a flowgram, where the intensity of the light is proportional to the number of nucleotides with the same base that have been incorporated consecutively. Therefore, the length of homopolymers is the limiting factor in the accuracy of the machine. Conversely, the long read length allowed by the pyrosequencing approach, compared to the other platforms, is the main advantage of the system (Tab. 1).

Company and Platform	ROCHE-454 GS-FLX+	Illumina HiSeq 2000	ABI SOLiD - 4
Sequencing chemistry	Pyrosequencing	Polymerase-based sequencing-by-synthesis	Ligation-based sequencing
Amplification type	Emulsion PCR	Bridge amplification	Emulsion PCR
Read length	700 bp	100+100	50 + 35
Paired-end / mate-pair	yes	Yes	yes
Mb / run	900 Mb	200 Gb	70 Gb
Time / run	18-20 hr	8 days	12 days
Cost / run	\$6200	\$20120	\$8128
Cost / Mb	\$7	\$0.10	<\$0.11
Main <i>pros</i>	Long read length	High throughput, low cost/Mb	Low error rate
Main <i>cons</i>	Limited throughput, high cost per Mb	Errors accumulating at the 3' end of reads	Color-based coding system

Table 1: Summary of the sequencing approaches and specifications offered to date by the three most common platforms (modified from Mardis 2008 and from Glenn 2011).

The approach followed by Illumina (former Solexa) since 2006, both in the first-born *Genome Analyzer* and in the current *HiSeq* platforms, is based on the so-called sequencing-by-synthesis (SBS) process. The single-strand DNA fragments are ligated at both ends to the internal surface of a glass cell (divided into eight lanes) thanks to oligonucleotide adaptors that bind complementary probes attached to the cell, forming a bridge-like structure. The fragments are then provided the amplification reagents and incubated, forming clonal clusters randomly located on the cell surface. In the sequencer, in the reads of every cluster the polymerase incorporates a single fluorescent nucleotide of the four provided at every cycle, which also carries a 3'-OH group in order to immediately terminate the extension and to be read by an imaging device. After every incorporation, the OH and the fluorochrome groups need to be removed before the starting of a new cycle. The time required for every step is therefore the limiting factor determining the read length, whereas the problem of homopolymer reading is strongly reduced compared to 454 (Tab. 1).

Also Applied Biosystems in 2007 developed its own platform, named *SOLiD* in order to designate the “Sequencing by Oligo Ligation and Detection” process used and the proclaimed accuracy of the machine. The amplification step follows an emulsion-PCR reaction where the DNA fragments are added an adaptor and ligated to magnetic beads by complementary oligos. The bead complexes are then fixed to a glass slide for the sequencing step, in which a set of semi-degenerated 8mer probes hybridize to the DNA fragments starting

from a primer annealed to the known adaptor sequence, every probe interrogating two positions, whose nucleotides combination is signalled by a specific fluorochrome on the probe. The fluorescent group is then removed, allowing the ligation of the following probe that will interrogate other two nucleotides 5 bp apart from the first ones, and so on until the end of the fragment. The whole cycle is repeated five times starting from different points (n, n-1, n-1, n, n-1) defined by distinct primers, permitting the reading of the whole fragment twice. The next step is deciphering the 2-bp code, starting from the known adaptor, that will be represented by a sequence of colors. This allows the discrimination of sequence differences (SNPs) from sequencing errors, therefore increasing the accuracy at the price of short reads and long running time (Tab. 1), but with a total throughput nonetheless many times higher than by 454.

However, beside these three main competitors, a range of new companies are emerging and developing -sometimes radically- new techniques that in the next few years could revolutionize the field, and potentially also the geography, of genomic research.

The emerging platforms

A revolutionary approach (D. J. Turner et al. 2009) was presented by the Helicos true single-molecule-sequencing (*tSMS*) technology, which does not require any amplification step before the sequencing (Braslavsky et al. 2003), therefore avoiding PCR-induced biases (CG content bias, phasing errors). It allows for the extension of about 800M of short (25-50 bp) fragments ligated by a poly-A tail to complementary poly-T adaptors on the cell surface. Then the sequences are read by adding single fluorescent nucleotides in a pyrosequencing fashion. Similarly to 454 instruments, it shows accuracy problems with long homopolymer sequences. The error rates can be reduced by reading the same fragments twice, but also increasing the costs, which can turn out to be already a limiting factor.

Another single-molecule reading platform is represented by the Pacific Bioscience *SMRT* (Eid et al. 2009), which is based upon the use of nucleotides labeled with reversible fluorophores that can be sequentially inserted by a single polymerase inside a nano-sized pore ('smart cell'); afterward their fluorescence can be read by contrast to the background noise (the so-called 'Zero mode waveguides'). It can produce an output of only *ca.* 40 Mb per run, but it is fast, cheap and the reads are longer than in many other platforms, although the same fragment should be read multiple times in order to get a satisfactorily low error rate.

Oxford Nanopore *BASE* platform, currently held by Illumina, is also based on a single-molecule approach (Maglia et al. 2008). The fragment sequence is deduced by the

conductivity changes perceived when a nucleotide, after being digested by an exonuclease, passes through a nanopore and binds to a cyclodextrin. The nanopores are placed into a double lipid layer onto a microwell hosted on a silicon chip.

The *IonTorrent* (Life) system has been incorporated by the ABI company, and also promises to be among the most competitive platforms on the market also thanks to the very limited size. The system is based on a sequencing-by-synthesis reaction on a silicon chip, where the incorporated nucleotides are read by nanoscopic pHmeters, rather than by camera; it can produce about 1Gb of output sequence in a simple and fast way.

Alternative approaches can be represented, for instance, by a ‘strobe’ sequencing, where 50 bp fragments are sequenced every 10 kb along the genome, allowing an optimal reconstruction of structural variants. In the future, the so-called “physical methods” will not require the use of any biological enzyme, but will be based on the physical properties of the DNA molecule itself, such as the different electrical signal produced by the nucleotides (Reveo), possibly read on a ‘DNA transistor’ (IBM). But this is the future, and their commercial production is still to come (see Glenn 2011 for an exhaustive review).

Whatever the platform, the exponential growth of the sequencing power will probably allow us to sequence millions of genomes in the next decade. But -beside strong information storage issues- the next problem will be: what to do with them?

The applications

Multiple studies have exploited the possibilities offered by the NGS platforms, and they can be grouped into three main categories based on their target information: DNA sequencing, RNA-based studies and, recently, methylome sequencing.

The first group is mainly represented by whole genome sequencing studies, in which the candidate genome is sequenced up to a sufficient coverage to allow its complete reading. Whereas this can be relatively simple for the less complex genomes, such as prokaryotes, the studies on animal and plant species are still relatively few. This approach is usually followed when the reference genome of a close relative species is not available, therefore is not possible to apply the same markers (e.g. known SNPs) to study of the target species. For the same reason, it usually requires a *de novo* assembly process, that can be time-consuming and difficult in the case of complex genomes, such as for polyploid plants. However, even the complete sequencing of a single individual usually allows the detection of millions of Single Nucleotide Variants (SNVs), that can be then used at a population level as markers, for example, in Genome-Wide Association Studies (GWAS) or in population genetics, and other

markers such as microsatellites (STRs) can be identified. If the coverage produced by the chosen platform is sufficiently high and the library preparation requirements allow it, more than one individual can be pooled on the same run. In this way, by comparing multiple genome, a higher number of SNVs and a list of genomic structural features, such as Structural Variants (SVs) and Copy Number Polymorphisms or Variants (CNVs) can also be detected, in addition to insertions and deletions (InDels) events, which are usually linked to repetitive and mobile elements (e.g. Short or Long Interspersed Elements, SINEs and LINEs).

Other members of this DNA-sequencing category are represented by several approaches of targeted sequencing (or re-sequencing): in these cases, only a portion of the candidate genome is selected and sequenced, in order to focus the efforts on a given set of regions of interest. The tools that allow the selection of genomic subsets are several, and they mainly include: Reduced Representation Libraries (RRLs, Altshuler et al. 2000); commercial or custom Targeted Capture arrays (Hodges et al. 2009); Complexity Reduction tools (CRoPS, van Orsouw et al. 2007). These approaches have been successfully applied to the study of large genomes (Burbano et al. 2010, Ng et al. 2009, Wiedmann et al. 2008), although capture arrays require the prior knowledge of the target sequence -or at least the one from a similar organism.

Intermediate between DNA and RNA sequencing approaches we find the transcriptome sequencing (also called mRNA-seq). Beside the fact that works with messenger RNA (mRNA) as starting material (then reverse-transcribed into cDNA), it is basically another way of focusing the sequencing efforts on a subset of the genome, namely its codifying portion: the exome. Its main advantages are that it allows to reach a much higher coverage at a much lower cost, since the size of the protein-coding elements is usually a small fraction of the total genome (*ca.* 1% in humans, Ng et al. 2009), it does not include complex structural features (such as repetitive elements) and the markers identified from it can be directly related to their possible biological function. If the read length and the coverage are sufficiently high (hundreds of bases) a self-assembly could be achieved without unsolvable nodes, therefore making these studies feasible even in absence of a reference genome.

The second group, including gene expression studies, directs its aim at evaluating which genes are expressed and the differences in the expression levels of the transcripts, usually by comparing multiple individuals of the same species or by pooling individuals from different groups to be contrasted. Also in this case, the mRNA is first selected out of the total RNA, then retro-transcribed into cDNA. The reads are usually aligned to reference genomes, or directly to other reference mRNAs. However, if the reads are sufficiently long to allow the

unique identification of the transcript they match to, they can be directly used for the quantification of the gene expression without the need of assembly them. However, the expression levels of the genes can vary by orders of magnitude, even from tissue to tissue, therefore requiring coverage much higher than for standard sequencing. Many of the pioneering studies (a few years ago, though) were based on the Expressed Sequence Tags (ESTs), an approach that allows to sequence only the terminal portions of each transcript, therefore saving precious sequencing power at the cost of losing sequence information at the central part of the transcripts. mRNA-seq can be an important step in the annotation of the genes.

The third group, although less common than the previous ones in the scientific literature, is addressing questions on the methylation status of candidate genomes, allowing to focus on some of the main features influencing the epigenetic regulation (Bossdorf et al. 2008). Briefly, it is based on the detection of the differences between the cytosine carrying a -CH₃ group (which can be turned into thymine if treated with sodium bisulfate during the library preparation) and the non-methylated ones. It is mainly applied to the study of embryonic development and carcinogenetics (Zhang and Jeltsch 2010).

But other important applications of NGS techniques have been recently targeted at the study of short RNAs (whose features and functions, beside gene regulation, are still being investigated), or like the so-called Nuc-seq (the study of which parts of DNA bind to the nucleosomes), or the new Chromosome Conformation Capture (CCC, or 3C, or Hi-C, that aims at reconstructing which portions of the DNA helix are actually adjacent and interact in the three-dimensional space).

The strategies

Almost every different study requires a different approach, and the combinations given by the starting molecule (DNA or RNA), library preparation (single fragments, paired end or mate pairs), sequencing platform, alignment method and analysis pipeline, that includes software and data storage, makes it hard to define a list of common solutions.

However, two steps are basic choices in every project design: the library preparation and the alignment method.

As we saw, the library preparation varies according the platform requirements in terms of fragmentation methods (mechanic or enzymatic) and read size. However, most of the libraries can be prepared aiming at sequencing either single fragments, or pairs of fragments separated by a known distance.

The single fragment libraries are the most common and simple ones. After basic quality controls on the DNA sample (quantity, concentration, integrity, etc.), the DNA is randomly sheared (often by nebulization), and the fragments with a suitable size, as required by the platform, are usually selected by a gel run. They will then be ligated to the adaptors following the manufacturer's protocols and simply run on the platform.

Conversely, an expanding approach is based on the possibility of sequencing pairs of fragments that lay on the same chromosome at a given distance. This is possible both by the construction of paired-end libraries as well as mate-pair libraries. The two differ for the size of the insert (usually shorter in the first case) and procedure to obtain them.

The paired-end (PE) reads are simply a selection of fragments being longer than what the machine will actually read by a known length, corresponding to the "insert" size; then the machine will only sequence their most external portions (usually corresponding to the length of a single fragment) on both sides, but in opposite directions. In the end, the two fragments will be therefore spaced by a known distance. This method is used for fragments less than 1kb apart. For longer distances, the most common approach is to build a mate-pair library. To do that, the DNA fragments with a selected length are first circularized by merging the two ends, then the DNA loop is fragmented again down to a desired size, the merged ends enriched, and the adaptors ligated to their opposite extremes, that will now be separated by a known distance.

However, whenever possible, a combination of different libraries can be useful in order to obtain the most complete information out of our genome, including both sequence and structural variants, and usually allowing a better assembly.

In addition, there are useful library preparation kits specifically designed to build "scaffold sequences" evenly spaced at large (>10kb) intervals, which can constitute an even better backbone for *de novo* genome assemblies.

Another important point to be addressed when planning a NGS experiment is the number of samples to be sequenced. Of course, the available resources (in terms of time, money and platforms) are the limiting factors. However, with the same exact funds, it is often possible to choose between running a single sample at a higher coverage versus running multiple samples at a lower coverage and, in this case, whether joining all the source DNAs or tagging the samples individually. Of course there is no universal answer to this problem, since it should be addressed for every single project. Generally speaking, the *pros* of sequencing a single genome at a higher coverage are the possibility of calling with a higher accuracy the heterozygote sites (given the error rates of the platform), of retrieving phase (haplotypes)

information, as well as the individual levels of expression in case of a gene expression study. On the other side, it will only allow the detection of sites (SNVs) that are heterozygous within that given individual, but may not be representative of the population variability. The opposite will be true if we decide to pool the samples from multiple individuals in the same run. Of course, the pros of both methods, with the exception of the individual coverage, will be retained in case we have the possibility to individually tag the different samples, giving us the possibility to reconstruct the sequence of every specimen using bioinformatics tools, but at the same time obtaining information about the inter-individual variability. In this case, the limiting factor is usually the number of tags that can be provided by the manufacturer (and their cost), in combination with the number of subdivisions that can be organized on the sequencing plates (lanes, gaskets, etc.)

Whatever the library preparation and number of samples chosen, once we get our sequence data the following step is to decide how to align our fragments. In order to do that, the two most common strategies are mapping to a reference genome or opting for a *de novo* (or self) assembly.

In the first case, the genome of a similar species should be already available and assembled. So far, the number of species is limited, especially among non-model organisms; however, projects such as the 10k Genome Project (G10K, <http://www.genome10k.org/>) “aims to assemble a genomic zoo - a collection of DNA sequences representing the genomes of 10,000 vertebrate species, approximately one for every vertebrate genus” (but a similar 5k Genome Project has been recently launched also for insects), therefore promising to widely expand the number of reference genomes that will be available in the next years. In this case, the software will “simply” try to align all our reads to the most similar region of the reference genome. Of course, giving a genome size in the order of magnitude of 10^9 bp, and a comparable number of reads for the current highest-throughput platform, this operation is far from simple, and in the next chapter we will see some of the problems related to the bioinformatics pipeline and hardware needed.

However, a much more complex procedure is represented by the self assembly of our reads without the support of a reference genome. In this case, our reads have to be compared one to the other, and the ones that are (almost) identical in a given portion will be used as starting point to build ‘contigs’ (sets of overlapping reads) and ‘supercontings’ (the largest contigs the software was able to reconstruct). Theoretically, we should be able to reconstruct a continuous sequence for each chromosome, but even with the most complete library preparation, that includes a proper scaffolding support, this is rare to achieve, with most

studies only reaching a large number of independent contigs. Of course, the number of combinations to be performed in order to compare every possible pair of reads is enormous, but the software dedicated to this task is improving at a fast speed, and the development of new informatics' methods is strongly supported also through competitions and prizes.

In the cases where a reference genome is available, but not being so evolutionarily close to the target species, a new approach, called assisted assembly, combines the support given by a reference genome to flexibly improve the accuracy and the speed of a *de novo* assembly.

The problems

When approaching NGS tools for the first time, the attention of the researcher will probably focus on the platform to choose, on the application that suits best the project aim, and on the sample preparation. However, additional problems will have to be faced in the early phases of a sequencing project other than these, and it is often harder to find appropriate information about them (Flicek and Birney 2010).

Whereas passing from the imaging step (that is, the primary output of the machine) to sequence data is usually a problem already coped by the manufacturer, and we will not have to directly deal with signal intensity, spot overlapping or background noise in the raw images, we will soon have to perform some basic quality controls on our fresh sequence data.

First of all, we will surely want to know what is the total output of our sequencing run, since it can widely vary on the base of the library preparation method, DNA quality, and run performance. Beside that, it is important that the average read length matches the expected one, and is as much as possible normally distributed, therefore excluding systematic errors during some phases of the process. After the assembly or mapping step, that we will discuss later in this paragraph, we will be interested in seeing what is the actual coverage of our genome, that is, on average, how many times a given base has been independently read by the machine. However, even if the average coverage can be satisfying, its variation can strongly affect our possibility to evenly represent our complete genome. Known factors affecting the variation in coverage are genomic features such as GC content (that can mainly bias the amplification success) and mappability (also indicated as alignability; namely, the presence of a given sequence in multiple locations of our genome, therefore affecting the possibility of uniquely mapping or assembly a read falling into them).

But most of the problems raise when we have to choose our bioinformatic pipeline, that is, the combination of statistical, mathematic and computational tools applied to solve a

biological problem, that in our case is mainly given by the assembly or mapping of our genome, but also performing the quality controls, some filtering steps, retrieving the information we need out of our sequences and produce reproducible results.

We can either chose to use the software provided by the manufacturer itself, or to opt for developing analysis pipelines specifically designed for our project (in which case, we will need to have a good command of the main programming languages such as C, Perl, Python, etc. in order to be able to interface the different tools developed within the scientific community). In every case, the factors to be considered for the choice of the software are mainly the running time (especially for large projects), its flexibility (the ability to be successfully and simply applied to different tasks), the maintenance, the documentation available (that is often very limited, also on the web), its popularity (it will be easier to publish a work by using a known software rather than presenting a new tools, unless you are a recognized genius in bioinformatics), its cost (it can dramatically change, especially for small institutions, being able to access free licensed software instead of purchasing a different piece of software for every analysis to be performed), and the hardware needed. However, the absence of standardized best-practices and fast-changing rate at which new software arises (basically every few months) makes it useless to deeply talk about the current available software in this dissertation, and it is better worth addressing the reader to the most recent publications at the time he will be interested in starting a NGS project.

On the contrary, we would rather spend some more words about the hardware infrastructures, which can be a limiting factor both in terms of computational performance and data storage.

Just to have an approximate idea, the primary output (the series of images produced during the sequencing run, the so-called Real Time Analysis, RTA) are the heaviest files and can weigh up to several terabytes, but they are usually discarded after they are turned into row sequence data, e.g. into .qseq, .fastq or .scarf formats.

But whereas the initial steps (primary output management) can be usually performed on the computers provided with the platform itself, the downstream analyses can still represent a bottleneck for the infrastructures commonly available to average-sized research institutions.

In fact, the complexity of the operations, especially the computational power for the assembly phase, usually requires dedicated facilities, multiple processors with dozens gigabytes of RAM, better if organized in a server, and terabytes of storage space for all the intermediate files that will be produced during the analyses. However, these facilities do not need to be necessarily on-site, but can be remotely accessed through common computers, better if speaking a common language such as UNIX.

In chapter 3.1, we will see more specifically what is the hardware and software utilized in a state-of-the-art sequencing project like the one performed at UCLA on the wolf genome.

Public Databases

In order to give the scientific community access to the sequence resources published by other groups in a coherent framework, several institutions organized publicly-available databases hosting sequence information and accessible through the net. Originally, they were hosting gene or transcript information produced by traditional Sanger sequencing, but nowadays they are trying to keep up with the impressive amount of data produced every year by NGS projects.

One of the best-known is GenBank (<http://www.ncbi.nlm.nih.gov/genbank/>; Benson et al. 2011), supported by NCBI and NIH. Its “annotated collection of all publicly available DNA sequence” nowadays accounts for “approximately 126,551,501,141 bases in 135,440,924 sequence records in the traditional GenBank divisions and 191,401,393,188 bases in 62,715,288 sequence records in the WGS division as of April 2011”. With its European (the European Molecular Biology Laboratory, EMBL) and Japanese (the DNA DataBank of Japan (DDBJ) counterparts it constitutes the International Nucleotide Sequence Database Collaboration. It also provides several sequence search and matching tools.

A particular focus to the genome automatic annotation issues is given by ENSEMBL, a joint project between European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute (WTSI). Its online interface, BioMart (www.ensembl.org/biomart/martview/), gives easy access to the available information.

Other genomic resources can be easily accessed online through the Genome Bioinformatics website set up by the University of California, Santa Cruz (<http://genome.ucsc.edu/index.html>), which hosts a Genome Browser that allows the graphical visualization of many features of all the complete genomes published up to date (P. a Fujita et al. 2011), beside other useful bioinformatics tools.

However, the managers of one of the most complete database are concerned about the future upload all the sequence information coming from the most recent genome sequencing projects, highlighting once again one of the limiting factors that will affect the NGS explosion in the next few years.

In the end, before considering a next-generation sequencing approach, other useful tools can be used in order to address many relevant biological questions at a genome-wide scale.

Just as an example, standard or customized SNP arrays have been successfully applied to describe a significant portion of the genetic variability intra or inter species, to detect signals of selection and to associate phenotypic traits to their causal variants (or at least identify their genomic positions). Of course, their development requires the previous knowledge of variation in the genome sequence; therefore they also widely benefit from the rise of NGS techniques.

1.3.The species: *Canis lupus*, Linnaeus 1758

It's always difficult to talk about the wolf in a strictly scientific way.

The idea of wolf, that is often very different from its real essence, always leads to strong feelings in people who have to deal with it.

And I am not immune to this.

The wolf (*Canis lupus* Linnaeus 1758) is one of the most fascinating -but at the same time hated- species all over the world and the links between wolf and human beings have always been incredibly close.

In many cultures the wolf was considered the ancestor of the whole population. It's the case of the legend on the origin of Romans, in which a female wolf looked after Romolo and Remo and saved them from starvation: 2500 years afterward, her statue is still the symbol of Rome. But the wolf is also considered the ancestor of Turkeys and is the totem animal of Mongolians and many Native Americans populations, on the opposite sides of the world.

Many kings and emperors chose to have a wolf on their effigy as a symbol of power and intelligence. Wolf is a symbol-species everywhere, indeed, and we can find its presence also in tales and allegories.

However, the idea of wolf changed throughout history (Ortalli 1988).

For the ancient Greeks and Romans it often represented a sign of *pietas* (adhesion to gods' will), but in the European Middle Age, when deep changes occurred in the organization of societies and in the way people looked at nature as a whole, it was assumed as an image of the devil itself, leading to adverse feelings and large persecutions.

Nowadays, with a much looser relation between people's everyday life and the environment, wolf has become a perfect character in cartoons, which strongly contributed to give a less frightening image of it, but once again far from its real nature.

1.3.1. Origin and distribution

The gray wolf (*Canis lupus* L.1758) is a carnivore belonging to the family of Canidae.

DNA sequencing (Vila et al. 1997, Leonard et al. 2002, Lindblad-Toh et al. 2005) and phylogenetic studies indicate that the gray wolf is the only ancestor of domestic dogs (*Canis familiaris*, or *Canis lupus familiaris*) (fig. 2).

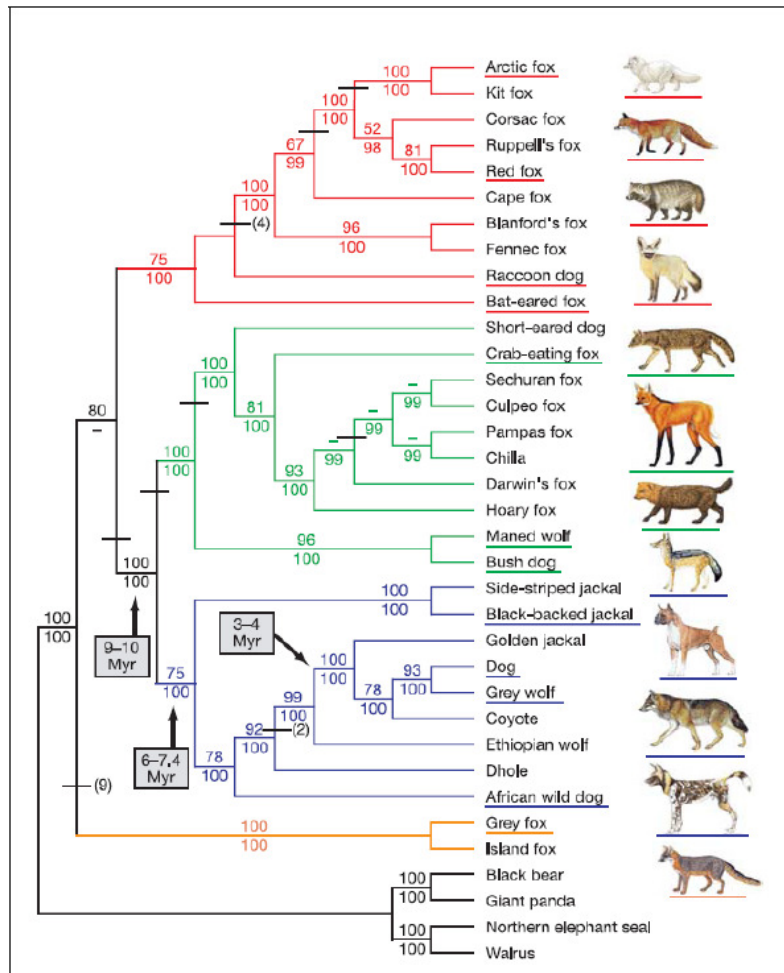


Figure 2. Phylogenetic relationships among the *Canidae* lineages inferred from nuclear sequence data (from Lindblad-Toh et al. 2005).

Probably wolf-like canids had their origin in Africa, since the two African jackals are the most basal members of this clade. South American *taxa* represent another large group of canids and are clearly rooted on the two most morphologically divergent species, the maned wolf and bush dog; the red fox-like canids, which are rooted on the fennec fox and Blanford's fox, also include the raccoon dog and the bat-eared fox. The grey fox lineage seems to be the most primitive and could suggest a North American origin of the living canids, which probably appeared about 10 million years ago.

According to these results, the first species of the genus *Canis* could have originated during the late Miocene, from 9 to 4.5 million years ago (Nowak 2003).

Canis sp. (Palombo et al. 2008) has been recorded in Africa at about 3.5 Mya and possibly a large-sized form could be present at Laetoli at about 3.7 Mya. Members of the genus *Canis* are thought to appear in Europe in the Late Pliocene or even in the Middle Pliocene. Fossil

record of the species *C. lupus* (Sommer and Benecke 2005) was found for the first time in Europe in assemblages of the Saalian Glacial by a very robust form, confirming the Palearctic Region to be the geographic origin of this species. The wolf is a member of the Late-Pleistocene *Mammuthus-Coelodonta* faunal complex and was possibly distributed in all parts of Europe during the Late-Pleistocene. Probably *C. lupus* ancestor originated in North America, moved to Eurasia and went back to the New World, which could have been reached several times (Leonard et al. 2002).

Wolves are highly adaptable and widely distributed in ecosystems ranging from Arctic tundra to Arabian deserts in the Old and New World (Mech 1970). Field observations, as well as population and genetic studies, indicate that wolves may disperse rapidly over long distances, either by recurrent dispersal or during waves of population expansion (Vila et al. 1999). Expanding wolf populations have rapidly recolonized suitable areas of their historical range in North America and Europe, and occasional events of long-distance dispersal have been described (Lucchini et al. 2002, Valière et al. 2003, Ciucci et al. 2009). However, permanent physiographic traits or anthropogenic habitat fragmentation may limit individual dispersal and gene flow. Wolves do not expand in agricultural landscapes, which, in contrast, are commonly used by other canids, like coyotes in North America and jackals in Eurasia (Wayne et al. 1992). Wolves were presumably widespread almost everywhere in Eurasia throughout the Holocene (Boitani 2000).

Human persecution, deforestation and the decrease of natural preys led wolf populations to decline in Europe during the last centuries (Delibes 1990). Large populations survived in the Balkans and Eastern Europe, while the species was eradicated in central Europe and Scandinavia, and only survived in fragmented populations in the Iberian and Italian peninsulas.

Studies on the control region of mitochondrial DNA (Randi et al. 2000, Vila et al. 1999) show an unexpected distribution of different haplotypes in Europe and Eastern Russia, probably due to several contraction-expansion periods and migrations, instead of a strictly geography-dependent scheme (Vila et al. 1999).

Italian wolves (Randi et al. 2000) have a mitochondrial haplotype (W14) that is unique (this fact partially supports the hypothesis made on morphological bases by Altobello in 1921 - almost one century ago – on the existence of a distinct Italian subspecies: *C. lupus italicus*).

The causes of this process have been accurately described by (Lucchini et al. 2004). Wolves disappeared from the Alps in the 1920s and drastically declined in Italy in the two decades after World War II. By 1973 there were approximately 100 surviving individuals, isolated in

the central Apennines (Zimen and Boitani 1975). Legal protection and the expansion of natural prey populations contributed to revert the wolf decline, and a census in 1983 suggested the presence of about 220 wolves (Boitani 1984). Thereafter, wolves expanded rapidly along the Apennines ridge, recolonizing the Western Italian and French Alps in 1992 (Breitenmoser 1998, Lucchini et al. 2002, Valière et al. 2003). Ciucci and Boitani (1991) estimated an annual population increase of 7% from 1973 to 1988, leading them to argue that wolves in Italy should now number about 600 individuals, although current estimates can better suppose the existence of about 1000 wolves.

Wolves in the Apennines (Lucchini et al. 2004) could have been, at least partially, genetically isolated from any other wolf population in Europe for some thousands of years, and not just for a few decades, as suggested by information on the species' historical distribution range. The Alpine ice caps at the last glacial maximum might have provided a geographical barrier that isolated wolves in refuge areas south of the Alps. Deglaciation and the expansion of extant ecosystems were completed only after the Younger Dryas cold spell (*c.* 10 000 years ago, Dawson 1996). Thus, the admixture of wolf populations expanding from different glacial refuges could have been relatively recent. Moreover, the Po River, which cuts the plain from the western Alps to the Adriatic Sea, was much more expanded during the last glaciation, because of the lower sea level and the presence of a north Adriatic land-bridge (Dawson 1996). For thousands of years in the Holocene, the Po River basin was flanked by extensive flooded alluvial plains and marshes, which were partially drained only in the last 2000 years (Sereni 1961). Admixture of Alpine and Apennines wolf populations could have been prevented also by deforestation and the concomitant eradication of wild ungulate populations, which were already widespread during the fifteenth century in northern Italy as a result of expanding sharecropping agricultural systems (Sereni 1961).

Despite the high potential rates of dispersal and gene flow (Lucchini et al. 2004) local wolf populations may not mix for long periods of time. Wolves from the Apennines are currently expanding, recolonizing parts of their historical range in the western Italian and French Alps (Lucchini et al. 2002, Marucco et al. 2009). Meanwhile, from the east, wolves with distinct mitochondrial haplotypes are moving from Slovenia towards the Italian border in the eastern Alps. It will be interesting to observe whether wolves expanding from the west (bearing Apennines haplotypes) and from the east (with Balkan haplotypes) will mix during the ongoing process of natural recolonization of the Alps.

1.3.2. Morphology and biology

Wolf weight and size can greatly vary worldwide.

In general, height varies from 0.6 to 0.95 meters at the shoulder and weight ranges from 20 to 62 kilograms. In Italy (Ciucci and Boitani 1998) the average weight of an adult male usually varies from 25 to 35 kg and it never overcomes 45 kg.

Wolves can measure from 1.3 to 2 meters from the nose to the tip of the tail, which itself accounts for approximately one quarter of the overall body length. The most remarkable dimensions can be found at high latitudes, with a maximum at about 60 degrees north

Wolves present sexual dimorphism, since females typically weight 20% less than males. They also have narrower muzzles and foreheads, smoother legs and less massive shoulders.

Wolves can cover long distances trotting at a pace of about 10 km/h, but can reach speeds approaching 65 km/h during a chase.

Wolves are digitigrades and their paws are able to tread easily on every kind of terrains. There is slight webbing between toes (Ciucci and Boitani 1998), which allows them to move on snow more easily than many preys. The front paws are larger than the hind paws, and have a fifth digit, the dewclaw, that is absent on hind paws, but never touches the ground.

The anatomical location of blood vessels – which allows a counter-current heat exchange - preserves paw pads from freezing and helps saving energy in cold climates. The same system has been maintained in ancient domestic dogs, such as Siberian husky.

Scent glands located among wolves toes leave on the ground trace of chemical markers, helping the wolf to orientate over wide territories and, in the meanwhile, to inform the other wolves of its position.

The coat of wolves consists of two layers: the first one is composed of tough guard hairs that repel water, while the second one is a dense undercoat that well insulates from external temperature. The undercoat is shed once a year in late spring or in early summer (Ciucci and Boitani 1998), increasing again from early winter months.

The coloration of the fur varies from gray to gray-brown, passing through white, red, brown, and black, sometimes according to the ecological adaptation to the habitat, as it occurs at the interface between taiga and tundra (Musiani et al. 2007).

Wolves have distinct winter and summer pelages that alternate in spring and autumn.

Italian wolves usually have black tips on tail and ears and black lines on the front legs, while abdominal parts are lighter or cream-white, as well as the face mask (Ciucci and Boitani 1998). This feature helps to emphasize certain gestures during social interactions.

Black wolves only occur in North America and Italy. Here, only few individuals have been found to have a completely black fur, which for a long time has been considered as the effect of introgression from dogs of the causative gene, β -defensin 103 (T. M. Anderson et al. 2009). However in most of the black wolves in Northern Apennines (Apollonio et al. 2004) no trace of hybridization was found, suggesting that the black coat colour can also derive from a natural combination of wolf alleles, or that the hybridization traces back thousands of years (T. M. Anderson et al. 2009). Compared to dogs, wolves show anatomical differences in the orbital angle (>53 degrees for dogs, <45 degrees for wolves), a lower frontal step, larger skull and brain capacity, as well as relatively larger paw size. Yellow eyes, longer legs, the presence of pre-caudal glands and longer teeth are other distinguishable features. Talking about dentition, the formula is I3, C1, P4, M2 / I3, C1, P4, M3 (Ciucci and Boitani 1998); the fourth upper premolars and the first lower molars are named carnassial teeth, specifically evolved for shearing flesh. The long canines (20 to 23mm), used to catch and hold the prey, can deliver a pressure up to 10,000 kPa. Teeth injuries are a serious danger for wolves, sometimes leading to starvation and death.

Hunting techniques rely on pursuit, which allows wolves to make a strong selection on physical and health conditions of preys: by choosing the most vulnerable ones, they can save energy and in the meanwhile effect a positive selection on the prey population.

The oestrus occurs once a year (whereas in dogs it occurs twice) and the mating takes place between January and April (in Italy, generally in March), according to the latitude and to the photoperiod, which regulates the hormonal production (Kreeger 2003): the higher is the latitude, the later it occurs.

The alpha pair is the only one to mate and, since a pack can usually support only one litter a year (the wolf has evolved as a K-strategy species), this dominant behaviour is beneficial in the long run and allows a continuous adaptation to the environment.

The gestation period lasts 60-64 days (Packard 2003) and adult females produce about 4-6 pups (with documented variations from 1 to 11, Mech 1974).

The pups, which weight about 0.5 kg, are born blind and completely mother-dependent. The father, often helped by others relatives or pack members, protects the home-site and carries food for the mother (Mech 1999). Pups reside for two months in the den, which is usually placed on high ground and near an open water source (Joslin 1967, Mech 1970), often at the center of the pack's territory, to minimize the hunting effort and the pups exposition to other packs (Mech and Boitani 2003). Its features can change according to the habitat and to the

ground type and it can be adapted from other species' den, or in a rocky cave. The pups begin to eat regurgitated food two weeks later (Packard 2003), when their milk teeth have emerged. Two months later, pups are moved to a *rendezvous* site (Joslin 1967), a safe place where to stay and wait for the adults during the hunts, until they will be able to join the chase (at about the 8th month of age).

The pups' fights for eating privileges produce a secondary ranking among them and practice them to the dominance/submission rituals.

Young wolves reach sexual maturity at two or three years, when many of them (mainly the males) leave their birth packs and look for their own territories where they establish and mate. Into the wild, wolves generally live from 6 to 8 years, although in captivity they can live up to twice that age.

The mortality rates in the wild are high. Pups die for food scarcity, pathogens or for falling prey of other predators. The most significant causes of mortality for adult wolves are human hunting and poaching, car accidents and wounds inflicted while hunting prey. Rival wolf packs are often their most dangerous non-human enemies, as 14–65% of wolf deaths can be inflicted by other wolves (Huber et al. 2002).

Wolves are social animals and communication plays a great role in every moment of their life. Wolves can communicate in several ways that can be grouped in acoustic, visual and olfactory communication.

The howling is the most widely known means of communication among wolves (and can be reasonably considered one of the main sources of human fear toward wolves). The howling is a deep sound, whose fundamental frequency can range from 150 to 789 Hz, up to the 12th superior harmonic (Theberge and Falls 1967). Different wolves can howl in different ways (Ciucci and Boitani, 1998), as well as different populations can use the howling differently. Howling allows the pack members to keep in touch through forested areas or over great distances and also to meet in a specific location before a chase (Harrington and Asa 2003). Howling is important as a declaration of territory, as shown in a dominant wolf tendency to respond to a human imitation of a "rival" wolf in an area that wolves consider their own. Wolves will also howl for communal reasons, as to strengthen the social bonds. A wolf howl can be heard for several kilometers, depending on weather conditions. Wolves howl more frequently during the breeding season (Harrington and Fred 2000) and in the first half of winter.

Other acoustic ways to communicate are represented by growling, barks and rallies.

Growls are signs of warning or threaten and are usually associated with a visual signal, whereas barks can denote a nervous mood and are much less frequently used than in dog's communication. The rally is a high pitched noise that is often used when the wolves of a pack meet, or even to denote submission.

Visually, wolves are always communicating one to another through body language, which comprehends body carriage, tail and ears postures and facial expressions, usually enhanced by the light mask around the muzzle. A large combination of these coded signals can share feelings and underline hierarchical relationships, as dominance, submission, anger, fear, aggression or defensive attitude, suspicion or tension, but also relaxation, happiness or playfulness.

Another crucial form of communication in wolves is the olfactory-based one (which is probably the most difficult for us to understand and even imagine), since smell is the most developed sense in wolves, with about 1000 genes dedicated to this function (Tacher et al. 2005).

Scent glands are present all over the body, especially at the base of the tail or among toes (Harrington and Asa 2003). Pheromones secreted by these glands can identify each single wolf, its health conditions, and its social and reproductive status. Alpha wolves scent-mark frequently, with both faeces and urine. Male and female alpha wolves usually urine-mark objects with a raised-leg stance (RLP) in order to enforce rank and territory, whereas other pack members usually squat. Defecation markers are particularly useful for spatial navigation and are often deposited along frequently used paths or in important crossroads (Barja et al. 2004), keeping the pack from traversing the same terrain too often and allowing each wolf to know of the whereabouts of its pack members. Ground-scratching is the main way to depose the scent of the inter-toes glands. Above all, scent marking is used to inform other wolves and packs that a certain territory is occupied.

Wolves live in packs (Mech 1970), hierarchically ruled social units that can comprehend from two to ten individuals (even more at high latitudes). Living in pack allows wolves to reach a good hunting and reproductive success. The ranking within the pack can be defined as a linear dominance hierarchy and the two individuals that lead the pack (one for each sex) are also called alpha male and alpha female. They are usually monogamous until the death of one of them and they are the only individuals that can reproduce in a pack, although multiple litters have been documented (Vonholdt et al. 2008). The members of the alpha pair have the greatest control over food resources, but also keep the pack cohesive and functional (Mech 1970), leading it in the everyday decisions and in territory defence, especially by the male

(Packard 2003). The ranking is decided on the base of ritual, agonistic behaviors and it can change from year to year, in particular before the reproductive period.

All the wolves of the pack assist in raising the pups. Some mature individuals can choose not to disperse and stay in their original pack helping rear pups.

As we said, usually wolves packs (Packard 2003) are considered to be organized as a linear hierarchy, but the concept of family may better describe the relationships and the dynamics among its members. A pack, indeed, is always composed by the mating pair, their pups (if any) of the year and of the prior years, plus some external individuals (the adoptees) that can come by dispersal events from other packs (Mech and Boitani 2003).

Dispersal reduces the resource exploitation in a single territory, prevents inbreeding and promotes natural selection and cross-breeding.

The size of the pack may change over time according to several factors, including habitat, food supply and even personalities of individual wolves within a pack. New packs are formed when a wolf leaves its birth pack, finds a mate, and claims a territory (Rothman and Mech 1979).

Territory size, as well as the number of members, can greatly vary (20 km^2 - 4335 km^2) and is negatively correlated with the available prey biomass, which is related to the habitat and also to the latitude. In Italy, the territory size can range from 20 to 300 km^2 (Apollonio et al. 2004; Ciucci and Boitani 1998).

Although the main prey is represented by large herbivores (wild boar, roe deer, red deer, moose, mouflon, even bison), which are chased with the cooperation of the whole pack and accurate attack techniques (e.g. at first to the legs, then to the neck, with precise bites that produce an hypo-oxygenation shock), they can also hunt rodents and other small animals. In northern America the bison is the largest prey that wolves use to hunt, while in northern Europe it is the moose. Hunting success seems to be related to the pack size and to the presence and age -therefore the experience- of an adult male (Sand et al. 2006). In Italy, also garbage has represented an important food source during the years of maximal reduction.

As keystone predators, wolves have a great impact on the trophic network, but at the same time they are also vulnerable to prey fluctuations.

Predations on livestock are quite common in rural areas, and surplus killing has been documented, but not completely explained.

1.3.3. Threats and legal status

As we saw, wolves have been progressively eradicated throughout Western Europe and in the Alps in the 18th and 19th centuries (Breitenmoser 1998), surviving in fragmented populations in Iberian peninsula and Italy (Boitani 2003). Wolves in Italy were confined south of the Po River since the turn of the last century, continuing to decline until the 1970s, when approximately 100 individuals ranged in two fragmented areas in central-southern Apennines (Zimen and Boitani 1975). The Italian wolf population suffered severe persecution until 1971, when wolf hunting was stopped and poison baits banned. This change of attitude was completed in 1976 when the species was given a fully protected status. This process was stimulated by WWF International that funded a long-term project called “San Francesco”, including a public educational campaign, scientific works and management solutions to protect wolves. Due to the more effective legal protection and, above all, substantial changes in the ecology of mountain areas (e.g. decrease of human density and increase of wild ungulates), this declining demographic trend quickly reversed in the 1980s, when wolves started to expand in Italy and in other European countries (Breitenmoser 1998; Boitani 2003). In Italy wolves crossed the northern Apennines and recolonized the south-western Alps, where genetic identification confirmed their presence in France and in Switzerland (Fabbri et al. 2007, Lucchini et al. 2002, Valière et al. 2003), and reappeared again in the central Italian Alps in 2000. Few years ago the Italian wolf population was guessed to number more than 600 individuals (Boitani 2003), being now probably closer to 1000 individuals.

The wolf is considered a species of Least Concern (2007 IUCN Red List of Threatened Species) by the World Conservation Union, but it is currently legally protected in Italy through international law, European law and Italian law.

For every detail, we refer to the National Action Plan for Conservation of Wolf (Genovesi 2002), that collects the best knowledge and practices on the conservation of wolves in Italy in order to coordinate the actions for its management.

The wolf is protected under international law, primarily under the Bern Convention on Conservation of Wildlife and Natural Habitats (1979), in appendix II (Strictly Protected Species). The convention forbids its catching, killing, possession and trade. However, many countries in Eastern and Northern Europe refused to fully protect the wolf, and Spain has recently authorized its hunting.

Also the Convention on International Trade of Endangered Species of Fauna and Flora (CITES, Washington 1973) strictly protects several wolf populations (the ones from Bhutan, India, Nepal and Pakistan) in Appendix I (species threatened with extinction which are or

may be affected by trade) and gives a lower protection to all the other populations in Appendix II (species that are not necessarily now threatened with extinction, but may become so unless trade in specimens of such species is subject to strict regulation).

The wolf is protected through European Law by the Council Directive 92/43/EEC on the conservation of natural habitats and of wild fauna and flora (HABITAT).

Other useful references are represented by the Large Carnivore Initiative for Europe planned by WWF and the European Action Plan for Conservation of Wolf (Boitani 2000).

The Italian law has received all the recent International and European directives on protection of wolf, but over 150 wolves are thought to be willingly killed in the last 20 years in Italy, with a single case of legal incrimination (Caniglia et al. 2010).

In the United States of America, the wolf has been recently delisted from the Endangered Species Act. As a consequence, after 25 years from the first reintroduction in Yellowstone, wolf hunting is allowed again outside the National Parks.

Human-caused mortality represents one of the main problems in conservation of wolves.

Recent studies (Lovari et al. 2007) on 154 carcasses found in Italy from 1990 and 2001 show that about half of the deaths is caused by road kills, and approximately 18% is related to poison or shots. Less than 15% and 10% are caused by intra specific strife and disease, respectively, although the sampling strategy could represent a source of bias (P. Ciucci et al. 2007).

Other primary conservation issues (Genovesi 2002) are competition and genetic pollution with free ranging dogs.

Habitat loss and fragmentation, human disturbance, demographic factors and range fragmentation represent other factors of concern, even though they can be considered of secondary importance.

The fear of extensive hybridization (Verardi et al. 2006) between declining wolf populations and widespread free-ranging domestic dogs in Europe has been a main concern for conservation biologists over the past 30 years (Boitani 1984; 2003, Randi and Lucchini 2002). Wolves and domestic dogs are isokaryotypic, fully interfertile and have been shown to mate successfully in captivity and into the wild when they co-occur (Wayne et al. 1995, Vilà and Wayne 1999).

Despite a substantial demographic recovery, wolves are still largely outnumbered by free-ranging dogs, which are estimated to be more than 1 million (Genovesi and Dupré 2000). There is serious concern that, as a consequence of such striking disparity in population size, the genetic integrity of wolf gene pool might be threatened.

The presence (Randi 2008) of anomalous morphological characters (i.e. black coat colour or dewclaws), has been observed in some wolves in Italy. Dewclaws (vestigial first toes) on the hind legs are common in some dog breeds, but never detected in wolves. Black wolves are widespread in some North America populations, but they were never been observed in Europe. Both these traits could have been introduced in the Italian wolf population via hybridization with free-ranging domestic dogs.

Analyses of diagnostic mitochondrial DNA (mtDNA) haplotypes failed to detect introgression of dog mtDNA in wolf populations, suggesting that either hybridization is rare or strictly unidirectional, or that F1 hybrids are not able to backcross into the wolf populations (Randi et al. 2000, Vilà and Wayne 1999).

Recent genetic studies (Verardi et al. 2006) led to identify 11 out of 220 wolf genotypes (5.0%) that were likely admixed with dogs, a proportion that is higher than in previous studies (one admixed over 107 genotyped wolves; Randi and Lucchini 2002) and suggested that dogs and wolves might have admixed during the last 70 (\pm 20) generations, that means 140–210 years (assuming a generation time in wolves of 2-3 years). Often, but not always, admixed wolves showed morphological signals of hybridization. It is interesting to notice that the admixed wolves were mostly confined to peripheral areas of the species distribution range in Italy. Despite hybridization, wolves and free-ranging dogs remain genetically distinct in Italy, suggesting that introgression in nature might be strongly counteracted by selection or by ethological factors (Randi and Lucchini 2002, Vilà and Wayne 1999). In conclusion, introgressive hybridization (Verardi et al. 2006), although perhaps protracted in time, is limited and seems to pose no serious threat on the integrity of the Italian wolf gene pool.

On the contrary (Randi 2010), in North America the presence of two groups of canids, whose morphological traits lead them to be classified as a different species (*Canis rufus*, or red wolf) or subspecies (the Great Lakes wolf, *Canis lupus lycaon*, or *Canis lycaon*), recently revealed a pattern of past extended hybridization between wolves and coyotes (Vonholdt et al. 2011), suggesting that these events, even between more distantly related species, such as coyote, can also give rise to distinct viable populations.

2. The Major Histocompatibility Complex: its variability in the Italian wolf population and its influence on mating choice and fitness traits

Studying how natural selection shapes the patterns of genetic diversity in wild and model populations has always been one of the most investigated topics in biology. Answering the underlying questions leads us to understand how the different species adapt to their environment through time and space, and -at some extent- to predict how likely they will successfully cope with future changes, such as global climate warming, habitat loss and fragmentation.

However, only a few genetic systems have been as thoroughly studied as the Major Histocompatibility Complex. Its unique properties (such as the extremely high variability and its implication in multiple biological pathways) always made it a perfect candidate for a wide variety of studies, ranging from immunology to conservation genetics and behavioral ecology.

2.1. Background

2.1.1. Structure and functions

The Major Histocompatibility Complex (MHC) is a set of genes implicated in immune response, both innate and adaptive, usually clustered together along a single or a few chromosomes.

The name takes its origin for the important role the MHC has in the tissue compatibility during transplantations (the first field where it was deeply studied and characterized), by discriminating self from non-self antigens.

In humans (MHC Sequencing Consortium 1999) it comprehends about 130 functional genes plus *ca.* 100 pseudogenes, mainly hosted on chromosome 6. However, the number of genes can widely vary from species to species, and it can be as low as 19 (Kaufman et al. 1999) in chicken (*Gallus gallus*), or an intermediate number between birds and placental mammals in marsupials, such as opossums *Monodelphis domestica*, with about 115 genes (Belov et al. 2006). Recent genomic studies (Star et al. 2011) revealed it has been largely rearranged in the Atlantic cod (*Gadus morhua*) compared to other teleosts. In dog (Yuhki et al. 2007) it includes more than 100 genes, only part of which are strictly deputed to immunity functions, mainly hosted in a telomeric 3Mb-region on chromosome 12.

The MHC genes encode for a series of glycoprotein receptors that have an important role in starting several biological pathways in response to pathogens and infectious diseases.

Their main role is to bind fragments of proteins (antigens) in the cells and to present them to the T-cells that will initiate a cascade of immune responses.

They are usually grouped into three main subfamilies, named class I, II and III, according to their structure and specific functions.

Class I molecules are implicated in the response against intracellular pathogens, mainly viruses, by binding endogenously derived peptides from proteins in the cytoplasm and presenting them to cytotoxic T-cells, possibly leading to the cell's apoptosis. Class I receptors are located on all the nucleated somatic cells, whereas class II molecules are only expressed on leukocytes, specifically on macrophages and B-cells. Class II receptors are mainly committed to the response to bacterial infections by presenting exogenous antigens to helper T-cells.

Both class I and class II proteins are heterodimers (Fig. 3), but they differ since class I molecules show three α -chain domains (each of which is encoded by a different exon from a single gene) linked to a single β -microglobulin peptide; the membrane anchoring is ensured by the α_3 -chain; differently, we find two α - and two β -chains in class II receptors, encoded by two different genes, with α_2 - and β_2 - chains attached to the cell surface.

The Antigen Binding Sites (ABS; also known as Peptide Binding Regions, PBR, or Antigen Recognition Sites, ARS), namely the sites that receive and bind the antigen residues, are ensured by the combination of specific portions of the α_1 - α_2 , or the α_1 - β_1 chains.

A third group (class III), although less well-studied, is implicated in other immune functions and consists in several proteins of the complement system but also cytokines, with roles of immune signaling, and heat shock proteins, for protecting cells from thermal stresses.

For each gene, MHC alleles are codominantly expressed, potentially leading to the incredibly high polymorphism of many MHC genes that we will see in detail in the next section.

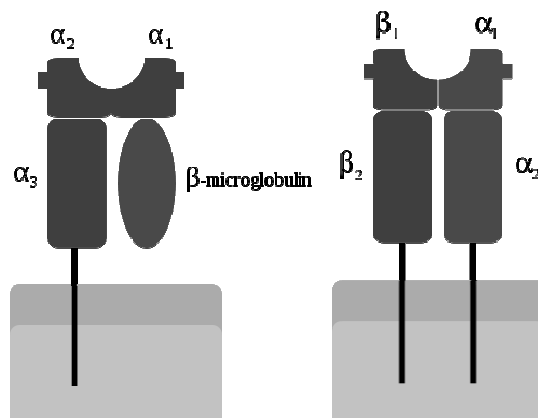


Figure 3: Schematic polypeptide structure of MHC class I (left) and class II (right) receptors (Modified from Wikimedia Commons under GNU license).

2.1.2. Genetic features and evolution

The MHC genes are among the most variable ones in the vertebrate genomes, with hundreds of alleles described in the human population to date; the metrics indicative of selective pressures (such as d_N/d_S ratio) are extreme for a set of protein-coding genes, but the levels of polymorphism and heterozygosity are exceptionally high even if compared to neutral markers. Their high variability has been explained by a number of theories (Bernatchez and Landry 2003, and therein references). The most general hypothesis is that allelic diversity at MHC genes is maintained by parasite-mediated balancing selection, therefore implying the assumption of host-pathogen co-evolution, which could be explained on the base of three models: the negative frequency-dependent selection hypothesis, the overdominance hypothesis, and the fluctuating selection.

The first one to be proposed (Clarke and Kirby 1966) was focusing on the role of selection on parasites. Since the ones able to circumvent the host defense by showing unrecognized antigens will be strongly selected, in a relatively short time they will also affect the fitness of the most common host genotypes, therefore leading to increase the relative fitness of the hosts carrying rare alleles able to defend from the parasite variant. Following this adaptive race, the allele frequency of both host and parasite will fluctuate through time, likely maintaining an extremely high polymorphism.

The second model (Doherty and Zinkernagel 1975) was taking into account the advantage of the heterozygote, better defined as the overdominance hypothesis. If a given population is exposed to a range of pathogens, the heterozygote individuals will be favored over the homozygote ones since they will be able to recognize and cope with a larger number of pathogens. On the other side, having the maximum levels of heterozygosity can not coincide with the maximal fitness, since a too high MHC diversity could lead to increased costs and inefficiencies along the downstream immune cascade (Nowak et al. 1992).

Finally, the fluctuating selection hypothesis (Hill 1991) differs from the rare-allele advantage one in considering the selection to be directional, and the parasite fluctuations to be driven by environmental factors other than the host. In this scenario, there is a fluctuating spatio-temporal variability in the pathogen distribution, therefore alternatively selecting for different MHC allele combinations. This would explain the high levels of differentiation between populations, and at the same time the genetic diversity within populations. This could occur in combination, or even without the need for, rare-allele or heterozygote advantage (Spurgin and Richardson 2010).

However, identifying which model of pathogen-mediated selection has the highest importance in shaping the genetic diversity in a given *taxon* is never trivial (Spurgin and Richardson 2010) and they are likely to act in combination.

Another range of explanations are related to the contributions that sexual selection can have in maintaining the high levels of genetic diversity at the MHC *loci*, and are however strictly linked to the hypotheses involving the role of pathogens. They consider the importance of the MHC genes in inbreeding avoidance, fitness advantages, and their possible role as ‘honest signals’ linked to other important traits.

In the inbreeding avoidance hypothesis (Potts and Wakeland, 1990), selection will favor reproductive mechanisms that lead individuals to avoid potential mates related or genetically similar to them, since this could strongly reduce their progeny’s fitness at several levels, especially in the expression of deleterious recessive mutations. In our case, MHC is likely perceived via olfaction or pheromone detection, or it influences the pleasantness of the perceived odors (Janeš et al. 2010), and acts as a potential signal of relatedness between individuals that share common MHC alleles. Disassortative mating based on the MHC system is therefore expected to be stronger in species that are more exposed to the risk of inbreeding (Jamieson et al. 2009).

Differently, models based on sexual selection as a tool for increasing the individuals’ fitness are once again mainly related to disease resistance. If it is true that heterozygote individuals have a higher resistance to diseases, parents choosing mates with alternative MHC alleles will be favored by natural selection in having a higher offspring survival. On the other side, the sexual selection could enhance the ability of the host to keep its defenses up-to-date with the parasite weapons (as for *Salmo salar* in Landry et al. 2001), as suggested by the ‘moving target’ hypothesis (Penn and Potts 1999).

As a third sexual selection-related model, it has been hypothesized that the increased fitness of the individuals that opt for an MHC-based mating choice is not directly linked to the favorable effects of MHC genes themselves, but it is likely to be caused by other genes linked to them. In this case, the MHC recognition would operate as an ‘honest signal’, in the same way the antlers indicate the general fitness of a potential mate in white-tailed deer (*Odocoileus virginianus*; Ditchkoff et al. 2001).

Beside its influence on sexual selection, the perception of MHC similarity through odors can play a role also in the evolution of kin altruism (Lewis 1998).

Nonetheless, other significant events at the molecular level could have contributed in providing the raw material for the MHC variation through time. They are mainly represented

by the processes of duplication, *interlocus* recombination and gene conversion (Van Oosterhout et al. 2006). Both mechanisms could increase the variability of the MHC gene, respectively by escaping the constraints of natural selection and by creating new combinations of sequences that are exchanged within and even across similar *loci*.

Whatever the most correct explanation underlying it, the way the variability is distributed among related species is another unique feature shown by the MHC genes, and it has been referred to as trans-species polymorphism. In other words, it occurs when a panel of alleles is shared across similar species even after a much longer time since their separation than expected on the basis of the coalescent theories. One of the explanations assumes that, if the polymorphism is mainly driven by host/pathogen selection, and the related species are still exposed to the same range of pathogens, this could lead to the maintenance of the same alleles through extended periods of time.

Recently (van Oosterhout 2009) a new theory has been proposed to explain why the MHC cluster of genes considered as a whole (including neighboring, non immunity-related genes), shows parameters that are hardly explainable by the traditional hypotheses alone, such as the strict association between pathogen susceptibility and given haplotypes, or the large differentiation between populations that coexists with the trans-species evolution. It has been named Associative Balancing Complex evolution, and it considers that given the low recombination rates in the MHC cluster and its high polymorphism, the purifying selection can act less efficiently in removing recessive mutations (the well-known Muller's ratchet effect) that will be rarely expressed and can accumulate as a 'sheltered load'. After they spread in the population, they can get to fixation and reinforce the linkage, given that recombinants will be selected against by epistatic interactions. This hypothesis requires lower selection coefficients that overdominance alone for explaining the MHC polymorphism.

2.1.3. Methods

Through time, several methods have been developed and applied to study the variability of MHC genes, ranging from standard sequencing to electrophoresis to next generation sequencing.

The most common method requires bacterial cloning of the selected genes and their traditional capillary sequencing. The cloning step can be sometimes skipped by the computational reconstruction of the most probable haplotypes (Bos et al. 2007) allowed by dedicated software. The problems raising in this case are given by possible gene duplications

(that can be relatively common in some *taxa*) and, for both the approaches, the costs of sequencing a large panel of samples. Therefore, a number of alternative methods have been developed, the most common ones being based on the physical separation of different alleles: the single-stranded conformational polymorphism (SSCP), or the denaturing gel gradient electrophoresis (DGGE). Other methods have been applied less frequently, such as Restriction Fragment Length Polymorphism (RFLP) and sequence-specific oligonucleotide probing (SSOP). However, they are not error-prone, and their efficiency can be strongly reduced in case of a large number of alleles to be discriminated, but they can be cheaper for large panels of samples. The reference strand mediated conformational polymorphism (RSCA) is the only method other than sequencing that allows the discrimination of single nucleotide differences (Kennedy et al. 2005).

Recently, next generation technologies have been applied to the study of MHC (Babik et al. 2009), by PCR-amplifying selected exons and pooling several individuals in a single run, then applying strict quality controls in order to accurately discriminate potential artifacts from true alleles. However, although the costs over hundreds of samples are strongly reduced and dedicated software has been developed (e.g. jMHC, Stuglik et al. 2011), the scalability of the experiment design is limited (Wegner 2009).

In the case we do not need to have direct access to the gene sequences, but we are only interested in estimating the levels of polymorphism to be compared between groups of individuals, and relate them to environmental or fitness measure, another useful tool (Aguilar et al. 2004) can be designing MHC-linked microsatellites. In this case, the limiting step is to identify and design specific primers around repeats that are variable in the population, which can be time-consuming in case we do not have a reference genomic sequence.

2.1.4. Studies

Whereas the first studies on the MHC focused on its importance in organ rejections during transplantations, many of the following ones investigated its roles in a range of questions. Among them, we can easily identify three main groups: the ones dedicated at describing the MHC variability in one or more species, the ones aiming at elucidating the relationship with the resistance to a given pathogen and the influence of the MHC on the fitness of individuals, and the ones investigating its possible influence on sexual selection and mating behavior.

As it commonly happens for many fields of research, the attention has not been evenly distributed among *taxa*, but it has been mainly focused on humans and, secondarily, on a

number of mammals and birds (Bernatchez and Landry 2003), and on a smaller set of teleosts. Therefore, the results obtained so far could be biased and limited to a panel of species, so they should not be automatically extrapolated to all vertebrates, and even less to the other *taxa* of the animal kingdom.

Homozygosity at MHC *loci* has been linked with higher exposure to severe human diseases, such as AIDS and hepatitis B, and its detrimental effect has been documented in many other pathologies (Horton et al. 2004).

One of the papers that mostly contributed to raise the attention on the MHC (even among the general public) was the one performed by Wedekind et al. (1995) in Bern, Switzerland. In this study, the investigators typed at three MHC *loci* a panel of students of both sexes, then asked the young women to score the pleasantness of the odor of the t-shirts worn by a set of males. The results indicate that the preference was markedly higher for potential mates having a dissimilar MHC allele set (this trend being reversed in the women using oral contraceptives, which simulate a status of pregnancy), therefore suggesting the MHC can be actively implicated in human mate choice. This would be coherent with several other studies indicating a higher incidence of abortions and a lower fertility in couples with a larger sharing of MHC antigens (Berger et al. 2010).

However, several following studies (reviewed in Havlicek and Roberts 2009) obtained contrasting results, sometimes with significant results only in female odor perception (Santos et al. 2005), indicating that the role of MHC on the mate choice in humans has not been univocally addressed, or that its influence can vary across populations, with an expected influence of their level of inbreeding (Piertney and Oliver 2006).

The same controversies also appear from studies on model (mouse) and non-model species, with particular attention to non-human primates (Setchell and Huchard 2010).

In a semi-free-ranging population of mandrill (*Mandrillus sphinx*), whose social system is based on polygyny, the probability of reproduction for males increased when both overall genetic and MHC dissimilarity with the mother increased, although reproductive success also increased with the male background heterozygosity (at microsatellite *loci*) and MHC diversity (Setchell et al. 2010).

Trying to clarify the role of CD8 T lymphocyte immune response to HIV, O'Connor et al. (2010) examined viral loads in Mauritian cynomolgus macaques (*Macaca fascicularis*) infected with simian immunodeficiency virus (SIV). They found clear evidence of heterozygote advantage, since chronic viremia in MHC-homozygote macaques was 80 times higher than in MHC-heterozygote macaques.

One of the most recent works (Thoss et al. 2011) showed in a semi-natural mouse population (*Mus musculus domesticus*) an increased fitness in individuals with higher heterozygosity at two MHC *loci*, especially in combination with intermediate levels of background heterozygosity. The reproductive individuals show higher heterozygosity at the MHC compared to non-reproductive ones, suggesting an increased fecundity or mating success.

In birds, a female ‘mating-up’ process (Griggio et al. 2011) has been shown for house sparrows: in this case, there is no evidence for females to chose mates with a high MHC diversity, unless they show a limited variability themselves, in which case they can have a clear gain by selecting the most heterozygote partners.

Previously, Richardson et al. (2005) similarly showed that in the Seychelles warbler (*Acrocephalus sechellensis*) there is no direct MHC-based mating choice, rather an increased probability of females having low MHC variation to gain an extra-pair paternity with males possessing higher MHC variability than the pair mate. Further studies on the same species (Brouwer et al. 2010) showed an association between MHC diversity and juvenile survival, particularly enhanced when a given allele was possessed. Ekblom et al. (2010) found that in the great snipe (*Gallinago media*) certain MHC alleles were associated with higher male mating success, especially if locally adapted, whereas there was no evidence of enhanced reproductive success for males with locally rare alleles (contrasting with the rare-allele advantage).

In teleosts, several studies indicate evidences of direct MHC-based mating choice. Neff et al. (2008) showed that in Chinook salmon (*Oncorhynchus tshawytscha*) females preferentially chose mates that allowed producing offspring with greater genetic diversity at the MHC, but without preferences with respect to male background genetic relatedness.

In Quebec rivers, Dionne and colleagues (2009) found that salmon (*Salmo salar*) susceptibility to myxozoan infection was correlated to the frequency of specific alleles, supporting the hypothesis of pathogen-driven, rather than heterozygote, advantage. The same team also showed (Dionne et al. 2007) that amino acid variability at the MHC, especially in the ABS, increased with the bacterial pressure, that is proportional to the river temperature, suggesting local adaptation driven by the amount of pathogens. Combining artificial crossing experiment followed by reintroduction with observations in the wild, Consuegra and de Leaniz (2008) observed that the offspring of wild salmon were more MHC-dissimilar than the ones produced by artificially crossed salmon, and that fish more dissimilar for MHC were carrying a lower parasite load. The authors’ conclusion is that disassortative MHC-based mate

choice and parasite-driven selection act in combination to maintain MHC diversity and individual fitness.

In three-spined stickleback (Eizaguirre et al. 2009), female sticklebacks preferred to mate with males sharing an intermediate MHC diversity compared to their own MHC profile, but also that a given MHC haplotype in males was associated with body size and resistance to a common parasite, and enhanced the probability of being chosen by the females and increasing the offspring. This finding gives additional support to the fact that assortative mating is a means to support 'good genes' and respond to parasite-driven selection.

2.1.5. MHC in canids

As expected, MHC has been thoroughly studied in canids as well.

However, most of the works (Angles et al. 2005, Fliegner et al. 2008, Francino et al. 1997, L J Kennedy et al. 1998, 1999a, 1999b, 2005, Runstadler et al. 2006, Wagner et al. 1996) investigated the variability of the MHC in dogs (commonly named DLA, for Dog Leukocyte Antigens), discovering from dozens to hundreds of alleles at the most variable *loci*, with more than a thousand dogs from tens of different breeds analyzed so far using different methods. However, the levels of diversity within single breeds were sometimes limited (Angles et al. 2005), whereas significant variation is retained in some feral dog populations (Runstadler et al. 2006).

This large amount of data required to be organized according to standardized methods (Ellis et al. 2006, Kennedy et al. 2001, 1999, Robinson et al. 2003), although some levels of overlap and discrepancy still remains (present work, chapter 2.3-2.4).

MHC variation has been newly associated with a number of diseases, such as leishmaniasis (Quinnell et al. 2003), hypothyroid disease in Doberman Pinscher (Kennedy et al. 2006), canine transmissible venereal tumor (Murgia et al. 2006), canine juvenile generalized demodicosis (It et al. 2010), chronic superficial keratitis in German Shepherd (Jokinen et al. 2011) and canine necrotizing meningoencephalitis in pug dogs (Barber et al. 2011).

A more limited number of studies focused so far on wild canids.

In wolves, Seddon and Ellegren (2002) investigated the variability of MHC class II *loci* in some European populations, comparing it to that of North American wolves and dogs. The high amount of variation (with up to 17 alleles found at DRB1) was mostly shared between dogs and wolves at *locus* DQA1, with some traces of trans-species polymorphism with coyotes (*Canis latrans*) and a likely past recombination event between *loci* DRB1 and DQB1.

Subsequently (2004) the authors also investigated the evolution of the same *loci* in the Scandinavian population, finding a reduced number of alleles per *locus* (coherent with only three founders for the whole population) and no traces of departures from neutrality, therefore concluding that bottleneck, fragmentation and genetic drift were masking or excluding evidences of balancing selection.

Furthermore, Berggren and Seddon (2005, 2008) explored the promoter regions of the same *loci*, finding variation at the DQB1 promoter in wolves, plus traces of balancing selection. There was a strong linkage with exon 2 alleles, but a weaker haplotype association was found in dogs than in wolves, suggesting different selective pressures and a possible reason for some common dog autoimmune diseases.

In the highly endangered and bottlenecked Mexican wolf (*C. l. baileyi*) population (Hedrick et al. 2000), some variability at DRB1 *locus* was retained, with five different alleles (although a single one was found in one of the three extant lineages). They also showed (Hedrick et al. 2003) a strong correlation between heterozygosity and resistance to canine parvovirus or canine distemper outbreak in the reintroduced population.

North American wolves have been exhaustively studied by Kennedy et al. (2007), finding many new alleles but a limited sharing between wolf and dog haplotypes, leading to interesting suggestions about the possible dog ancestors.

In closely related species, coyote and red wolf (*Canis rufus*, whose admixed origin has been recently clarified in Vonholdt et al. 2011), most of the alleles found in the latter were also present in the former, showing a higher contribution of coyotes than wolves to the gene pool of red wolves, and the single private allele found in red wolf being only one nucleotide different from a coyote allele. Interesting patterns, showing higher than expected heterozygosity and deviations from neutrality, thus suggesting trace of balancing selection, were also observed (Hedrick et al. 2002).

In the endangered Ethiopian wolf (*Canis simensis*), whose number is limited to less than 500 individuals, Kennedy et al. (2011) recently showed that a given haplotype (out of the seven ones found) was significantly associated with a lower post-vaccination immune response, during a severe rabies outbreak that affected one of the two existing populations. Therefore, this clearly indicates how even a limited level of variation at the MHC could be important for the survival of a species.

Although being the closest relative to the species of the genus *Canis* after the dhole (*Cuon alpinus*), the African wild dog (*Lycaon pictus*) does not seem to share with them any of the alleles known to date (Mardsen et al. 2009). However, the variability at the DQA1 and

DQB1 *loci* appears to be strongly reduced, with one and two alleles, respectively. This could suggest past population bottlenecks and declines resulted in loss of genetic variation, and potentially expose the species to higher risks of extinction given the actual population size (*ca.* 6000 individuals) and fragmentation.

A strong example of genetic impoverishment is given by the Island fox (*Urocyon littoralis dickeyi*), particularly the San Nicolas population, in which previous studies found a complete lack of variation at commonly variable markers. On the contrary, Aguilar et al. (2004) found some level of variation at MHC *loci* (DRB1, DQB1, and three MHC-linked microsatellites), which requires strong coefficients of balancing selection on the MHC to recover variability after a recent bottleneck event that monomorphized the examined neutral *loci*.

However, significant additional information still needs to be gained, especially aiming at obtaining a better view over the variability in the remnant threatened populations and the effects of the MHC on fitness and mating system in the wild, as we seek in the present study.

2.2.Aims

As we saw, the Italian wolf population has been threatened in the recent years by direct human persecution and reduction of the natural prey. This combination of factors strongly reduced its population size down to less than a hundred individuals in the '70s, confined to central Apennines and without any possibility of external gene flow. However, a much longer isolation from the other European populations could have occurred (Lucchini et al. 2004) for the past thousands of years, resulting in genetic reduction and differentiation, with a single mitochondrial (mtDNA) haplotype carried by the whole population (Randi et al. 2000). In other endangered species, such as cheetah (*Acynomix jubatus*, O'Brien et al. 1985), Tasmanian devil (*Sarcophilus harrisii*, Siddle et al. 2007), Florida panther (*Puma concolor coryi*, Roelke et al. 1993), and to a lesser extent in panda (*Ailuropoda melanoleuca*, Wan et al. 2006), MHC variation revealed to be strongly reduced (Radwan et al. 2010), therefore summing up to other conservation concerns (isolation, past bottleneck and general loss of genetic diversity or inbreeding, etc.).

Therefore, our main aim is to describe for the first time the MHC variation in the Italian wolf population, trying to verify which effect, if any, has been caused by the long bottleneck and isolation. Our hypothesis is that, compared to other wolf populations, the levels of polymorphism can be reduced, but that a certain level of variability could be maintained by selection. To do that, we will study the heterozygosity at three MHC *loci* and we will compare it to supposedly neutral microsatellite markers.

The second main concern for the preservation of the genetic health of the Italian wolf population is given by the possible hybridization or gene introgression with domestic dog. Multiple studies based on the analysis of mitochondrial and nuclear markers (Randi and Lucchini 2002, Randi et al. 2000, Randi 2008, Verardi et al. 2006) revealed several events of detectable hybridization, although representing a limited number in the population (*ca.* 5%). In Italy, feral or free-roaming dogs outnumber wolves by three orders of magnitude (about one million *vs.* one thousand wolves), although the cases of hybridization were only detected at the boundaries of wolves' distribution, and widespread gene introgression seems to be unlikely. However, recent studies (Caniglia et al. submitted) also taking into account functional markers responsible for the black coat coloration in wolves and dogs (β -defensin, Anderson et al. 2009, Candille et al. 2007), verified that their presence in the population is much more common, suggesting past hybridization and gene introgression events - if it will be confirmed that the mutation originated in dogs.

Therefore, in order to explore how these events could have influenced a highly functional and adaptive gene cluster such as MHC, we dedicated the second part of this study to the analysis of the presence of dog-derived MHC alleles in the wolf population. To do that, we compared all the alleles described to-date in the *Canis* genus with the ones found in a sample of admixed individuals and in a group of black wolves carrying the β -defensin mutation. According to several models, the allele frequency of functional genes under balancing selection are expected to be less divergent than the ones at neutral *loci*, but this trend can be reversed in case of local adaptation (van Oosterhout 2009). Therefore, assuming different selective pressures on wolf and dog populations, MHC genes could be potential candidate markers to better discriminate the origins of admixed or introgressed individuals.

In addition, wolf societies are among the most highly organized ones in the animal world (Bekoff and Pierce 2009), with comparably levels of social complexity found in a few species of primates, cetaceans, plus elephants and hyenas. The mating system (Geffen et al. 1996) is strictly monogamous and based on hierarchical levels, with a single mating pair usually reproducing once a year in every pack.

Therefore, it is interesting to investigate for the first time whether the MHC has any role in the mating choice within wolf packs, and if it can affect the fitness of individuals.

To test that, we will use data collected throughout the last ten years of non-invasive genetics studies in a subset of the Italian population located in the northern Apennines, whose pedigrees has been carefully reconstructed based on a number of microsatellite markers, and their main fitness traits have been deduced, resulting in an almost unique dataset in the world, with the only exception of Yellowstone wolves and few other cases.

Our hypothesis is that high MHC polymorphism (compared to the background levels) can be maintained by disassortative mating choice, thus maximizing the offspring's heterozygosity. In addition, this could be reflected on the fitness of the individuals, with higher values expected for the most heterozygote wolves.

2.3.Methods

Sampling and laboratory procedures

The samples were chosen among the ones available in the large database of wolf and dog genotypes that is being implemented at the Laboratory of genetics at ISPRA, the Italian Institute for Environmental Protection and Research (formerly National Wildlife Institute, INFS), located in Ozzano Emilia, Bologna; the database is developed in compliance with European Community and national laws that require that wolf populations are actively monitored (Boitani 2000, Genovesi 2002). The samples were obtained by professional operators (veterinaries, Forestry Corps agents) from autopsies of wolves died for natural reasons, car accidents or illegally killed (Caniglia et al. 2010) throughout the population range (P. Ciucci et al. 2007, Lovari et al. 2007), or from biopsies of live-trapped wolves (e.g. in Ciucci et al. 2009), and sent to ISPRA in the last 15 years; for most of the wolves, phenotypic information was recorded, such as estimated age and health conditions at the time of death, as well as its sex and morphological abnormalities, e.g. dewclaw or darker-than-usual coat color. However, given the non-systematic fashion in which these data were recorded, information concerning health status and causes of death has not been taken into account as indicative of pathologies nor as estimates of individuals' fitness during life.

In addition, non-invasive samples (feces, urine and blood traces) were also included in the database, coming from monitoring projects performed in the Apennine from 2000 to 2009 (Caniglia et al. 2010b, Galaverni et al. 2012) and in the western Alps from 1999 to 2004 (Fabbri et al. 2007).

Muscular tissues were stored at -20°C in 10 volumes of 95% ethanol, whereas scat samples in 95% ethanol were frozen for at least 10 days at -80°C in order to kill parasites and *Echinococcus* eggs, and then stored at -20°C until DNA extraction.

DNA was extracted using the Qiagen DNeasy Blood and Tissue Kits (QIAGEN) with a robotic liquid handling system MultiPROBE IIEX (PerkinElmer). Fecal samples were processed in a room dedicated to non-invasive genetics, always adding blank controls (no DNA in PCR) to check for possible contamination and according to a *multitube* protocol as in Caniglia et al. (2010).

All the samples were amplified and sequenced at 350 bp of the mtDNA control-region, which contains diagnostic mutations for the identification of the Italian wolf haplotype W14 (Randi et al. 2000). Subsequently, they were genotyped at 12 canine microsatellite *loci* that were selected for their high polymorphism in the Italian wolf population: FH2004, FH2079,

FH2088, FH2096 and FH2137 (Francisco et al. 1996), CPH2, CPH4, CPH5, CPH8 and CPH12 (Fredholm and Winterø 1995), C09.250 and C09.253 (Ostrander et al. 1993), as in Randi and Lucchini (2002), at the optimal PCR conditions for each primer. This panel of microsatellites allows determining the individual genotypes with a probability of identity $PID = 7.1 \cdot 10^{-9}$, and an expected PID among full sib dyads $PID_{sibs} = 3.1 \cdot 10^{-4}$ in the Italian wolf population (Fabbri et al. 2007, Lucchini et al. 2002). Whenever unknown, the sex of the genotypes was determined by PCR-RFLP of diagnostic ZFX/ZFY sequences (Garcia-Muro et al. 1997, Fabbri et al. 2007). Male individuals were also amplified at three Y-linked microsatellites: MS34A, MS34B, MS41B (Iacolina et al. 2010, Sundqvist et al. 2001). Furthermore, every genotype was also tested for the presence of a 3-bp deletion at the *K-locus* (Caniglia et al. submitted) indicative of the mutation at the β -defensin 103 (CBD103) gene that induces the black coat color in wolves.

PCR products were analyzed in an automated sequencer ABI 3130XL (Foster City, CA), using the software SEQUENCING ANALYSIS v.3.7 and SEQSCAPE v.2.5 for sequences, and GENESCAN v.3.7 and GENMAPPER v.4.0 for microsatellites.

Aiming at describing the variability of the MHC and the presence of dog alleles in the Italian wolf population, genotypes from 92 unrelated wolves of both sexes were randomly chosen, after including all the possible individuals for which atypical phenotypic features had been recorded, or that have been considered of admixed origin according to previous studies (Ciucci et al. 2003, Randi and Lucchini 2002, Verardi et al. 2006). Therefore, the proportions of individuals with a possible admixed origin are not proportional to the real frequency in the population, but are likely being increased. According to their location, the samples from the selected subset were assigned to four groups: Alps (Al), northern (nAp), central (cAp) and southern Apennine (sAp), whose territories include all the current Italian wolf distribution.

The software STRUCTURE v. 2.2 (Falush et al. 2003) was used to assign individuals to baseline wolf or dog populations, independent of any prior non-genetic information, on the basis of the 12 genotypes microsatellite *loci*. As a reference, we included the genotypes determined in other 154 randomly selected tissue samples from wolves in the database having the typical Italian wolf coat color pattern and not showing any detectable phenotypic and genetic signal of hybridization. A reference dog population was composed by the genotypes determined from 116 blood samples collected from dogs living in rural areas in Italy.

We ran Structure with a burn-in period of 10^4 iterations followed by five repetitions of 10^5 iterations, selecting the 'admixture model' (each individual may have ancestry in more than one parental population) and the 'I model' (independent allele frequencies), with the

population flag option activated (updating the allele frequencies with the POP flag). The optimal number of populations, or better the value that maximized the posterior probability of the data, was set at $K = 2$ according to previous studies (Randi and Lucchini 2002, Verardi et al. 2006). We then assessed the average proportion of membership (Q_i) of the sampled populations to the inferred clusters. Individuals showing a proportion of membership higher than the minimum value observed in the reference wolf population (also considering the values of the 90% interval of confidence, C.I. 90%) were assigned to the wolf cluster as pure wild-type wolves (Wt); individuals showing lower values were considered as admixed (H), as well as the ones showing mtDNA haplotypes different from W14 (Randi et al. 2000), or Y-chromosome microsatellite *multilocus* haplotypes different from the ones described in the Italian wolf population (Caniglia et al. submitted, Iacolina et al. 2010) Individuals that have been genetically assigned to the wolf cluster, but showed atypical phenotypic features were assigned to a third group (Ph).

For all these samples, the second exon of three DLA class II genes, DRB1, DQA1 and DQB1, was analyzed. They were amplified with primers used in Hedrick et al. (2002), after Kennedy et al. (1998) for DRB1, and in Kennedy et al. (2006), after Wagner et al. (1996) for DQA1 and DQB1 (Tab. 2). All the primers are intronic and *locus*-specific, yielding a product of 280 bp for DRB1, 345 bp for DQA1 and 300 bp for DQB1.

Locus	Primer name	Primer sequence (5'-3')	Annealing temperature and time	No. of cycles
DRB1	DRB1F	CCGTCCCCACAGCACATTTTC	62-52°C * 60'' (TouchDown)	20 TD+20
	DRB1R	TGTGTCACACCTCAGCACCA		
DQA1	DQAin1	TAAGGTTCTTTTCTCCCTCT	57°C * 30''	30
	DQAin2	GGACAGATTCAGTGAAGAGA		
DQB1	DQB1B	CTCACTGGCCCGGCTGTCTC	66°C * 45''	30
	DQBR2	CACCTCGCCGCTGCAACGTG		

Table 2: Primer sequences and amplification conditions for the studied MHC class II *loci*.

Amplification reactions were carried out in a 10 μ l mix, including 2 μ l genomic DNA, 1 μ l BSA 2% and 0.2 μ l of each 10 μ M primer plus 0.25 units Taq. After the initial denaturation at 94°C, each cycle was performed with a step at 94°C for 30'', an annealing step with conditions specific for each primer (Tab. 2) followed by an extension for 45'' at 72°C. A final extension at 72°C for 10min was performed once the optimal number of cycles was completed. PCR products were purified with Exo/SAP-IT, then the sequencing reactions were performed in both directions using BigDye Terminator v1.1, according to the manufacturer's protocol. PCR products were analyzed in an automated sequencer ABI 3130XL with the software SEQSCAPE

v.2.5, using as references the sequences DLA-DRB1*03101 (AF336108.1), DLA-DQA*014012 (AJ316220.1) and DLA-DQB1*05601 (FM246843.1).

Genetic variability

The allele identification was performed after a computational reconstruction with Phase (Stephens et al. 2001; Stephens and Donnelly 2003), in DnaSP v5.10 (Librado and Rozas 2009), using the 'recombination' model (-MR0) and 1000 iterations after 100 burn-ins. Compared to similar software, Phase is able to cope with tri-allelic states (represented as numeric STR markers) that are commonly found in MHC sequences. When the probability of reconstruction of the alleles was lower than 0.9, with multiple combinations of alleles being possible, the sample was discarded.

The alleles were then matched via BLASTn at NCBI (Johnson et al. 2008) to the ones available in GenBank for all the species of the genus *Canis*, which were downloaded and aligned in Geneious v.5 (Drummond et al. 2011). In addition, we also included all the sequences available on the Immuno Polymorphism-MHC Database (IPD; Robinson et al. 2010) on the EBI-EMBL website (<http://www.ebi.ac.uk/ipd/mhc/dla/index.html>) that were not found in GenBank. Sequences that were matching along all the analyzed regions, but have been assigned multiple names, were grouped and assigned a single name respecting the rules defined in the official ISAG reports (Ellis et al. 2006, Kennedy et al. 2001).

The alleles in our samples were accepted if they matched previously described alleles. Otherwise, they were considered as new alleles and submitted to GenBank only if they were observed in homozygous state in at least two different samples. Otherwise, when a new allele was observed in a single sample or only in heterozygous state, if it was a single nucleotide different from already described alleles it was considered a possible sequencing error and named after it, otherwise discarded.

Given the high linkage between the *loci*, *multilocus* haplotypes were also reconstructed, following the subtractive approach method described in Kennedy et al. (2007). The haplotype reconstruction was then confirmed computationally in PHASE (Berggren and Seddon 2008) by concatenating the gene sequences prior to the phasing step, and applying the recombination model with two hot-spots (-MR2) corresponding to the boundaries between adjacent genes (DRB1/DQA1/DQB1).

For both microsatellites and MHC genes, the number of alleles, the allele frequencies (AF) by population and by *locus*, the observed (H_o) and expected (H_e) heterozygosity, F statistics and departures from Hardy-Weinberg equilibrium (HWE) were assessed in GENALEX 6.4 (Peakall

and Smouse 2006). Considering every variable site as a single marker, we also computed the same F statistics and departures from HWE SNP by SNP, in order to identify which specific site was responsible for the larger effects on these metrics.

The AF at each *locus* were compared between groups by computing the R^2 values from their regression plots in Excel, as well as by a χ^2 test in which we compared the AF for each group to the ones of the whole population. A pairwise Kolmogorov-Smirnov test was also computed and represented as a cumulative fraction plot after Kirkman (1996; in <http://www.physics.csbsju.edu/stats/>, accessed on December 21st, 2011). The AF by geographic groups were also computed and represented for wild-type wolves, in addition to an allele discovery rarefaction curve.

Average observed heterozygosity levels at both STR *loci* and MHC genes were compared with the ones expected by a Ewen-Watterson statistics of heterozygosity implemented in BOTTLENECK 1.2.02 (Cornuet and Luikart 1996), under the assumptions of: 1) an infinite allele mutation model (IAM); 2) a two-phase mutation model (TPM) with 90% single-step mutations. The test computes the difference (DH) between the observed and expected heterozygosity values, and divides it by the SD of gene diversity, retrieving the corresponding p values after simulating 1,000 iterations per *locus*. In populations where a recent bottleneck occurred, as it is the case (Fabbri et al. 2007, Lucchini et al. 2004), both the allele numbers (k) and gene diversity (He, or Hardy-Weinberg heterozygosity) at polymorphic *loci* are reduced, but at a faster pace for the allele number, leading to an observed gene diversity higher than the expected equilibrium gene diversity (Heq) computed from the observed number of alleles under the assumption of a constant-sized population (Cornuet and Luikart 1996).

DNASP v.5 was used to compute for each MHC gene the number of segregating sites, the haplotype diversity (Hd), and the nucleotide diversity (Pi), both on average and in sliding windows of 25 bp with step size of 5 bp. We also computed the average pairwise ratio (d_N/d_S) of the number of non-synonymous mutations per non-synonymous site (d_N) to the number of synonymous mutations per synonymous site (d_S), as well as the Tajima's D test and the Fu and Li's test along sliding windows.

MEGA v.4 (Tamura et al. 2007) was used to reconstruct the phylogenetic relationships of all the available sequences for each gene, using a Neighbor-Joining method with 5000 bootstrap replicates based on the Kimura 2-parameter substitution model. As outgroup, one corresponding sequence from *Macaca fascicularis* and one from *Macaca mulatta* were randomly chosen and included. Whenever present, gaps were excluding from pairwise comparisons. Only bootstrap values above 60 were represented on the trees. Based on the

extant bibliography, the species in which every allele was found were indicated aside each branch.

However, given the low resolution of phylogenetic trees applied to MHC genes, in order to better resolve the topology and distances of the alleles we also reconstructed *single-locus* haplotype networks in NETWORK v.4.6.1 (using values of $\epsilon = 10$) after creating the input Roehl Data Files (.rdf) in DnaSP. Specific networks were also constructed only including the alleles found in the Italian population, whose frequencies were split into the three groups of assignment (Wt, Ph, H) and represented with proportional sizes of the nodes.

Pack reconstruction

For the analyses on the influence of the MHC variability on the mating choice and on the fitness in the wild, we started based on the work from Caniglia et al. (2010; submitted a;b) and Galaverni et al. (2012). These works led to identify, through a non-invasive genetic approach, a number of stable wolf packs inhabiting the Northern Apennine. Individual genotypes from the ISPRA wolf database, all sampled from 2000 to 2011, were obtained using the same markers and methods previously described. According to its sampling locations, each genotype was assigned an individual Minimum Convex Polygon (iMCP). Individuals that have been sampled more than four times and for at least two years (Frequently Sampled Individuals, FSI), were considered as potential candidates for being reproductive members of a pack (Caniglia et al. submitted, Galaverni et al. 2012), excluding cases where subsequent sampling locations were exceeding 20 km of linear distance, since they could represent cases of dispersal. Whenever an iMPC was overlapping one or more FSI iMPCs, they were merged in a multiple MPCs (mMPCs). All the individuals that were sampled (even only once) within a given mMPC or within a surrounding area of 15 km were considered as potential members of the same pack. The most likely familial groups were reconstructed through a maximum-likelihood approach implemented in COLONY v. 2.0 (Wang and Santure 2009), considering all the individuals as candidate parents. COLONY was run with allele frequencies and PCR error rates as estimated from the whole reference population, considering a probability of including fathers and mothers in the candidate parental pair of 0.5. The best genealogies reconstructed by COLONY were then verified in PARENTE v. 1.2 (Cercueil et al. 2002), and only highly matching parent–offspring combinations were retained (only 1/24 allele disparity was allowed, corresponding to a match >95%). All the genealogies were compared to the patterns of temporal and spatial sampling, plus, whenever available, to

other field information, such as snow-tracking, wolf-howling and camera-trapping (Caniglia et al. 2010; Galaverni et al. 2012).

The Queller & Goodnight's relatedness (r) between reproductive individuals, and between pack members and non-members, was evaluated.

Following these findings, we selected 66 genotypes likely to belong to reproductive individuals from 34 different pack pairs (with multiple breeding pairs being possible through time in the same pack).

We also included 10 random individuals shown to be the offspring of different pairs, plus 18 individuals belonging to a given pack, but unrelated to the reproductive individuals. This allowed us to check for the correct assignment of the alleles and haplotypes in the breeding pairs by trio comparisons with their offspring, and to evaluate the mean heterozygosity values in breeding vs. non-breeding individuals.

MHC and mating choice

The possible influence of the MHC variability on mating choice was tested in several ways. First, given the allele frequency in the Northern Apennine population, calculated from both invasive and non-invasive samples, we calculated the probability for each breeder of mating (mating probability, Mp) with each one of the following classes of individuals, based on HWE expected frequencies: with a mate sharing both *multilocus* haplotypes ($Mp = p*q$), with a mate with a single haplotype in common ($Mp = [p*(1-q)+q*(1-p)]$), or with both haplotypes being different ($Mp = [(1-p)*(1-q)]$), where p and q are the frequencies of the two haplotypes in the first genotype. We then evaluated the difference between the observed and expected number of mating events in the three classes (χ^2 Test). In order to detect the effects of a potential sex-biased choice, we replicated the tests by considering male ($Mp\♂$) and female ($Mp\♀$) breeders independently. Without a prior knowledge of the effect of any particular MHC gene, the same test has also been replicated independently for the single *loci* rather than for the whole haplotypes.

However, since the conditions required to meet the HWE expectations could be not met in our sample, especially given the low population size and the potential gene flow, we tried to empirically assess the probability of non-random mating by applying a permutation procedure based on the values of Queller and Goodnight's relatedness (r) at the three MHC *loci* between the members of the real and the potential pair combinations, as implemented in PERM 1.0 (Duchesne et al. 2006). The software was run for 5000 permutations repeated for 10 iterations, also considering the individuals' sex in defining the potential mates' groups, on

the hypothesis that the relatedness between mates is lower than expected by random, according with a disassortative mating scheme. To better evaluate the effects of similarity between mates at the peptide level, the same procedure was applied considering the average number of amino acid (AA) differences between mates, both for single genes and in total.

However, at least two confounding effects can limit the power of detection of any trace of assortative or disassortative mating.

The first one is given by the fact that, despite the high mobility of wolves (Valière et al. 2003), especially males, the actual gene pool of potential mates is more likely to be limited to the adjacent packs' members of the opposite sex (in particular for females) or to the pack members themselves. Therefore, we compared the average levels of allele sharing (number of alleles in common) and protein divergence (average number of AA differences) between the members of actual pairs with the ones between reproductive wolves and ten unrelated potential mates of compatible sex from the same packs. In addition, by using PERM in order to overcome the different sample sizes, the mean H_o of breeders has been compared to that of unrelated non-breeding individuals, supposing that reproductive wolves show a higher heterozygosity level than non-reproductive ones.

The other confounding effect could rise from considering packs of new foundation, particularly in areas of recent wolf expansion or colonization. In this cases, given the low number of individuals present in the area, an actual choice of the mate is not possible or less likely to occur. Therefore, we repeated the analyses on mating probabilities (both at the gene and haplotype levels) without including the breeders from packs whose foundation was documented to be occurred in the same year of the reproduction or in the previous one, either by non-invasive sampling or field observations.

In addition, we also tested the hypothesis of a mating-up model, where the individuals with a lower heterozygosity tend to compensate this potential handicap for their offspring by choosing mates with higher heterozygosity values.

We therefore calculated: i) the proportions of homozygote females mating with heterozygote males compared to the proportion of the heterozygote ones; ii) the number of alleles in common between mates with respect to their heterozygosity levels at the MHC; iii) the heterozygosity values of male versus female mates and *vice versa*; iv) the level of protein divergence, computed as the number of amino acid (AA) differences, between mates compared to their heterozygosity levels. The latter point was also used in order to test for the divergent-allele hypothesis that is, supposing that individuals tend to choose mates with the

most different allele sequence, and that this will enhance the progeny's fitness. In each of the four cases, we expected an inverse correlation in the case of a mating-up scheme.

As a comparison, the number of AA differences and the number of shared alleles were also compared to the background relatedness, based on the 12 microsatellite *loci*.

Nonetheless, a direct effect of the MHC variability on the mating choice in the wild, given the limited number of packs that was feasible to analyze, given the elusiveness of the species, is not easy to detect – or maybe just does not occur.

On the other hand, the effects of MHC heterozygosity on the fitness of the individuals in the wild are an interesting and useful measure of their level of adaptation to the environment, and can reflect the strength of current selective pressures on the population.

MHC and its effects on fitness traits

Although a meaningful fitness estimate in relation to MHC variation could be given by the levels of parasites affecting a host, or the rate of pathogen-driven mortality, exhaustive and methodologically coherent data in the Italian wolf population are not currently available, although the presence of parasitic infections such as mange (*Sarcoptes scabiei*) has been repeatedly documented (Apollonio et al. 2004, Galaverni et al. 2012, Lovari et al. 2007)

Therefore, in the present study we included what we considered being other good estimates of fitness in wolves: the total number of offspring of an individual (hereafter: 'total offspring', TO), the time it has been sampled ('sampling time', ST), the years as documented breeder ('reproductive years', RY), and the average litter size per year ('litter size', LS).

TO is a good measure of the reproductive success of an individual, and probably the most informative fitness trait at all. ST can reflect the wolf survival, RY the duration as top-ranked, breeding individual, LS can be an indirect measure of the fecundity of a given pair.

Of course, these parameters are only deduced as indirect estimates from the non-invasive genetic sampling, and can be biased by a number of factors: the sampling time and intensity, which were not homogeneous throughout the study area, but also environmental and temporal variations, which can influence the food availability, the energy consumption, etc. Nonetheless, there are three reasons why we think they are worth using: 1) to our knowledge, they are the best estimates so far available for a representative wolf population in Italy and, with the exception of Yellowstone wolves, among the best ones worldwide; 2) these measures, although potentially biased, should still reflect the real ones, being proportional to them or at least good underestimates; 3) every sampling bias should be irrespective and independent from the MHC variation, which is actually the main variable we are going to

relate to the fitness traits; therefore any source of error should be randomly distributed in relation to the MHC.

In order to take into account the possible influence of sampling heterogeneity, however, we divided the packs into three geographical and altitudinal groups, respectively: Eastern (E), Central (C) and Western (W) Northern Apennine; High (above 800m a.s.l., H), Intermediate (400 to 800m a.s.l., I), and Low (below 400m a.s.l., L) altitude, reflecting potential and described ecological partitions (e.g. distribution of beech vs. oak forests). We also took into account the year in which the first reproduction of each breeder occurred.

However, we did not consider in these analyses the breeding pair coming from Maremma Regional Park, since living in a widely different environment (Mediterranean coastal forest) and being of putatively admixed wolf*dog origin (Caniglia et al. submitted).

Fitness estimates have been compared to all the variables deduced from genetic data on breeding wolves: the H_o (MHC), both haplotypic, at all *loci* and at each *locus*; the background H_o (STR); the difference in relatedness at MHC and STRs; the average number of AA differences between alleles for each individual (for every MHC *locus*, the total number at three *loci*, and the total number in β -chains, namely DRB1 plus DQB1); the relatedness (r) between mates, both r (MHC) and r (STR); the average number of AA differences between mates (also for every MHC *locus*, the total at three *loci*, and the total in β -chains).

Then, the breeding wolves have been ranked according to each fitness trait, and the ones from the first quartile ($n=11$) have been compared to the ones in the last quartile, then the differences in the means of their genetic parameters have been compared (both with a t test for independent samples, C.I. 95%, and a comparison of means, with 10 iterations of 1000 permutations, in PERM).

In order to take into account the cumulative or interactive effects of the variables, for each fitness trait the best Linear Model has also been reconstructed in R v.2.9.2 (R Development Core Team 2009; <http://www.R-project.org>) with the user interface implemented in R Commander (Fox 2005).

2.4. Results

MHC variability in the Italian wolf population

We obtained good quality sequences from which we reconstructed reliable allele combinations at all three MHC *loci* in 74 out of 94 samples (79%).

From STRUCTURE analysis, all the reference wolf individuals showed a proportion of membership to the wolf cluster $Q_i > 0.95$, with the inferior limit of the 90% C.I. higher than 0.79. Therefore, since 26 of the tested individuals (35%) showed Q_i values lower than those, they were considered as admixed (and labeled as ‘H’). All the remaining individuals were considered as putative genetically-genuine wolves, and split into wild-type (‘Wt’, $n=38$) and phenotypically-unusual individuals (‘Ph’, $n=10$), according to their documented appearance.

All the *loci* (100%) turned out to be polymorphic. DRB1 showed 9 different alleles, DQA1 had 6 alleles, DQB1 8 alleles (Tab. 3). The number of segregating sites ranged from 43 in DRB1 to 39 in DQB1 but only 8 in DQA1.

All the alleles matched previously described sequences, except for two DRB1 alleles, which showed a single difference to known alleles found in European (Calu-DRB1*13 allele, with a G to A mutation at site 255) or North American wolves (DRB1*09201 allele, with a C to A mutation at nucleotide 60). The latter has also been sampled in our population, in individuals from Southern or Central Apennine, whereas Calu-DRB1*13 allele was not found in our samples. However, the two new alleles were the most frequent in the Italian population, and found in homozygote state in more than four individuals each, therefore meeting the criteria established by the ISAG nomenclature committee (Ellis et al. 2006, Kennedy et al. 2001). Their sequences are going to be submitted to GenBank and assigned official names. Meanwhile, they will be described in the present study as Calu-DRB1*13-newItaly and DRB1*09201-newItaly.

DRB1	Nomenclature	n	Freq	Described in:	GenBank name	AN
1	Calu-DRB1*13-newItaly	56	0.38	Never	-	-
2	DLA-DRB1*09201-newItaly	40	0.27	Never	-	-
3	DLA-DRB1*03601	21	0.14	We, Wa	03601	AF336110.1
4	DLA-DRB1*02001	10	0.07	D	D20	U58684.1
5	DLA-DRB1*03901	6	0.04	Wa ,Rw	03901	AF343740.1
6	DLA-DRB1*01501	5	0.03	D	DRB1-W	DQ056281.1
7	DLA-DRB1*03701	5	0.03	Wa	03701	AF343738.1
8	DLA-DRB1*00101	3	0.02	D	DRB1-U; DRB1-Q	DQ056278.1; DQ056274.1
9	DLA-DRB1*09201	2	0.01	Wa	09201	AM408904.1

DQA1	Nomenclature	n	Freq	Described in:	GenBank name	AN
1	DLA-DQA1*005011	101	0.68	We, Wa, D, Wm	DQA3	U44787.1
2	DLA-DQA1*01201	21	0.14	We, Wa, D, C	01201	AF343734.1
3	DLA-DQA1*00401	11	0.07	Wa, D	DQA4	U44788.1
4	DLA-DQA1*00201	6	0.04	We, Wa, D	DQA9	U75455.1
5	DLA-DQA1*00601	6	0.04	We, Wa, D	DQA6	U44790.1
6	DLA-DQA1*00101	3	0.02	We, Wa, D, C, Wm	DQA2	U44786.1

DQB1	Nomenclature	n	Freq	Described in:	GenBank name	AN
1	DLA-DQB1*03901	56	0.37	We	03901	AY126651.1
2	DLA-DQB1*00701	45	0.31	D, Wm, We, Wa	DQB4	AF043149.1
3	DLA-DQB1*03501	22	0.15	We, Wa, D	03501	AJ311107.1
4	DLA-DQB1*01303	9	0.06	D, Wm, We, Wa	DQB7	AF043152.1
5	DLA-DQB1*02901	6	0.04	We	02901	AY126648.1
6	DLA-DQB1*00301	5	0.03	D	DQB6	AF043151.1
7	DLA-DQB1*00201	3	0.02	D	DQB3	AF043148.1
8	DLA-DQB1*02002	2	0.01	Wa, D	DQB19	AF043164.1

Table 3: Official names and frequencies of the alleles found at each *locus* in the Italian wolf population, with corresponding GenBank names and accession numbers (AN), and the canid populations where they were described to date (We=European wolf; Wa=North American wolf; Wm=Mexican wolf; Rw=red wolf; D=dog; C=coyote). ‘n’ indicates the number of chromosomes carrying a given allele.

However, the allele frequencies changed across groups, as shown in Tab. 4., with some private alleles being present in the admixed individuals’ group H. Except for allele DLA-DQA1*00101, described in a number of wolf populations, dogs and coyotes, all the other ones have been only detected in dogs, therefore are compatible with being alleles of dog origin retained in admixed individuals.

Locus \ Group	Wt (2n=76)		Ph (2n=20)		H (2n=52)		Described in
DRB1 allele	n	freq	n	freq	n	freq	
Calu-DRB1*13-newitaly	28	0.37	12	0.60	16	0.31	Never
09201-newitaly	28	0.37	6	0.30	6	0.12	Never
DLA-DRB1*03601	11	0.14	1	0.05	9	0.17	We, Wa
DLA-DRB1*02001	3	0.04	0	0.00	7	0.13	D
DLA-DRB1*03901	3	0.04	1	0.05	2	0.04	Wa ,Rw
DLA-DRB1*01501	0	0.00	0	0.00	5	0.10	D
DLA-DRB1*03701	2	0.03	0	0.00	3	0.06	Wa
DLA-DRB1*00101	0	0.00	0	0.00	3	0.06	D
DLA-DRB1*09201	1	0.01	0	0.00	1	0.02	Wa
DQA1 allele	n	freq	n	freq	n	freq	
DLA-DQA1*005011	58	0.76	18	0.90	25	0.48	We, Wa, D, Wm
DLA-DQA1*01201	11	0.14	1	0.05	9	0.17	We, Wa, D, C
DLA-DQA1*00401	3	0.04	0	0.00	8	0.15	Wa, D
DLA-DQA1*00201	3	0.04	1	0.05	2	0.04	We, Wa, D
DLA-DQA1*00601	1	0.01	0	0.00	5	0.10	We, Wa, D
DLA-DQA1*00101	0	0.00	0	0.00	3	0.06	We, Wa, D, C, Wm

DQB1 allele	n	freq	n	freq	n	freq	
DLA-DQB1*03901	29	0.38	12	0.60	15	0.29	We
DLA-DQB1*00701	29	0.38	6	0.30	10	0.19	D, Wm, We, Wa
DLA-DQB1*03501	12	0.16	1	0.05	9	0.17	We, Wa, D
DLA-DQB1*01303	2	0.03	0	0.00	7	0.13	D, Wm, We, Wa
DLA-DQB1*02901	3	0.04	1	0.05	2	0.04	We
DLA-DQB1*00301	0	0.00	0	0.00	5	0.10	D
DLA-DQB1*00201	0	0.00	0	0.00	3	0.06	D
DLA-DQB1*02002	1	0.01	0	0.00	1	0.02	Wa, D

Table 4: Allele frequencies by group at each *locus* ('Wt'=wild-type wolves, 'Ph'=phenotypically-unusual wolves; 'H'=admixed wolf*dog individuals). Alleles that are private to one group are highlighted in bold in the corresponding column.

The effective number of alleles (N_e) is maximum at DRB1. Both the observed (H_o) and expected heterozygosity (H_e) are higher at DQB1 and minimum at DQA1, with the first parameter being slightly lower than the second (Tab. 5, upper).

Locus	N	Na	Ne	I	Ho	He	UHe	F
DRB1	74	9	4.074	1.677	0.689	0.755	0.760	0.087
DQA1	74	6	2.020	1.070	0.486	0.505	0.508	0.037
DQB1	74	8	3.777	1.565	0.716	0.735	0.740	0.026
Mean		7.667	3.290	1.437	0.631	0.665	0.669	0.050
SE		0.882	0.641	0.186	0.072	0.080	0.081	0.019

Locus	Group	N	Na	Ne	I	Ho	He	UHe	F
DRB1	Wt	38	7	3.374	1.423	0.579	0.704	0.713	0.177
	Ph	10	4	2.198	0.967	0.700	0.545	0.574	-0.284
	H	26	9	5.753	1.941	0.846	0.826	0.842	-0.024
DQA1	Wt	38	5	1.648	0.798	0.447	0.393	0.399	-0.137
	Ph	10	3	1.227	0.394	0.200	0.185	0.195	-0.081
	H	26	6	3.347	1.459	0.654	0.701	0.715	0.068
DQB1	Wt	38	6	3.139	1.307	0.632	0.681	0.691	0.073
	Ph	10	4	2.198	0.967	0.700	0.545	0.574	-0.284
	H	26	8	5.474	1.840	0.846	0.817	0.833	-0.035

Table 5: no. Alleles (N_a), no. Effective Alleles ($N_e = 1/(\sum p_i^2)$), Shannon's Information Index ($I = -1 * \sum (p_i * \ln(p_i))$), Observed ($H_o = \text{No. of Hets} / N$), Expected ($H_e = 1 - \sum p_i^2$) and Unbiased Expected ($UHe = (2N / (2N-1)) * H_e$) Heterozygosity, and Fixation Index ($F = (H_e - H_o) / H_e$), Where p_i is the frequency of the i^{th} allele in the N individuals analyzed, in total (upper table) and across groups (lower table).

These values vary across groups, with the maximum heterozygosity found at every *locus* in the admixed individuals' group (Tab 5, lower) that, on average, also shows a higher number of alleles (Fig. 4).

However, the mean values across *loci* resulting from the F statistics are close to zero, with $F_{is} = -0.036 \pm 0.015$, $F_{it} = 0.027 \pm 0.025$, $F_{st} = 0.061 \pm 0.013$.

The genotype frequencies at all *loci* resulted not significantly different from what expected from HWE, except for *locus* DRB1 in the Wt group ($p < 0.001$), which showed a particular excess of the most common homozygote genotype (DRB1*09201-newItaly).

We further investigated this skew by splitting the allele sequences into their segregating sites, and treating them as separate markers. In this way, we identified the nucleotide mostly responsible for the departure from the equilibrium, namely the base in position 60 of the sequence ($p = 0.007$), corresponding to the mutation that discriminates the newly-discovered DRB1*09201-newItaly allele from its closest sequence, DRB1*09201. The only other nucleotide departing from the equilibrium in the same group, although less significantly ($p = 0.03$), was the one at position 97.

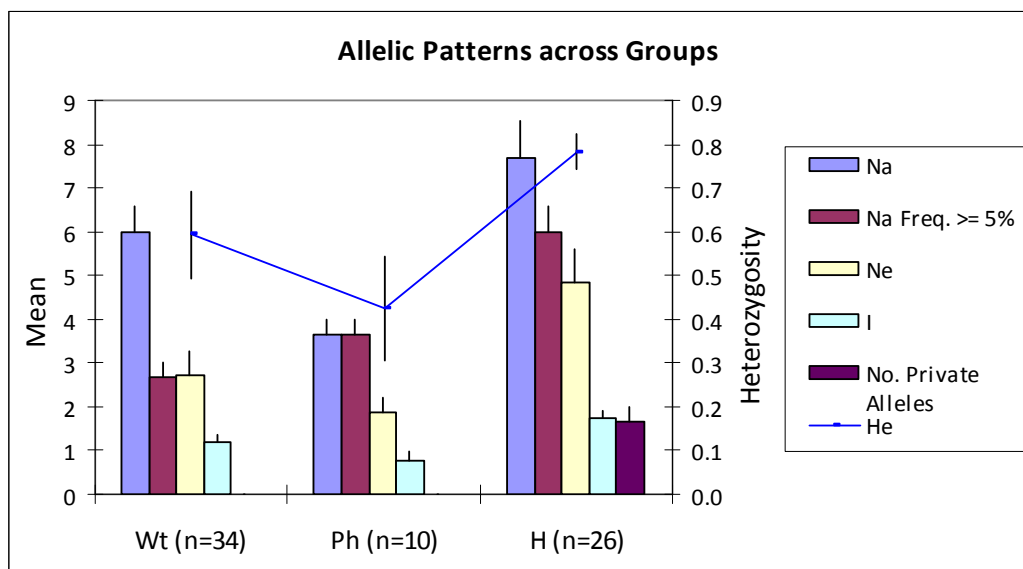


Figure 4: no. Alleles (Na), no. of Common alleles (Freq. $\geq 5\%$), no. of Effective Alleles (Ne), Shannon's Information Index (I), no. of Private Alleles and Expected Heterozygosity (He), averaged across *loci* for each group.

When compared to the microsatellite variation, the mean heterozygosity values at the three MHC *loci* turned out to be higher than the ones averaged over the 12 neutral markers, both considering observed and expected values, and number of alleles per *locus* (Tab. 6).

Parameter	STR		MHC	
	Mean	SE	Mean	SE
N	71.667	0.667	74.000	0.000
Na	5.750	0.827	7.667	0.882
Ne	3.099	0.386	3.290	0.641
I	1.216	0.125	1.437	0.186
Ho	0.562	0.049	0.631	0.072
He	0.613	0.055	0.665	0.080
UHe	0.617	0.055	0.669	0.081
F	0.071	0.024	0.050	0.019

Table 6: Mean no. Alleles (Na), no. Effective Alleles (Ne), Shannon's Information Index (I), Observed (Ho), Expected (He) and Unbiased Expected (UHe) Heterozygosity, and Fixation Index (F), averaged across the 12 microsatellite (STR) and the three MHC *loci* (SE=standard error).

This is particularly apparent when looking at the values across groups, where the admixed individuals show the highest excess of heterozygosity at the MHC *loci* compared to the analyzed microsatellites (Tab. 7).

	Group	Ho STR	Ho MHC	He STR	He MHC
	Wt	0.526	0.553	0.558	0.593
Mean	Ph	0.517	0.533	0.518	0.425
	H	0.636	0.782	0.670	0.782
	Wt	0.060	0.055	0.066	0.100
SE	Ph	0.061	0.167	0.052	0.120
	H	0.038	0.064	0.040	0.040

Table 7: Mean Observed (Ho) and Expected Heterozygosity (He) at the MHC and microsatellite (STR) *loci*, averaged in each group (SE=standard error).

The Ewen-Watterson statistics showed that the heterozygosity levels in the wild-type group at the STRs were higher than expected (Tab. 8), but their significance changed according to the model (Wilcoxon test, one tail for Ho excess $p=0.004$ under the IAM; $p=0.15$ under the TPM). Conversely, at the MHC we did not find any trace of significant excess (Wilcoxon test, one tail for Ho excess, $p=0.812$ under the IAM, and $p=1.000$ under the TPM). These contrasting results could suggest that if a reduction in the allele diversity during the population decline occurred, it did not influence with the same intensity the neutral and the functional *loci*, with the MHC being less severely affected.

locus	observed			under the I.A.M.				under the T.P.M.			
	n	ko	Ho	Heq	S.D.	DH/sd	Prob	Heq	S.D.	DH/sd	Prob
2004N	72	6	0.734	0.600	0.136	0.987	0.144	0.698	0.087	0.420	0.398
2079N	72	4	0.642	0.454	0.168	1.122	0.133	0.559	0.123	0.681	0.296
2088N	74	4	0.668	0.452	0.170	1.267	0.076	0.561	0.121	0.873	0.192
2096N	74	3	0.648	0.345	0.180	1.675	0.024	0.436	0.152	1.391	0.042
2137N	72	10	0.818	0.763	0.083	0.668	0.265	0.828	0.041	-0.244	0.343
cph2	74	4	0.405	0.450	0.168	-0.265	0.365	0.561	0.122	-1.275	0.119
cph4	74	3	0.309	0.338	0.177	-0.161	0.450	0.438	0.152	-0.850	0.217
cph5	74	3	0.660	0.341	0.176	1.819	0.005	0.442	0.150	1.449	0.027
cph8	74	5	0.783	0.529	0.158	1.612	0.003	0.650	0.098	1.361	0.015
cph12	74	3	0.413	0.341	0.177	0.406	0.405	0.436	0.152	-0.150	0.374
u250	70	4	0.676	0.456	0.165	1.335	0.056	0.570	0.120	0.888	0.178
u253	70	2	0.029	0.202	0.166	-1.039	0.220	0.240	0.170	-1.242	0.161
MHC	n	ko	Ho	Heq	S.D.	DH/sd	Prob	Heq	S.D.	DH/sd	Prob
DRB1	76	7	0.713	0.646	0.125	0.534	0.366	0.746	0.070	-0.472	0.255
DQA1	76	5	0.399	0.533	0.152	-0.888	0.198	0.641	0.103	-2.348	0.035
DQB1	76	6	0.691	0.599	0.140	0.656	0.310	0.702	0.080	-0.149	0.374

Table 8: Results from the Ewen-Watterson test, under an Infinite Allele Model (I.A.M) or a Two-Phase model (T.P.M), after 1,000 iterations (N = sample size; ko = observed number of alleles; He = observed heterozygosity; Heq = heterozygosity expected at equilibrium; DH = Ho/Heq).

The reconstruction of the most likely MHC *multilocus* haplotypes revealed the presence of 13 combinations of alleles, with the three most common haplotypes accounting for almost the 80% of the total frequencies (Tab. 9).

Haplotype	Nomenclature (DRB1 / DQA1 / DQB1)	n	freq.
1	Calu-DRB1*13-newItaly / DQA1*005011 / DQB1*03901	54	0.36
2	DRB1*09201-newItaly / DQA1*005011 / DQB1*00701	39	0.26
3	DRB1*03601 / DQA1*01201 / DQB1*03501	21	0.14
4	DRB1*02001 / DQA1*00401 / DQB1*01303	9	0.06
5	DRB1*03901 / DQA1*00201 / DQB1*02002	6	0.03
6	DRB1*03701 / DQA1*005011 / DQB1*00701	5	0.04
7	DRB1*01501 / DQA1*00601 / DQB1*00301	4	0.03
8	DRB1*00101 / DQA1*00101 / DQB1*00201	3	0.01
9	DRB1*09201 / DQA1*00601 / DQB1*02002	2	0.01
10	Calu-DRB1*13-newItaly / DQA1*005011 / DQB1*00701	2	0.01
11	DRB1*01501 / DQA1*00401 / DQB1*00301	1	0.02
12	DRB1*02001 / DQA1*00401 / DQB1*03901	1	0.01
13	DRB1*09201-newItaly / DQA1*005011 / DQB1*03501	1	0.01

Table 9: Haplotype counts and frequencies across the whole population. The two most common haplotypes include the two new-found alleles at DRB1, one of which is also present in a low frequency combination, but always associated to the most common DQA1 allele.

Also reflecting the distribution of private alleles, three haplotypes were only present in the H group (Tab. 10, Fig. 5), and two in the Wt group, although in the latter being the least common ones.

Haplotype	Nomenclature (DRB1 / DQA1 / DQB1)	Wt (n=34)		Ph (n=10)		H (n=26)	
		2n	freq	2n	freq	2n	freq
1	Calu-DRB1*13-newItaly / DQA1*005011 / DQB1*03901	27	0.36	12	0.02	15	0.29
2	DRB1*09201-newItaly / DQA1*005011 / DQB1*00701	27	0.36	6	0.02	6	0.12
3	DRB1*03601 / DQA1*01201 / DQB1*03501	11	0.14	1	0.01	9	0.17
4	DRB1*02001 / DQA1*00401 / DQB1*01303	2	0.03	0	0.00	7	0.13
5	DRB1*03901 / DQA1*00201 / DQB1*02002	0	0.00	0	0.00	4	0.08
6	DRB1*03701 / DQA1*005011 / DQB1*00701	3	0.04	1	0.00	2	0.04
7	DRB1*01501 / DQA1*00601 / DQB1*00301	2	0.03	0	0.00	3	0.06
8	DRB1*00101 / DQA1*00101 / DQB1*00201	1	0.01	0	0.00	1	0.02
9	DRB1*09201 / DQA1*00601 / DQB1*02002	0	0.00	0	0.00	1	0.02
10	Calu-DRB1*13-newItaly / DQA1*005011 / DQB1*00701	1	0.01	0	0.00	1	0.02
11	DRB1*01501 / DQA1*00401 / DQB1*00301	0	0.00	0	0.00	3	0.06
12	DRB1*02001 / DQA1*00401 / DQB1*03901	1	0.01	0	0.00	0	0.00
13	DRB1*09201-newItaly / DQA1*005011 / DQB1*03501	1	0.01	0	0.00	0	0.00

Table 10: Haplotype counts and frequencies by group. Haplotypes that are private to one group are highlighted in bold in the corresponding column.

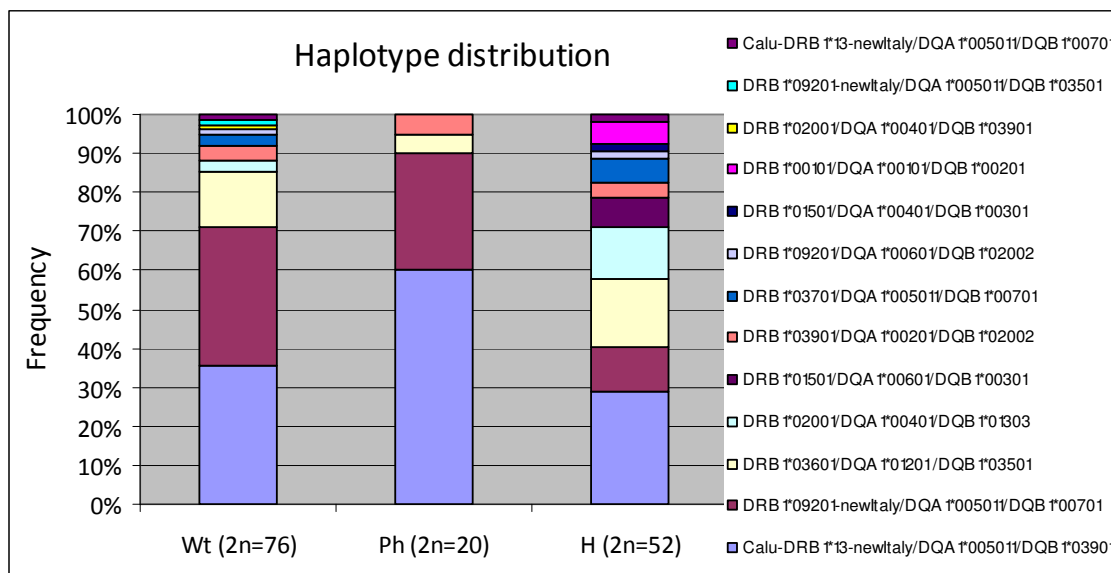


Figure 5: Distribution of the haplotype frequencies by group. Some of the haplotypes are private to the Wt or to the H group.

The highest genetic distance (Tab. 11) was shown by the H group, as expected from their admixed origin, whereas the Ph individuals were closer to the Wt rather than to the H group, confirming their genetic assignment to the wolf cluster based on STRs.

	Wt	Ph	H
Wt	–	0.026	0.032
Ph	0.014	–	0.075
H	0.058	0.095	–

Table 11: Genetic distances between population at the MHC *loci*. Fst values are showed above the diagonal, Unbiased Nei Genetic Distance values below.

Similarly, when comparing the haplotype frequencies between groups, the linear correlation was maximum between Wt and Ph wolves ($R^2 = 0.8396$), and lower between H and the other two groups ($R^2 = 0.5875$ and $R^2 = 0.6379$ with Wt and Ph, respectively), indicating a higher frequency similarity of the phenotypically-unusual wolves to the wild-type ones rather than to the admixed individuals.

This was also confirmed by comparing each group's frequencies to the ones from the whole population, with only the H group frequencies being significantly different from the general ones ($p=0.03$, χ^2 test).

The pairwise Kolmogorov-Smirnov test on the frequency distributions returned a maximum difference between the cumulative distributions between the Ph and H groups, although not significant ($D=0.5$, $p=0.066$).

Haplotype	Nomenclature (DRB1 / DQA1 / DQB1)	A (n=6)		nAp (n=10)		cAp (n=11)		sAp (n=11)	
		2n	freq	2n	freq	2n	freq	2n	freq
1	Calu-DRB1*13-newItaly/DQA1*005011/DQB1*03901	4	0.33	9	0.45	10	0.45	4	0.18
2	DRB1*09201-newItaly/DQA1*005011/DQB1*00701	7	0.58	5	0.25	7	0.32	8	0.36
3	DRB1*03601/DQA1*01201 /DQB1*03501	1	0.08	1	0.05	3	0.14	6	0.27
4	DRB1*02001/DQA1*00401/DQB1*01303	0	0.00	2	0.10	0	0.00	0	0.00
6	DRB1*03901/DQA1*00201/DQB1*02002	0	0.00	1	0.05	1	0.05	1	0.05
7	DRB1*03701/DQA1*005011/DQB1*00701	0	0.00	0	0.00	0	0.00	2	0.09
8	DRB1*09201/DQA1*00601/DQB1*02002	0	0.00	0	0.00	0	0.00	1	0.05
10	Calu-DRB1*13-newItaly / DQA1*005011 / DQB1*00701	0	0.00	0	0.00	1	0.05	0	0.00
12	DRB1*02001/DQA1*00401/DQB1*03901	0	0.00	1	0.05	0	0.00	0	0.00
13	DRB1*09201-newItaly/DQA1*005011/DQB1*03501	0	0.00	1	0.05	0	0.00	0	0.00

Table 12: Haplotype distribution among geographic groups, only considering wild-type wolves (Wt). Haplotypes that have only been found within one group are highlighted in bold in the corresponding column.

From the geographic distribution of the haplotypes in the Wt wolves (Tab. 12, Fig. 6), we can see a maximum number of haplotypes in the Northern (nAp, 7 haplotypes) and Southern Apennine (sAp, 6 different haplotypes), slightly lower in the Central Apennine (cAp, 5 haplotypes) and minimum in the Alps (A, 3 haplotypes). Within geographic groups, all *loci* and haplotypes combinations are not significantly different from the expected HWE distribution.

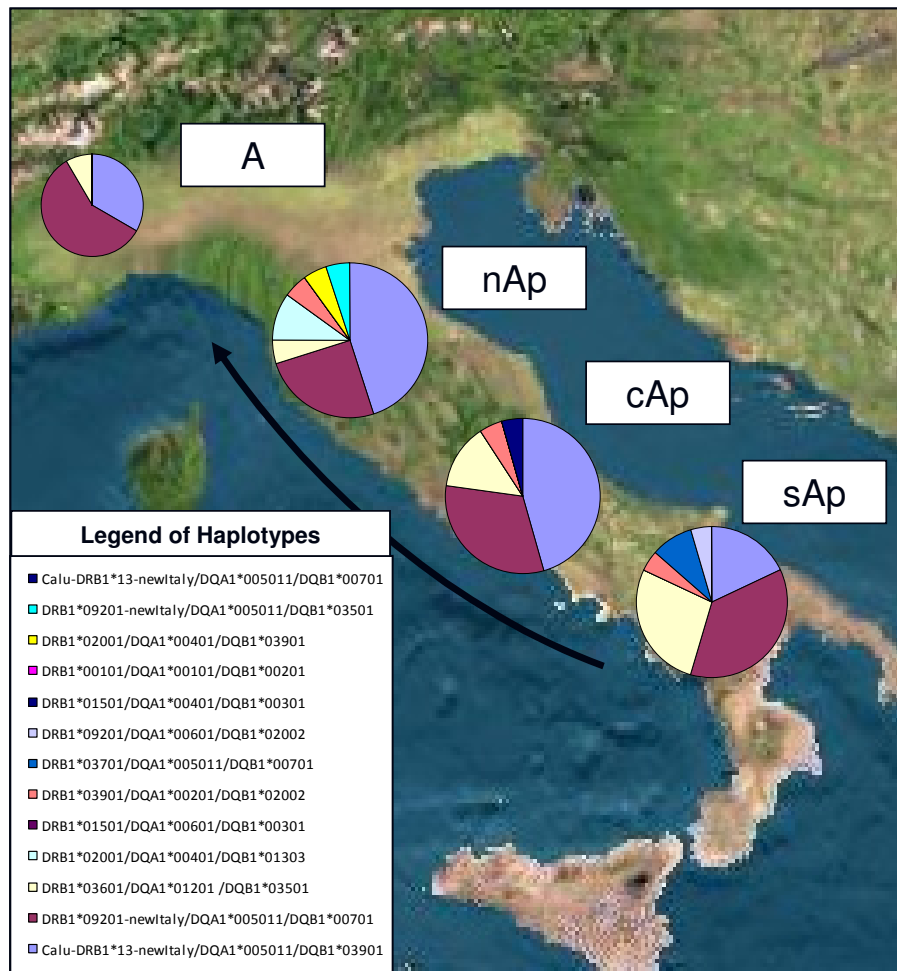


Figure 6: Haplotype distribution in wild-type wolves, split by their geographic origin: A=Alps ($2n=12$); nAp=Northern Apennine ($2n=20$); cAp=Central Apennine ($2n=22$); sAp=Southern Apennine ($2n=22$).

Looking at the rarefaction curves in the number of described alleles, all the geographic groups look close to (but can have not reached) a plateau, with similar patterns throughout the Apennine and a lower variability in the Alps (Fig. 7).

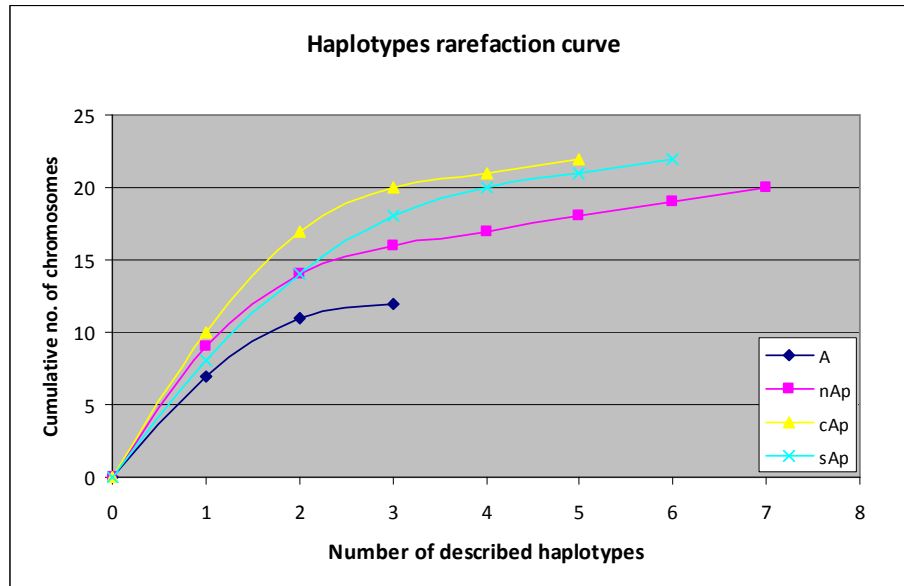


Figure 7: Distribution of the described haplotypes relative to the cumulative number of analyzed samples for each geographic group (A=Alps, nAp=Northern Apennine, cAp=Central Apennine, sAp=Southern Apennine).

The alignment of the corresponding portion of all the sequences available at the three MHC *loci* for the *Canis* genus highlighted some overlaps in the sequence names, which have been grouped whenever two alleles could not be resolved as different in the considered region.

For each allele, we identified all the species or *taxa* in which it has been described so far in literature.

Using the same sequences, the phylogenies reconstructed at each *locus* in MEGA show that the alleles found in the Italian wolf population are dispersed throughout the trees, not clustering in any specific clade (Fig. 8 to 10).

The two newly described alleles at DRB1 appeared to be respectively basal (Calu-DRB1*13-newItaly) and terminal (DRB1*09201newItaly) relative to the closest ones described in previous studies. Therefore, the latter is more likely to be really unique to the Italian wolf population.

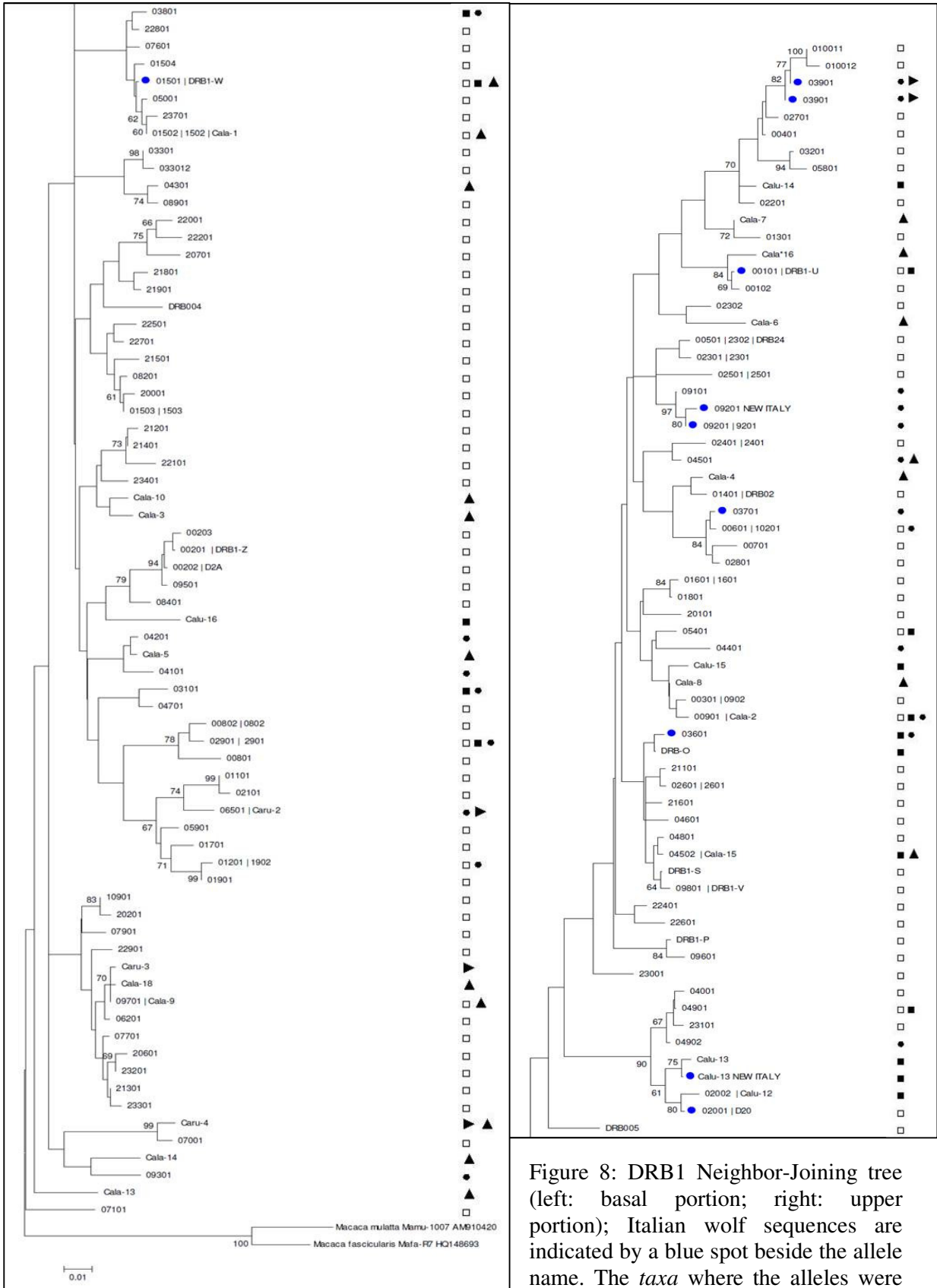


Figure 8: DRB1 Neighbor-Joining tree (left: basal portion; right: upper portion); Italian wolf sequences are indicated by a blue spot beside the allele name. The *taxa* where the alleles were described to date are indicated on the right column (Dog □; European wolf ■; American wolf ●; coyote ▲; red wolf ►; Ethiopian wolf ◄)

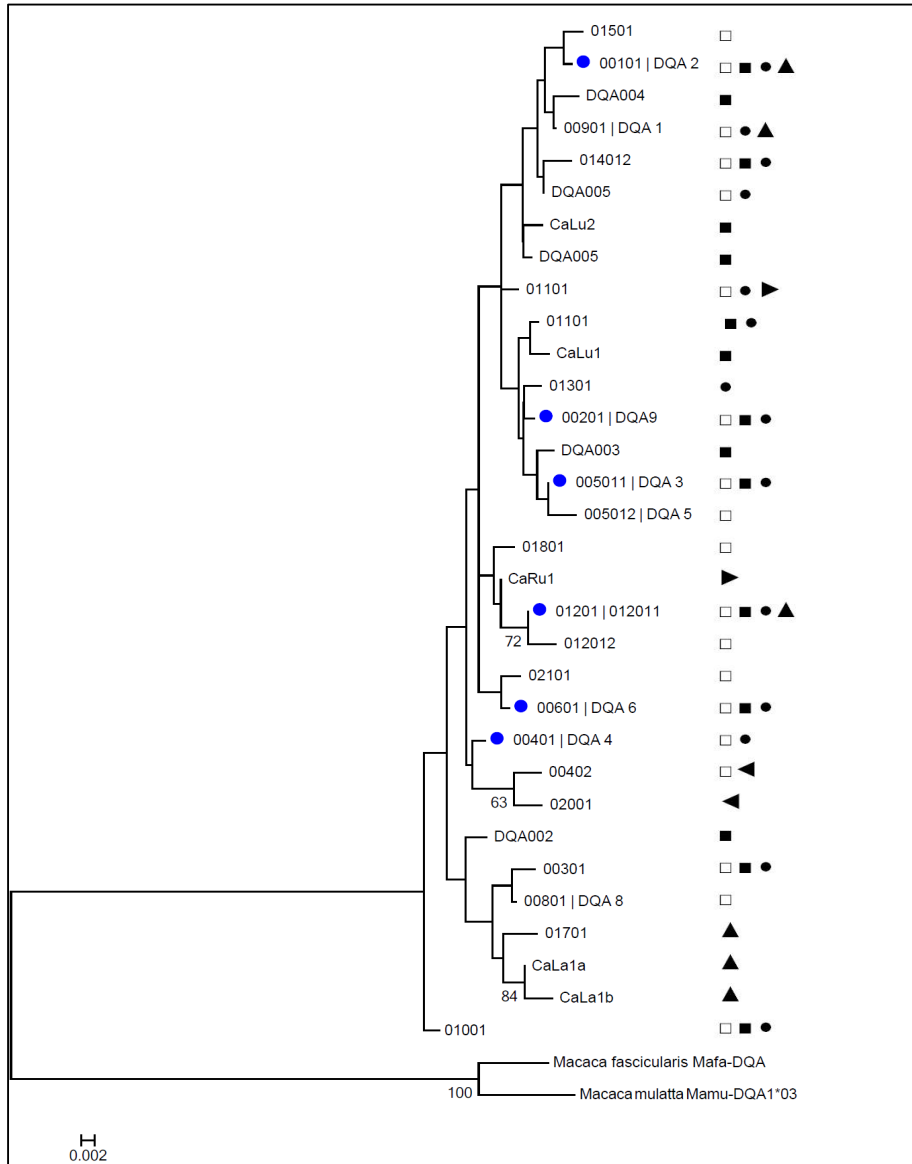


Figure 9: DQA1 Neighbor-Joining tree; Italian wolf sequences are indicated by a blue spot beside the allele name. The *taxa* where the alleles were described to date are indicated on the right column (Dog □; European wolf ■; American wolf ●; coyote ▲; red wolf ►; Ethiopian wolf ◄)

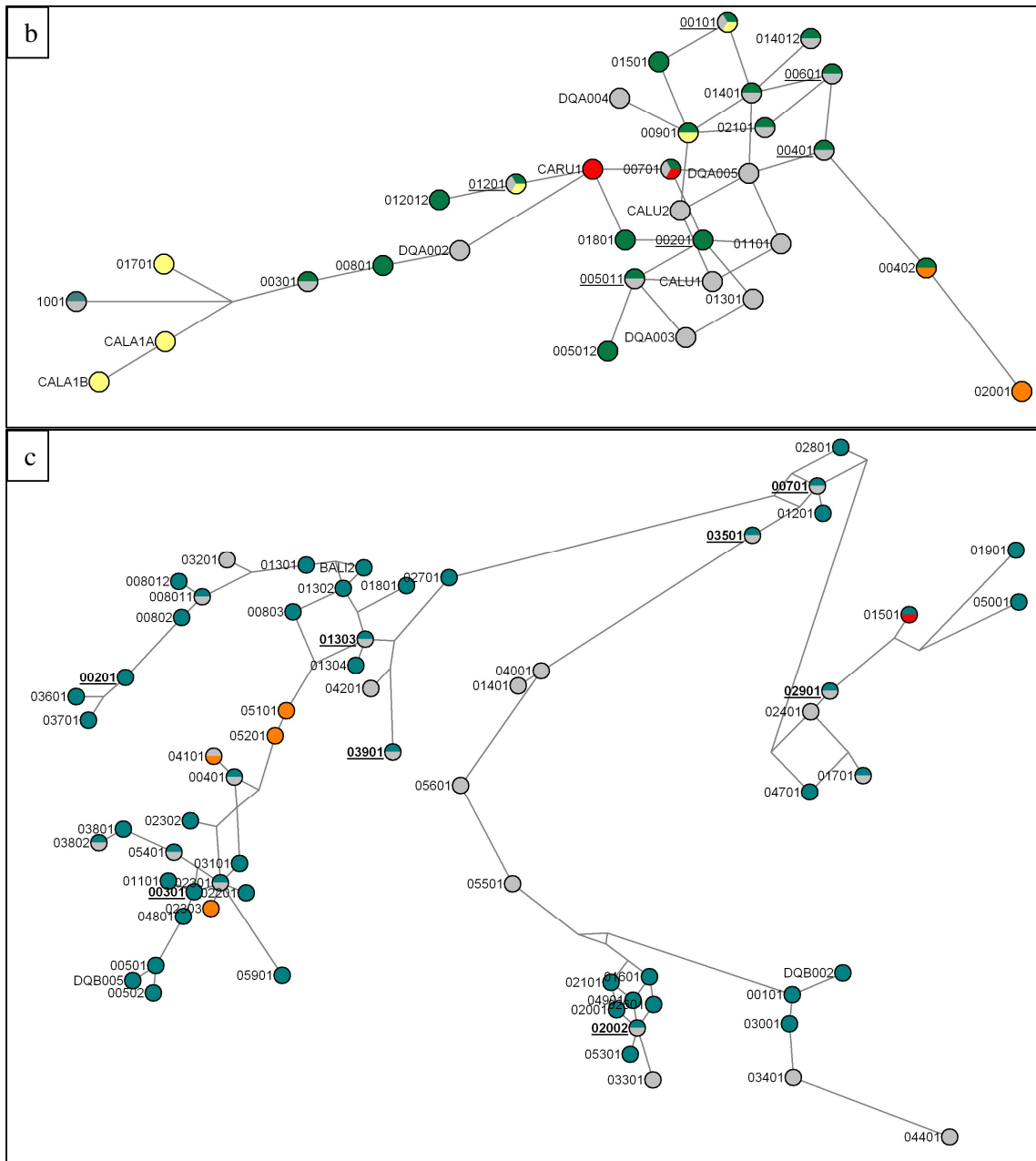


Figure 11: a) DRB1; b) DQA1; c) DQA1 Network; Italian wolf alleles are underlined. The alleles described in multiple *taxa* are indicated by circles, where the *taxa* are indicated by different colors in the slices (green= dog; grey = wolf; yellow = coyote; red = red wolf; orange = Ethiopian wolf). It is interesting to note the branch leading to DRB1-09201 (with its newly described form).

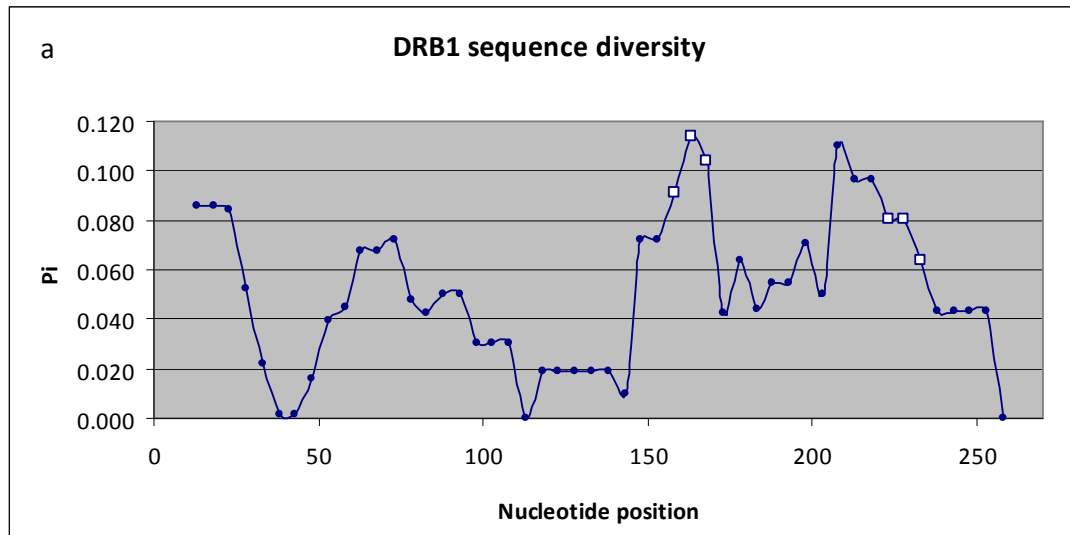
The calculated d_N/d_S values were higher than one at each *locus* (Tab. 13), with DQA1 having zero synonymous mutations, therefore confirming the general pattern of an historical positive selection.

Locus	SynDif	SynPos	d_s	NSynDif	NSynPos	d_N	d_N/d_s
DRB1	2.16	61.74	0.04	11.18	205.26	0.06	1.59
DQA1	0.00	56.13	0.00	1.93	189.87	0.01	N/A
DQB1	1.87	64.44	0.03	10.97	202.56	0.06	1.92
Total	4.03	182.32	0.02	24.07	597.68	0.04	1.82

Table 13: distribution of Synonymous (SynDif) and Non-Synonymous differences (NSynDif), their proportions (d_S ; d_N) relative to the total number of Synonymous (SynPos) and Non-Synonymous sites (NSynPos), and their ratio (d_N/d_S), both by gene and total across *loci*.

The average nucleotide diversity (π) was higher at DRB1 (0.04939) and DQB1 (0.04808) than at DQA1 (0.00783), although varying across the sequences and being maximum in correspondence with some of the Peptide Binding Sites (PBR) of DRB1 and DQB1 (Fig. 12).

Tajima's D values were not significantly deviating from the ones expected under neutrality (DRB1 $D= 1.41844$; DQA1 $D=0.78098$; DQB1 $D= 0.99091$; all $p>0.1$). However, by replicating the test along sliding windows, some of the sites turned out to be significantly ($p<0.05$) deviating from neutrality (white spots in Fig. 12a and 12c), once again in windows close to the PBRs of the β -chains, therefore suggesting the selective pressures specifically occurred in the most functionally-active portions of the sequences.



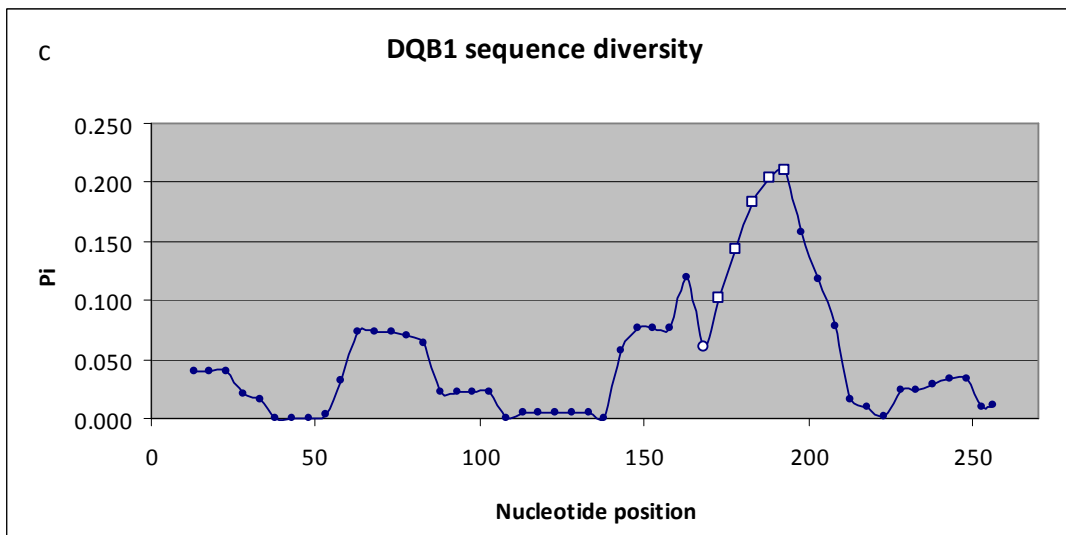
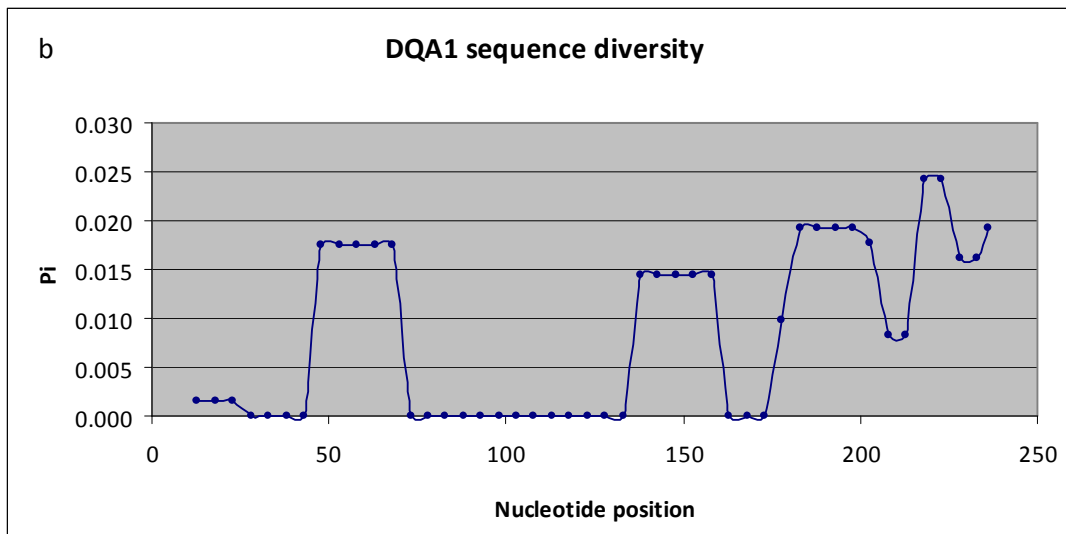


Figure 12: a) DRB1; b) DQA1; c) DQB1. Nucleotide diversity (the average number of nucleotide differences per site between sequences, P_i) as computed in sliding windows of 25bp size (step size=5bp). Windows where the Tajima's D values are significantly ($p < 0.05$) deviating from neutrality are indicated by white-filled squares.

MHC and mating preferences

We successfully sequenced and phased at all MHC *loci* the alleles of 52 breeders, 7 of their offspring and 10 unrelated individuals, for a total of 69 individuals belonging to 19 different packs. Two additional reproducers and three offspring were already sampled and sequenced in the first part of the study, summing up to 74 individuals of known pack membership. Within some of the packs, more than one breeding pair was identified through the years, therefore leading to reconstruct a total of 26 different pairs, whereas the remaining 8 pairs deduced

from pedigree reconstruction were not complete, with only one of the mates successfully sequenced at all *loci*.

Accordingly with the increased number of samples, two low-frequency ($\text{freq} < 0.05$) alleles not previously found in the Italian wolf population were identified (DLA-DQA1*00901 and DLA-DQB1*001019) (Tab. 14), as well as seven new rare ($\text{freq} < 0.05$) haplotype combinations (Tab. 15). Therefore, we described a total of 9 alleles in the Italian wolves for DRB1 and DQB1, and 7 for DQA1, combining into 20 different haplotypes.

DRB1 allele	n	freq	Decribed in	GeneBank_name	AN
Calu-DRB1*13-newItaly	61	0.43	Never	-	-
09201-newItaly	36	0.25	Never	-	-
DLA-DRB1*02001	20	0.14	D	D20	U58684.1
DLA-DRB1*03601	11	0.08	We, Wa	03601	AF336110.1
DLA-DRB1*09201	6	0.04	Wa	09201	AM408904.1
DLA-DRB1*01501	5	0.04	D	DRB1-W	DQ056281.1
DLA-DRB1*03901	3	0.02	Wa ,Rw	03901	AF343740.1
DLA-DRB1*03701	0	0.00	Wa	03701	AF343738.1
DLA-DRB1*00101	0	0.00	D	DRB1-U; DRB1-Q	DQ056278.1;DQ056274.1

DQA1 allele	n	freq	Decribed in	GeneBank_name	AN
DLA-DQA1*005011	97	0.68	We, Wa, D, Wm	DQA3	U44787.1
DLA-DQA1*00401	19	0.13	Wa, D	DQA4	U44788.1
DLA-DQA1*01201	11	0.08	We, Wa, D, C	01201	AF343734.1
DLA-DQA1*00601	7	0.05	We, Wa, D	DQA6	U44790.1
DLA-DQA1*00901	5	0.04	Wa, D, C	DQA1	U44785.1
DLA-DQA1*00201	3	0.02	We, Wa, D	DQA9	U75455.1
DLA-DQA1*00101	0	0.00	We, Wa, D, C, Wm	DQA2	U44786.1

DQB1 allele	n	freq	Decribed in	GeneBank_name	AN
DLA-DQB1*03901	55	0.39	We	03901	AY126651.1
DLA-DQB1*00701	41	0.29	D, Wm, We, Wa	DQB4	AF043149.1
DLA-DQB1*01303	20	0.14	D, Wm, We, Wa	DQB7	AF043152.1
DLA-DQB1*03501	12	0.08	We, Wa, D	03501	AJ311107.1
DLA-DQB1*02002	6	0.04	Wa, D	DQB19	AF043164.1
DLA-DQB1*00101	5	0.04	D	DQB19	AF043164.1
DLA-DQB1*02901	3	0.02	We	02901	AY126648.1
DLA-DQB1*00301	0	0.00	D	DQB6	AF043151.1
DLA-DQB1*00201	0	0.00	D	DQB3	AF043148.1

Table 14: Official names and frequencies of the alleles found at each *locus* in the non-invasively sampled Northern Apennine population, with corresponding GenBank names and accession numbers (AN), and the canid populations in which they were described to date: We=European wolf; Wa=North American wolf; Wm=Mexican wolf; Rw=red wolf; D=dog; C=coyote ('n' indicates the number of chromosomes carrying a given allele).

Haplotype	Nomenclature (DRB1 / DQA1 / DQB1)	Tissue samples		Non-invasive samples		Total	
		n	freq	n	freq.	n	freq
1	Calu-DRB1*13-newitaly / DQA1*005011 / DQB1*03901	9	0.45	53	0.38	62	0.39
2	DRB1*09201-newitaly / DQA1*005011 / DQB1*00701	5	0.25	35	0.25	40	0.25
3	DRB1*03601 / DQA1*01201 / DQB1*03501	1	0.05	11	0.08	12	0.08
4	DRB1*02001 / DQA1*00401 / DQB1*01303	2	0.10	16	0.12	18	0.11
5	DRB1*03901 / DQA1*00201 / DQB1*02002	1	0.05	0	0.00	1	0.01
6	<i>DRB1*03701 / DQA1*005011 / DQB1*00701</i>	0	0.00	0	0.00	0	0.00
7	<i>DRB1*01501 / DQA1*00601 / DQB1*00301</i>	0	0.00	0	0.00	0	0.00
8	<i>DRB1*00101 / DQA1*00101 / DQB1*00201</i>	0	0.00	0	0.00	0	0.00
9	DRB1*09201 / DQA1*00601 / DQB1*02002	0	0.00	5	0.04	5	0.03
10	Calu-DRB1*13-newitaly / DQA1*005011 / DQB1*00701	0	0.00	5	0.04	5	0.03
11	<i>DRB1*01501 / DQA1*00401 / DQB1*00301</i>	0	0.00	0	0.00	0	0.00
12	DRB1*02001 / DQA1*00401 / DQB1*03901	1	0.05	0	0.00	1	0.01
13	DRB1*09201-newitaly / DQA1*005011 / DQB1*03501	1	0.05	0	0.00	1	0.01
14	DRB1*01501 / DQA1*00901 / DQB1*00101	0	0.00	5	0.04	5	0.03
15	DRB1*03901 / DQA1*00201 / DQB1*DQB1*02901	0	0.00	3	0.02	3	0.02
16	Calu-DRB1*13-newitaly / DQA1*005011 / DQB1*01303	0	0.00	1	0.01	1	0.01
17	Calu-DRB1*13-newitaly / DQA1*00601 / DQB1*03901	0	0.00	1	0.01	1	0.01
18	DRB1*09201-newitaly / DQA1*00401 / DQB1*01303	0	0.00	1	0.01	1	0.01
19	DRB1*02001 / DQA1*005011 / DQB1*01303	0	0.00	1	0.01	1	0.01
20	DRB1*02001 / DQA1*00601 / DQB1*02002	0	0.00	1	0.01	1	0.01

Table.15: Haplotype distribution among 69 individuals non-invasively sampled in the Northern Apennine, plus 10 wild-type wolves previously sequenced in the same region. Haplotypes that have been only found in this additional sample are highlighted in bold, the ones that are not present in the Northern Apennine are indicated in italic.

Contrary to what described across the whole population, when considering only breeding individuals the allele frequencies significantly deviated from HWE at all *loci* ($p < 0.05$ at DRB1, $p < 0.01$ at DQA1 and DQB1, χ^2 test), suggesting the presence of features departing from the equilibrium.

Among the successfully sequenced wolves, we were able to reconstruct 10 full parents-offspring trios, in all cases confirming the Mendelian inheritance of the alleles. In two of the cases, however, the haplotype combinations in the offspring were different from the most probable ones phased in the parents, therefore suggesting the presence of alternative - although less-probable- haplotypes in the parents, or the occurrence of a recombination event between DQA1 and DQB1 *loci*.

With surprise, the test on the levels of allele sharing between mates (one, two or no alleles in common as expected from the allele frequencies of each mate) turned out to be significantly different from the expectations at all *loci*, but with an excess of cases where one or both alleles were shared (DRB1 and DQB1, $p = 0.016$ and $p = 0.034$, respectively; χ^2 test), and both alleles were in common between mates (DQA1, $p = 0.025$; χ^2 test). The same test was

significant also when considering the haplotypes as a whole, showing the same excess of one or two haplotypes being shared between mates.

Combination	Haplotypes		DRB1		DQA1		DQB1	
	Observed	Expected	Observed	Expected	Observed	Expected	Observed	Expected
2 different	22	30.8	16	26.2	8	12.7	20	27.9
1 in common	26	18.3	28	20.9	20	24.1	24	20.0
2 equal	4	2.9	8	4.9	24	15.3	8	4.1
total	52	52	52	52	52	52	52	52
p value	0.046		0.016		0.025		0.034	

Table 16: Number of occurrences of breeders sharing one, two or both alleles or haplotypes with their actual mates, compared to the ones expected under a random chance of mating according to each individual's frequencies. The p values are the ones computed by a χ^2 test. Higher-than-expected combinations are shown in bold.

The same unexpected pattern was partially confirmed by looking at the values of relatedness between mates compared to that from all the possible combinations among breeders. The sum of the r values between mates was higher than the average ($S=1.704$; Fig. 17), although not significantly ($p=0.080\pm 0.004$ S.D.). However, the same probability would decrease to $p=0.034\pm 0.003$ S.D if considering the wolves that reproduced in more than one pack or pair as independent mates.

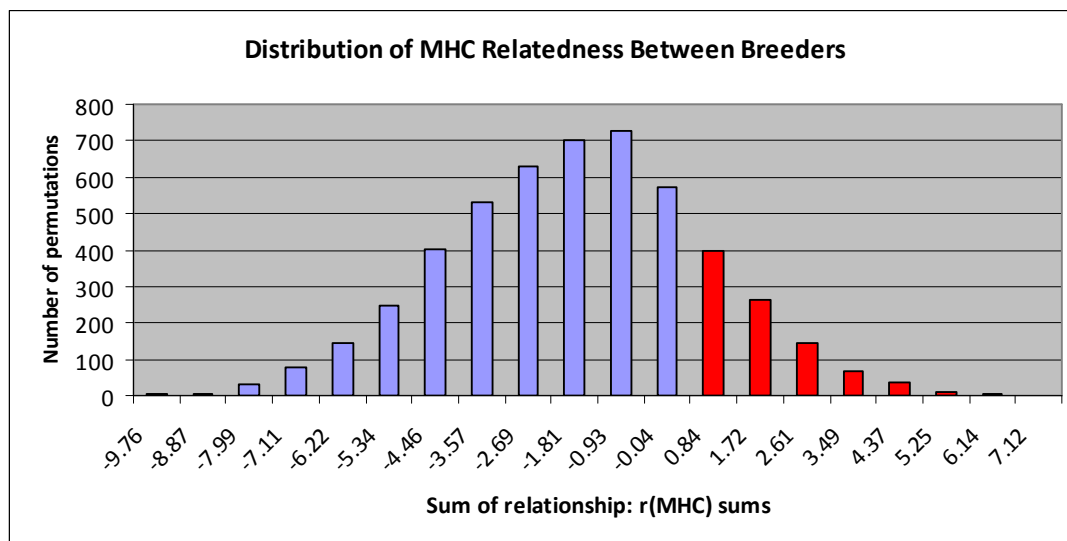


Figure 17: Distribution of the sum of pairwise values of relatedness between wolves at the 3 MHC *loci*, computed within the actual mates' group, and within random groups obtained by permuting individuals in the pairs. Males and females have been permuted independently. The red bars indicate values greater than the ones observed in the actual mates' groups in the last of 10 iterations. The p values were computed on 5000 permutations at each repetition.

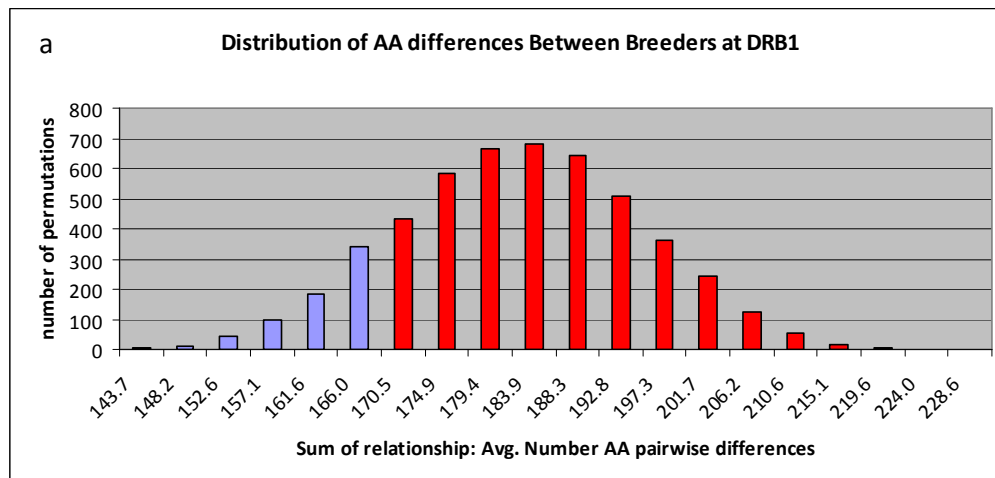
When the test was performed using r values computed at the 12 microsatellite *loci*, the sum of relatedness within the breeders' group was not significantly different ($p= 0.218\pm 0.004$ S.D) from the random one obtained through permutations.

Therefore, we further investigated this interesting pattern by considering the number of pairwise AA differences between mates at each *locus*. This metric reflects the divergence at the functional level better than a mere qualitative difference between allele as described by relatedness.

Locus	DRB1	DQA1	DQB1	Total
Mean value S	182.5	45.2	203.2	431.1
S.D.	12.5	3.4	15.7	26.4
Observed value S	172.8	38.8	173.3	384.8
Mean P Value:	0.221	0.030	0.026	0.039
S.D.	0.006	0.003	0.002	0.002

Table 17: Observed values of the sum S of pairwise AA differences between actual mates at the 3 MHC *loci* and in total, compared to the ones within random groups obtained by permuting individuals in the pairs, whose mean values are shown. The p values and their standard deviations were computed on 5000 permutations for 10 iterations.

The number of pairwise AA differences between actual mates compared to a random mating was significantly lower than expected for DQA1 and DQB1 (Tab. 17), not significantly lower for DRB1, and once again significantly lower for the total across *loci* (Fig. 14a, b, c, d).



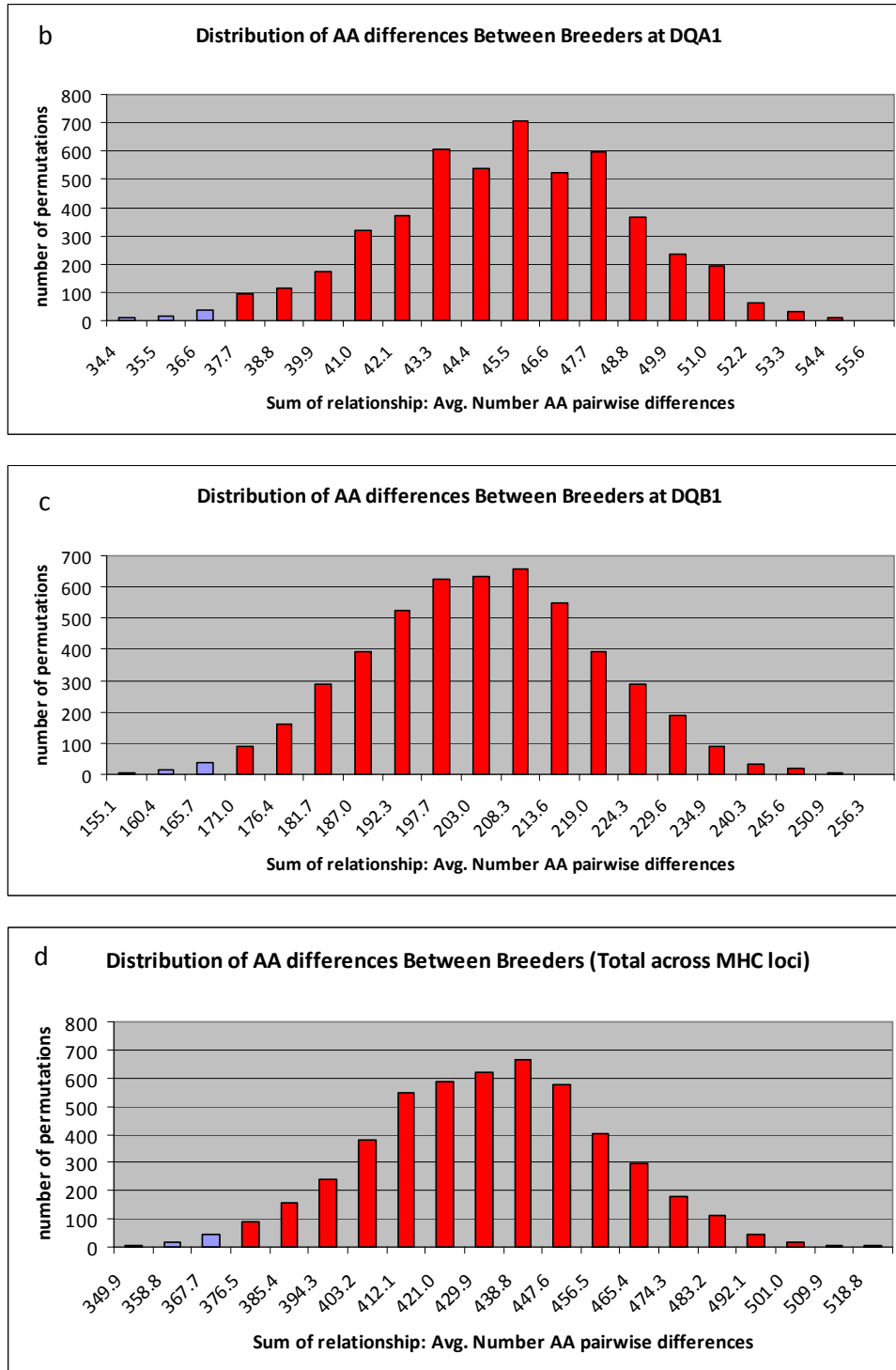


Figure 14: Distribution of the sum of pairwise AA differences between wolves at the 3 MHC *loci* and across them, computed within the actual mates' group, and within random groups obtained by permuting individuals in the pairs. Males and females have been permuted independently. The red bars indicate values greater than the ones observed in the actual mates' groups in the last of 10 iterations. The p values were computed on 5000 permutations at each repetition.

We were not able to repeat the test by excluding the two known new-founded packs, since the quality of their sequences did not allow us to reconstruct reliable alleles and haplotypes for both mates; therefore the potential bias coming from new-founded packs was already avoided.

A local genetic structure can be excluded by the results based on the background relatedness at the 12 STR *loci*. However, we wanted to better check if the levels of protein divergence between mates were at least higher than the ones with 10 unrelated wolves of the opposite sex being present in their pack. But also in this case, on average, the protein divergence between mates was lower, and not higher, than between a breeder and its potential alternative mate ($p>0.05$, t-test; Fig. 15).

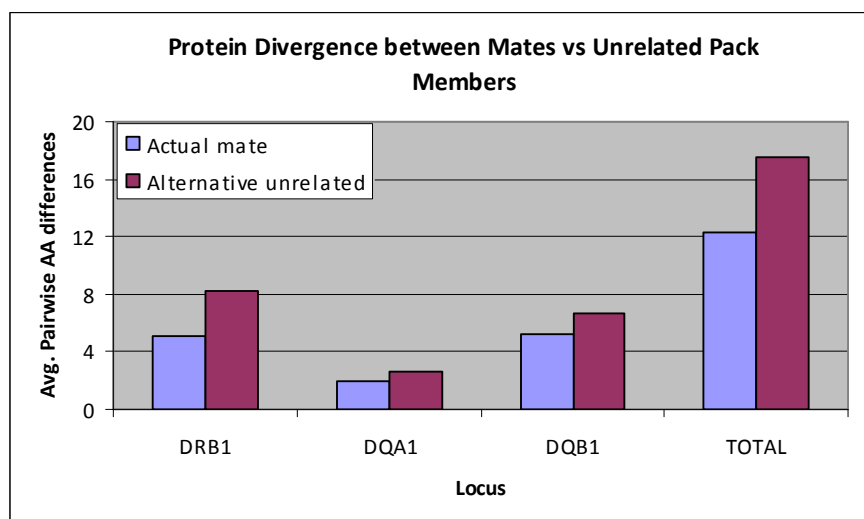


Figure 15: Average number of AA differences between 10 actual breeding pair members, and between 10 breeders and an unrelated potential mate of the opposite sex living in the same pack.

Looking at the heterozygosity levels, we also wanted to test the hypothesis that at least the less heterozygous wolves would benefit from breeding with a mate having higher heterozygosity.

However, also in this case this was not likely to occur, since the average heterozygosity levels of their mates, for both males and females (Fig. 16), were higher for heterozygote than for homozygote breeders, although not significantly different (t-test $p>0.05$).

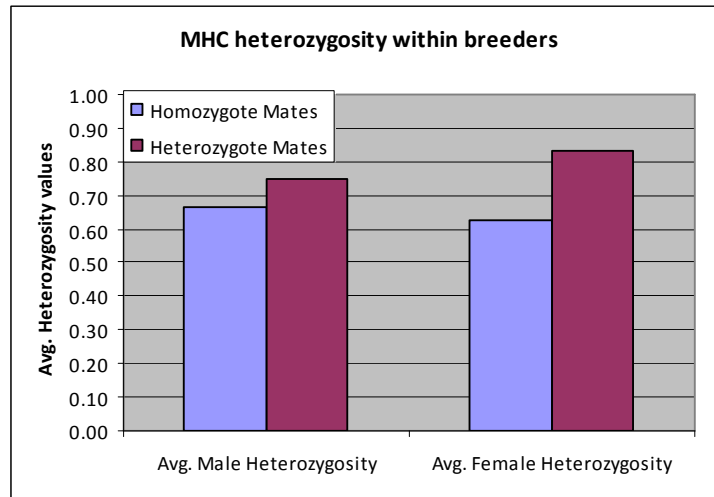


Figure 16: Average levels of heterozygosity of the respective mates in homozygote vs. heterozygote breeders, for each sex.

The same pattern emerged when considering all the four possible classes of *3-loci* heterozygosity in breeders, and the average number of AA differences with their mates (Fig. 17).

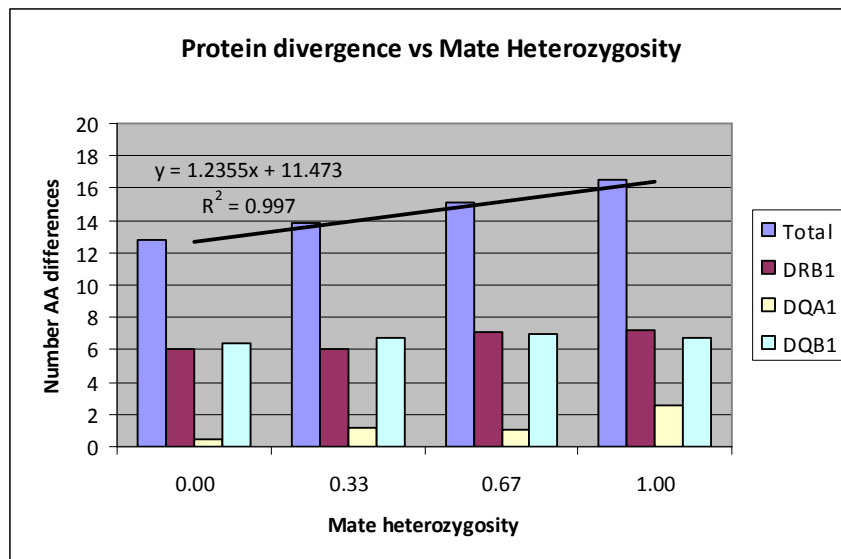


Figure 17: Average number of AA differences between breeders correlated to their heterozygosity classes. The inverse correlation expected under a mating-up process is not observed.

In the end, we tested whether the breeding individuals have, on average, a higher level of heterozygosity compared to unrelated, non-reproducing ones. Also in this case, contrary to the expectations, the reproductive individuals have a lower heterozygosity at the MHC than non-breeding individuals, although not significantly different ($p=0.26$, t-test). On the contrary,

the values are identical when comparing the background heterozygosity at the 12 STR *loci* (Fig. 18).

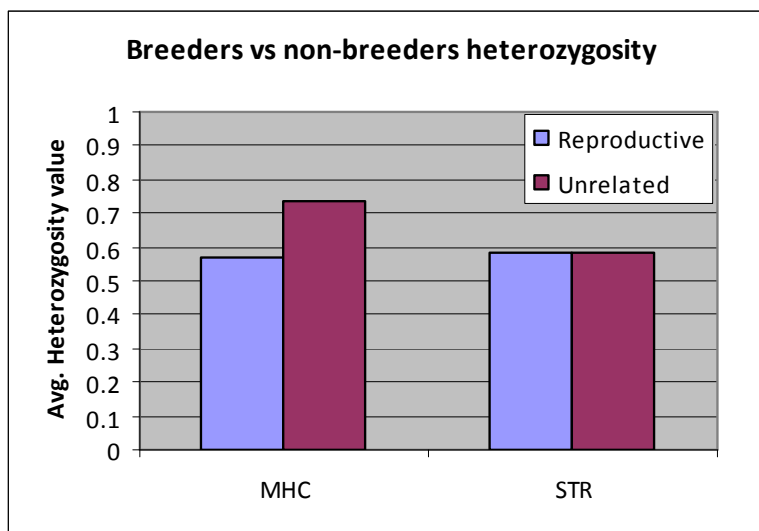


Figure 18: Average levels of heterozygosity of breeders compared to the ones of 10 non-breeding, unrelated individuals living in the same packs, both at the 3 MHC *loci* and the 12 microsatellites.

MHC and its effects on fitness traits

First of all, we checked for any correlation between our fitness traits. As expected, the total offspring (TO) was strongly correlated with the number of years as reproducers (YR) (Pearson's correlation $c=0.76$, $p=1.024e-09$), similarly to what observed between the sampling time (ST) and YR ($c=0.42$, $p=1.836e-03$), or ST and TO ($c=0.32$, $p=0.014$).

However, although linked, the measures are still likely to represent different components of the fitness; therefore they were all kept in the subsequent analyses in order to better address different hypotheses.

Secondly, we took into account the possible confounding factors affecting the chosen fitness traits. The gender, that was known to show slight differences between females and males in the sampling time (Caniglia et al. submitted), did not lead to significant differences in our metrics (t-test, $p>0.05$), although being higher in females compared to males for each trait (Fig. 19).

Also the geographic location of the packs (in the Western, Central or Eastern portion of the study area), did not show significant differences (t-test, $p>0.05$, Fig. 20).

The difference is more marked when considering the mean altitude at which the packs are located. In this case, the pairs living at intermediate altitudes produce a higher total offspring and, consequently, a higher litter size per year (Fig. 21). However, also this difference is not

strictly significant between High (n=18) and Intermediate (n=6) altitude pairs (t-test, respectively $p=0.072$ and $p=0.091$), mainly because of the limited number of observations to be compared (and not computable for the Low altitude class, comprehensive of a single pair).

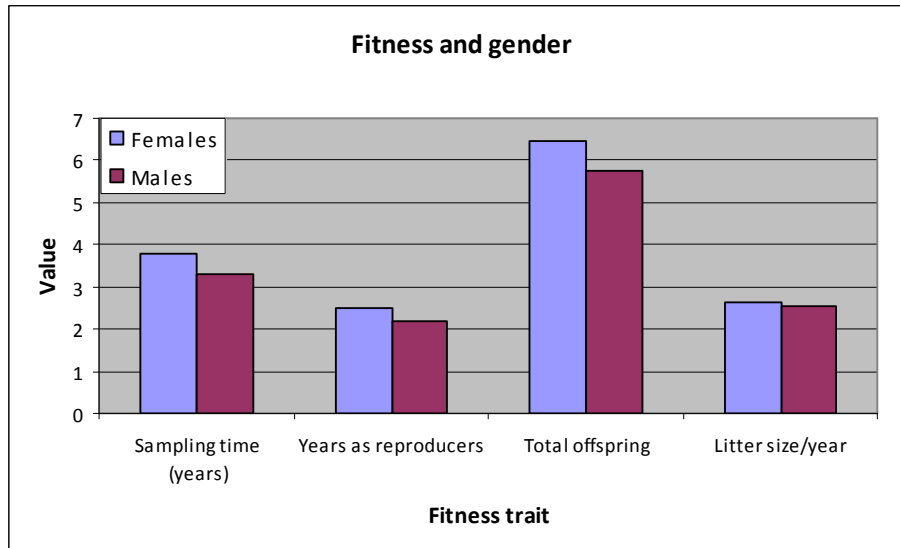


Figure 19: Comparison of the average values for each fitness trait between female and male breeding wolves. The differences in the distribution of the values are not significant (t-test).

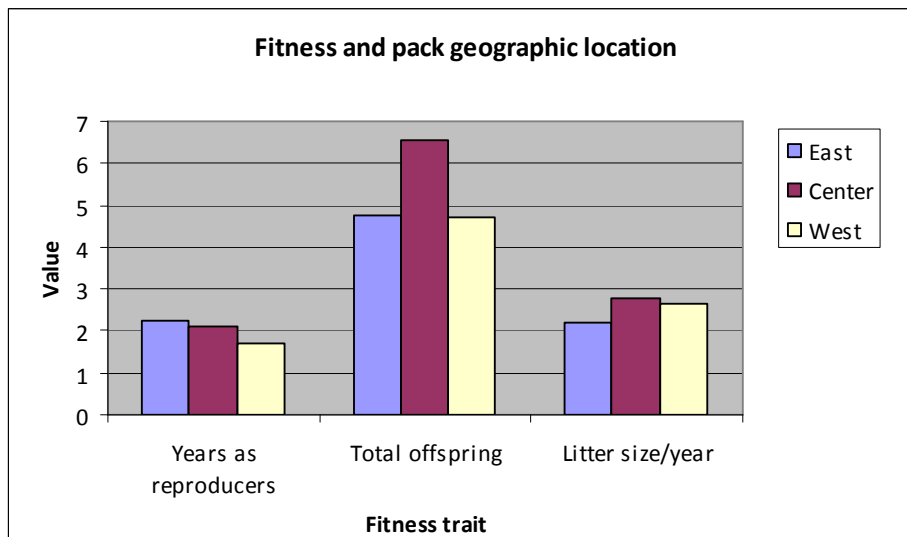


Figure 20: Comparison of the average values for each fitness trait between geographic locations within the study area: East (n=9 breeding pairs), Central (n=9) and West (n=7). The differences in the distribution of the values are not significant (t-test).

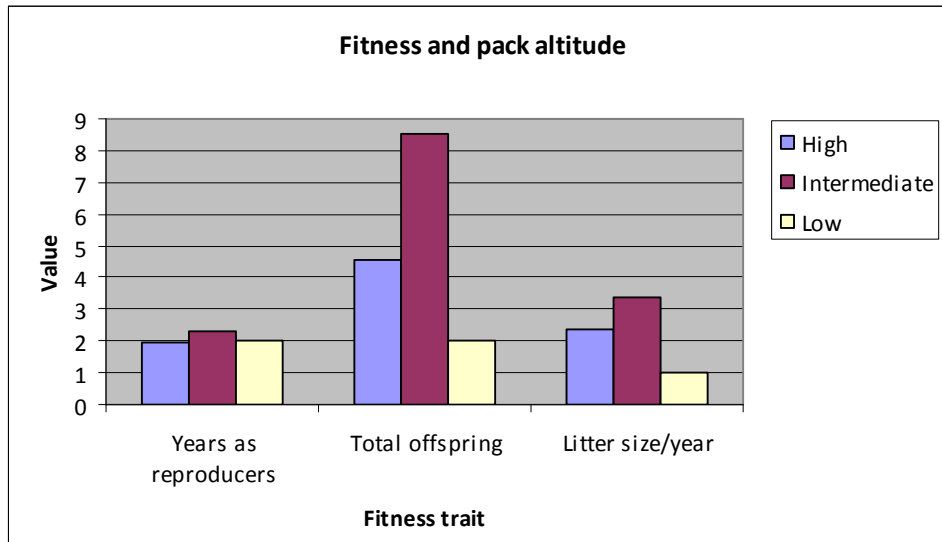
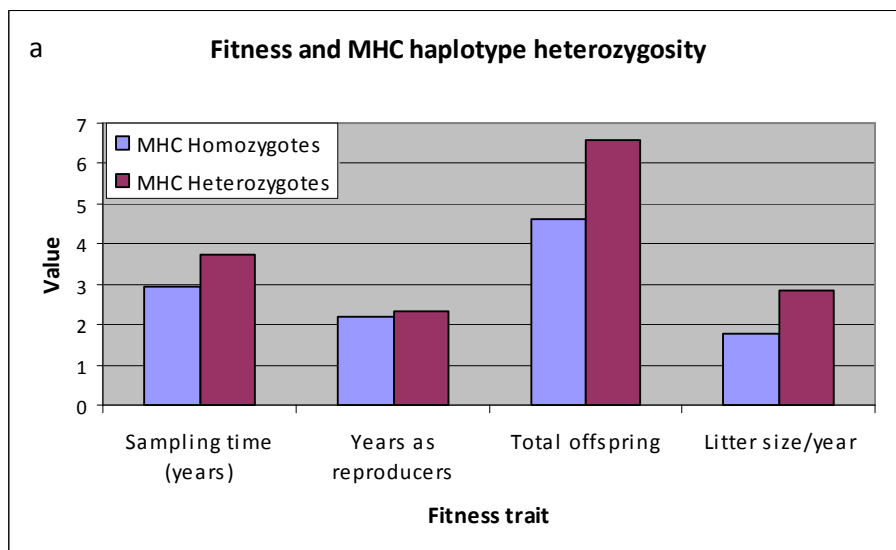


Figure 21: Comparison of the average values for each fitness trait between altitudinal ranges: High (above 800m, n=18 breeding pairs), Intermediate (400-800m, n=6), and Low (<400m, n=1). The difference in the distribution of the values are not significant (High vs. Intermediate $p > 0.05$, t-test; N/A for Low vs. High and Low vs. Intermediate).

The comparison of the fitness traits with the individuals' heterozygosity (both as haplotypes and single *loci*) this time showed higher values in heterozygote vs. homozygote individuals for all the traits but the years as reproducers (Fig. 22), thus concordant with previous results. However, these differences are only significant in the case of LS (Welch two-sample t-test, $t = -2.8904$, $p = 0.004$) when considering heterozygosity at the haplotype or at DQB1 (Welch two-sample t-test, $t = -2.8904$, $p = 0.003$).



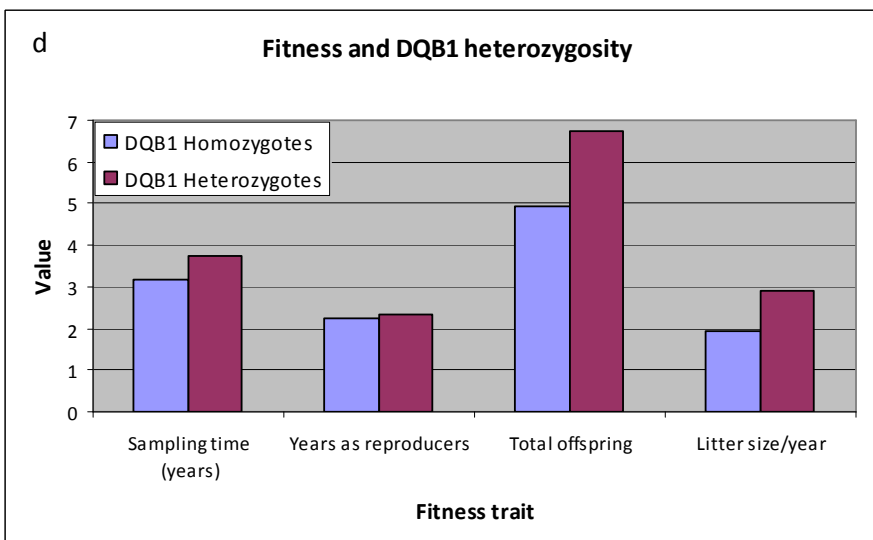
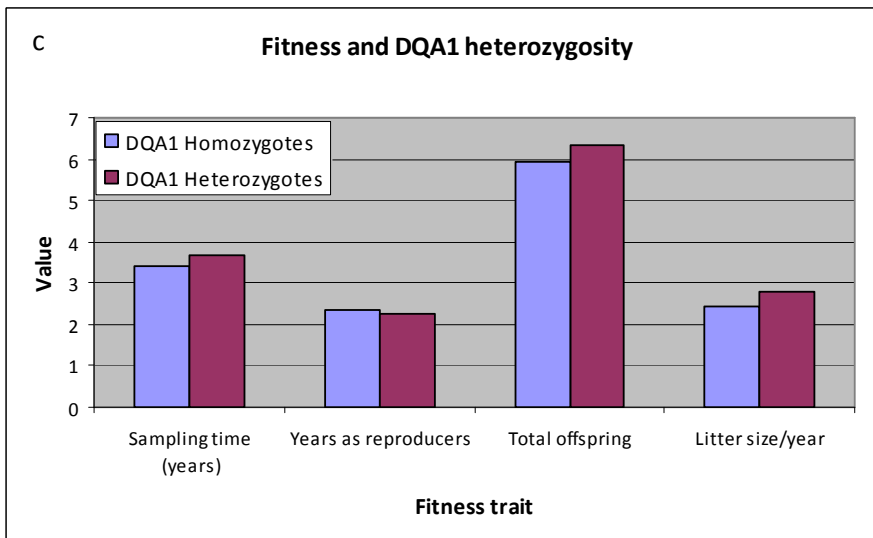
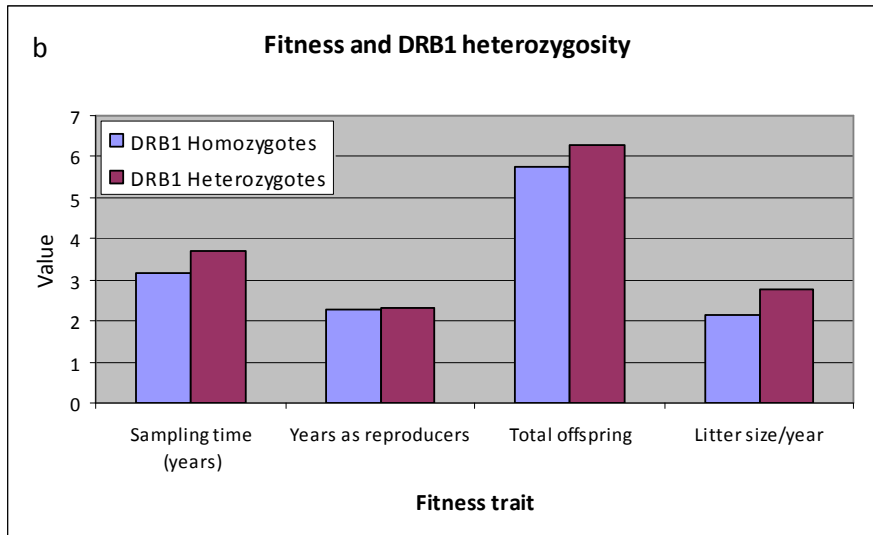


Figure 22: Comparison of the average values of fitness between homozygote and heterozygote wolves, both at the haplotype level (a) and at the single *loci* (b to d).

Also the background heterozygosity showed an effect on the fitness traits, with individuals having higher-than-average heterozygosity levels also showing higher fitness values in terms of TO and LS (Fig. 23), although not strictly significant (Welch two-sample t-test, $p > 0.10$).

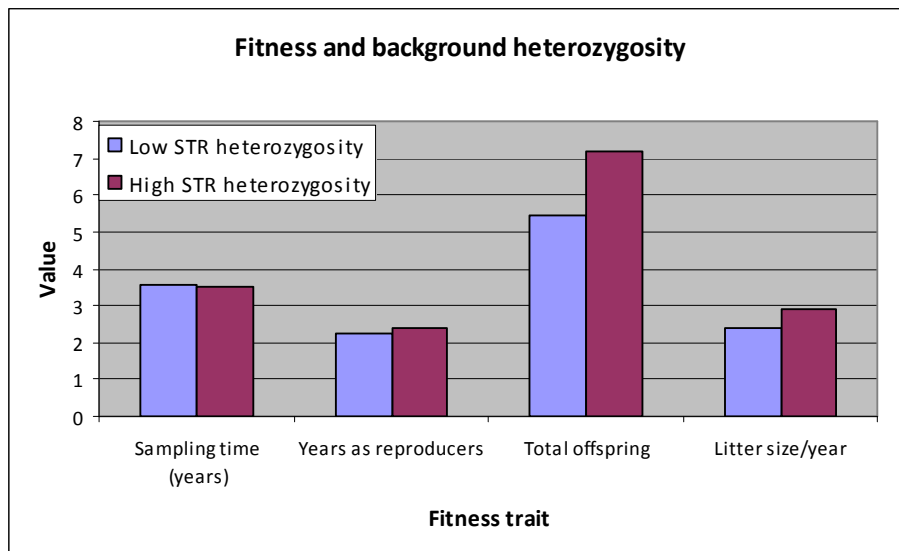
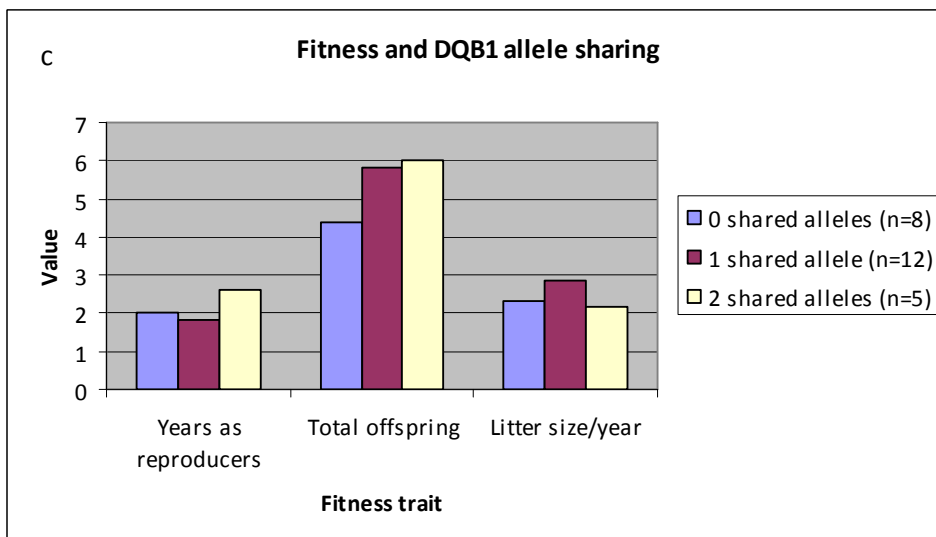
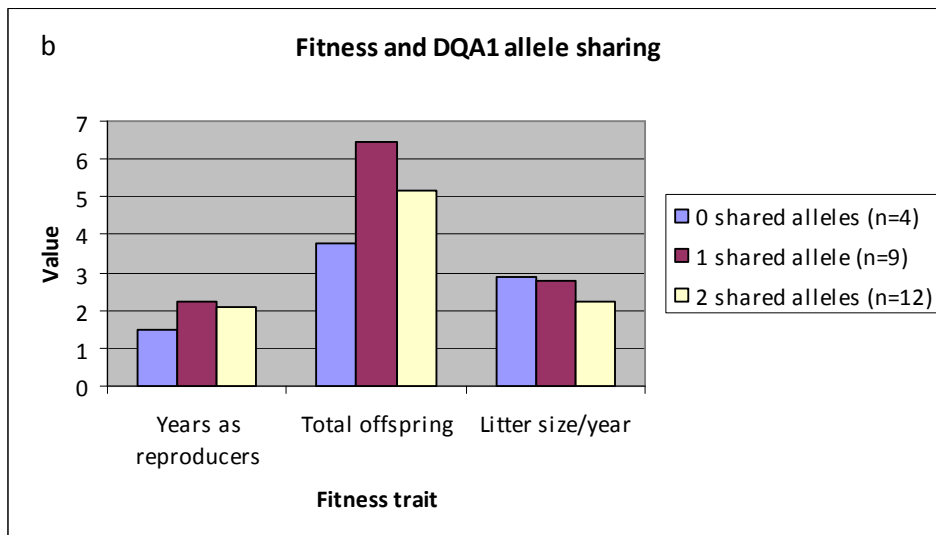
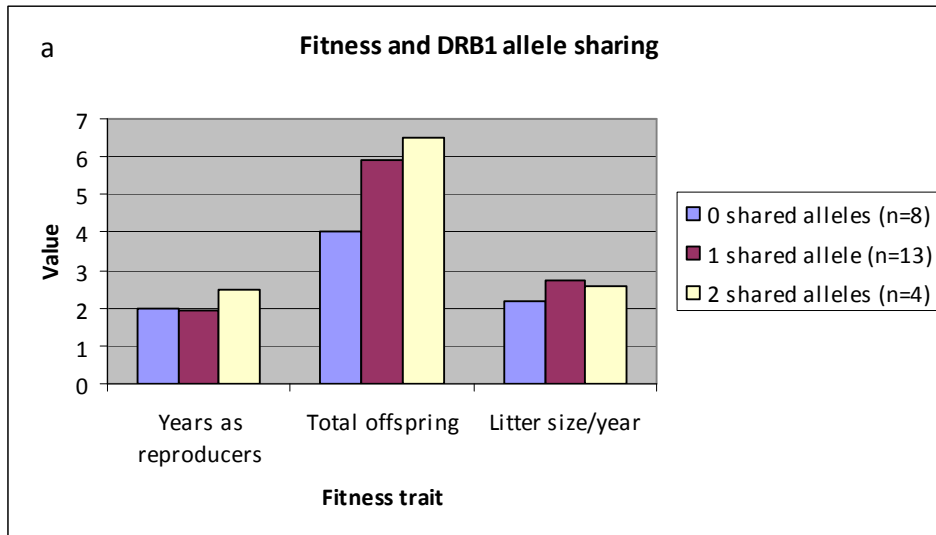


Figure 23: Comparison of the average values of fitness between wolves showing high and low background heterozygosity levels. The groups were determined by being higher ($n=17$) or lower ($n=28$) than the mean ($H_o=0.58$) at the 12 STR *loci*.

$H_o(\text{STR})$ and $H_o(\text{MHC})$ were not significantly correlated (Pearson's correlation $c=-0.08$, $p=0.5981$), suggesting independent contributions to fitness.

At the breeding pairs' level, an interesting pattern emerged when comparing fitness values between pairs sharing zero, one or two haplotypes, with the maximum total offspring and litter size values reached when one haplotype was shared (Fig. 24). However, also in these case, when considered alone these differences turned out to be not significant (pairwise Anova test on the means).



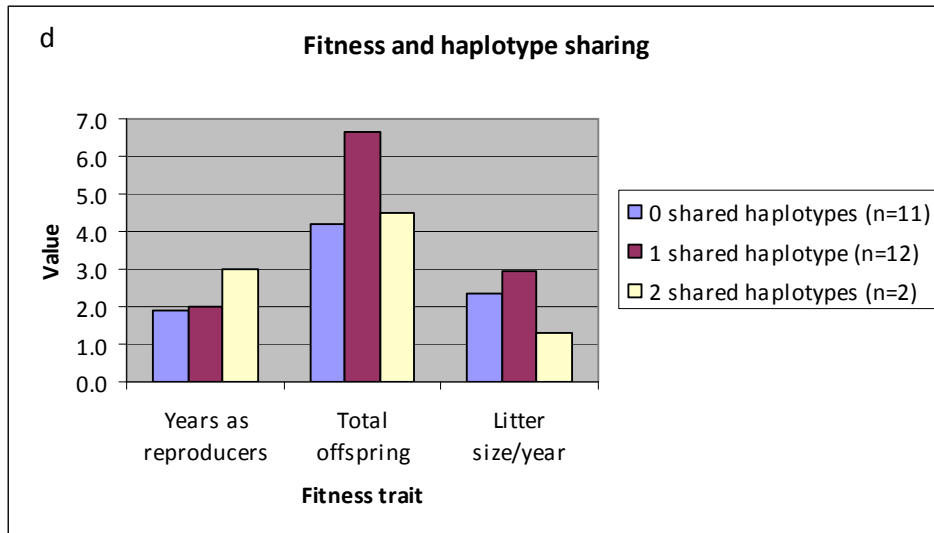


Figure 24: Comparison of the average values of fitness among breeding pairs when the members share zero, one or two alleles (a to c), or haplotypes (d).

Therefore, we tried to combine all the genetic measures into single Linear Models that could better explain the fitness variables. In this way, we were able to reconstruct three representative models ($p < 0.01$, F-statistic), respectively for the number of years as breeders (RY), the total offspring (TO) and the litter size per year (LS).

RY (Tab. 18) was best explained taking into account the relatedness between mates, both at MHC and STR *loci* (the latter showing the most significant effect), but with opposite sign (Fig. 25).

Linear Model	Reproductive Years ~ Avg r(STR) + Avg r(MHC)					
Coefficients	Estimate	Std.Error	t value	P(> t)	Significance	
(Intercept)	2.3303	0.1687	13.815	2E-16	***	
Avg r(STR)	-3.0647	0.8279	-3.702	0.000617	***	
Avg r(MHC)	0.6737	0.2552	2.64	0.0116	*	

Signif. codes:	0 '***'	0.001 '***'	0.01 '*'	0.05 '.'	0.1 ''	1
	value	d.f.	p			
F-statistic	7.789	2 and 42	0.001327			
Residual standard error	1.131	42				
Multiple R-squared:	0.2705					
Adjusted R-squared	0.2358					

Table 18: Linear Model best explaining the number of years as reproducers (RY) for the breeding wolves, based on the values of relatedness (r) at the background 12 STR and at the 3 MHC *loci*. When a wolf reproduced with more than one mate, the average value of r was considered. (d.f.=degrees of freedom)

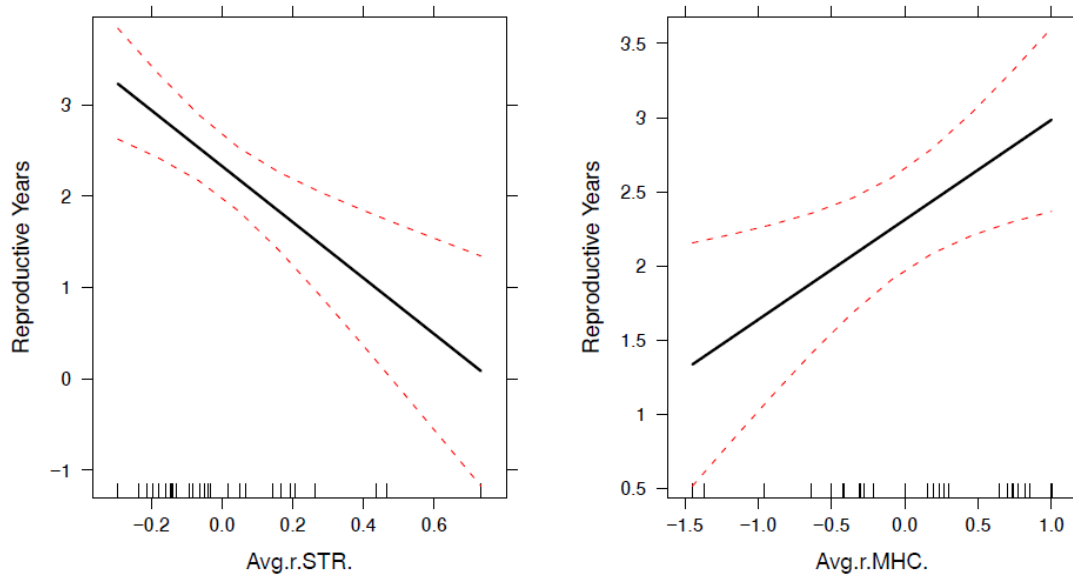


Figure 25: Effect Plot of the correlation between the number of years as reproducers (RY) for the breeding wolves and the two most significant explanatory parameters: the relatedness (r) at the background 12 STR and at the 3 MHC *loci*, clearly showing a negative and a positive correlation, respectively, with the fitness trait.

Similarly, TO was also largely explained by the relatedness at MHC and STR, but with significant contributions ensured by the heterozygosity levels, both at the MHC (avg. across *loci*) and in the background (12 STR) (Tab. 19).

Linear Model	Total offspring \sim Ho(MHC 3loci) + Ho(STR) + Avg. r(MHC) + Avg. r(STR) + Year first reproduction				
Coefficients:	Estimate	Std.Error	t value	Pr(> t)	Significance
(Intercept)	1417.5	571.0608	2.482	0.017469	*
Ho(MHC 3loci)	4.0446	1.6271	2.486	0.017321	*
Ho(STR)	9.6994	4.7111	2.059	0.046235	*
Avg. r(MHC)	3.2038	0.9562	3.35	0.0018	**
Avg. r(STR)	-10.5944	2.9251	-3.622	0.000832	***
Year first reproduction	-0.7081	0.2851	-2.484	0.01741	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
	value	d.f.	p		
F-statistic	5.684	5 and 39	0.000497		
Residual standard error	3.961	39			
Multiple R-squared	0.4215				
Adjusted R-squared	0.3474				

Table 19: Linear Model best explaining the total offspring (TO) for the breeding wolves, based on the values of heterozygosity (Ho) and relatedness (r) at the background 12 STR and at the 3 MHC *loci*, plus the year in which the first breeding occurred. When a wolf reproduced with more than one mate, the average value of r and year was considered. (d.f.=degrees of freedom)

Even in this case, there is a negative correlation between the fitness trait and the STR relatedness between mates, and a positive one in the other cases (Fig. 26). Also the year in which the first reproduction occurred, used as a temporal indicator, has a significant effect, with TO decreasing with the time.

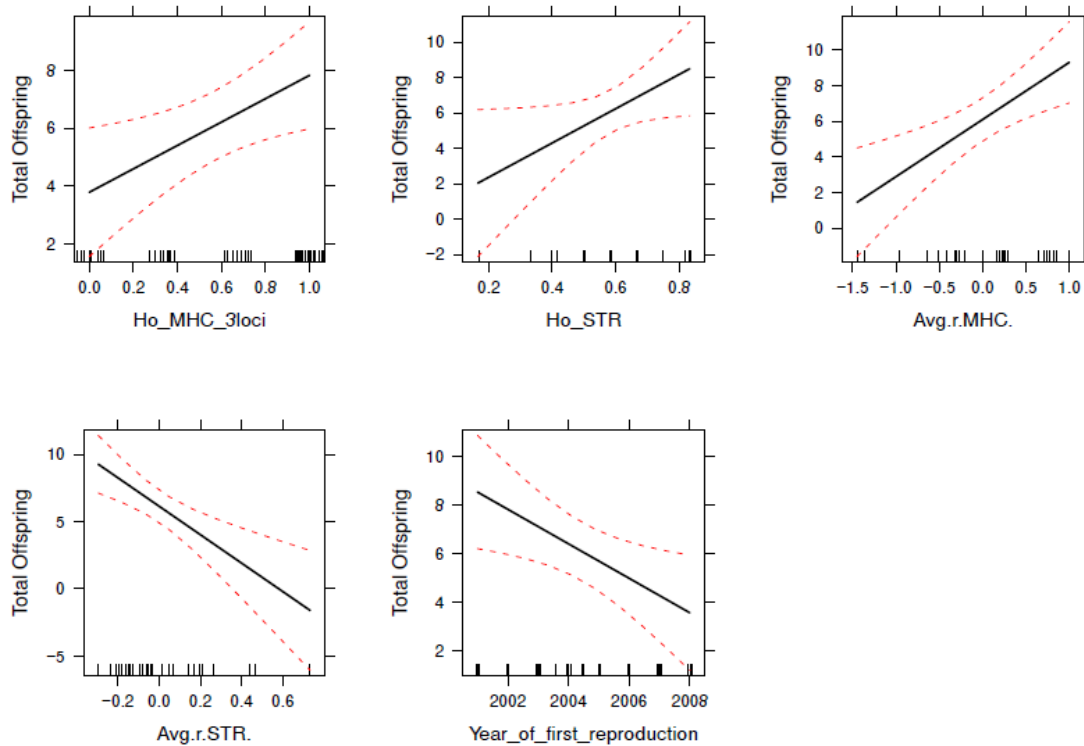


Figure 26: Effect Plot of the correlation between the number total offspring (TO) for the breeding wolves and the five most significant explanatory parameters: the heterozygosity (Ho) at both MHC and STR, the relatedness (r) at the 3 MHC and at the background 12 STR *loci*, and the year of first reproduction. The first three factors show a positive correlation with the total offspring, contrary to the latter two.

The last relevant model (Tab. 20) better explained the litter size per year (LS) on the basis of the heterozygosity levels of the mates, both at MHC haplotypes and in the background, and both with a positive contribution (Fig. 27).

Conversely, we were not able to find statistical support ($p > 0.2$) for the models explaining the sampling time (SO), although the same genetic features were partially correlated with it: positively for Ho(MHC), Ho(STR) and r (MHC), negatively for r (STR) (data not shown).

Linear Model		Litter Size ~ Ho(STR) + Ho(MHC haplotypes)				
Coefficients:	Estimate	Std.Error	t value	Pr(> t)	Significance	
(Intercept)	-0.225	0.8353	-0.269	0.789		
Ho(STR)	3.4466	1.3111	2.629	0.0119	*	
Ho(MHC haplotypes)	1.0365	0.3916	2.647	0.0114	*	

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ''	1
	value	d.f.	p			
F-statistic:	7.041	2 and 42	0.002306			
Residual standard error	1.129	42				
Multiple R-squared	0.2511					
Adjusted R-squared	0.2154					

Table 20: Linear Model best explaining the litter size per year (LS) for the breeding wolves, based on the values of heterozygosity (Ho) at the background 12 STR and at the 3 MHC *loci*. (d.f.=degrees of freedom)

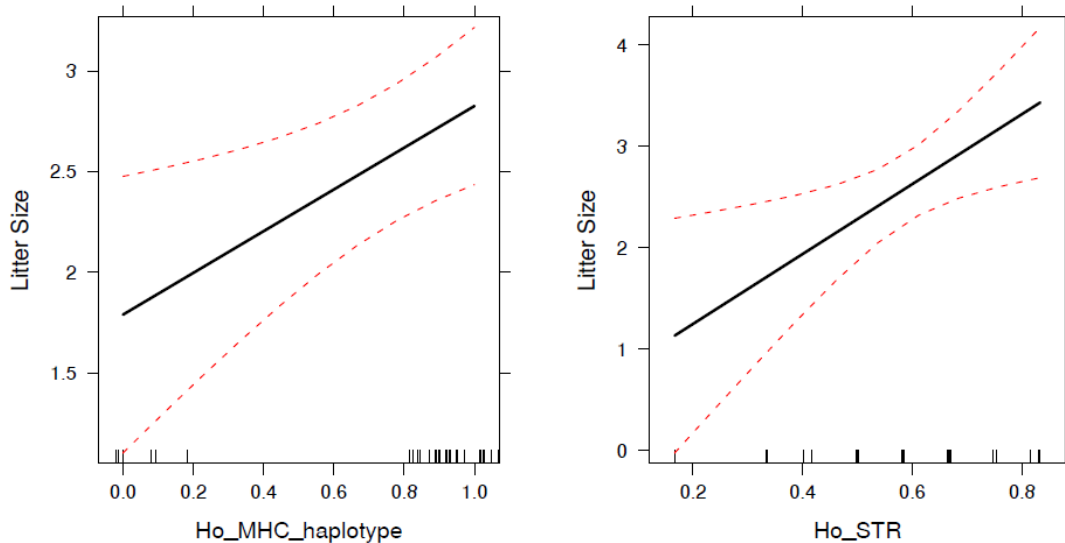


Figure 27: Effect Plot of the correlation between the average litter size (LS) for the breeding wolves and the two significant explanatory parameters: the heterozygosity (Ho) at MHC and at background STR, both positively correlated with the fitness trait.

2.5. Discussion and implications

Although thoroughly studied, the role of the Major Histocompatibility Complex in the way vertebrate species cope with natural selection is still to be univocally addressed (Bernatchez and Landry 2003, Sutton et al. 2011).

Several studies revealed its importance in the response of hosts to pathogens (Dionne et al. 2007), in maintaining genetic variability in otherwise monomorphic *taxa* (Aguilar et al. 2004), in influencing the mating scheme of mammals (Setchell et al. 2010), birds (Griggio et al. 2011) and teleosts (Evans et al. 2011), although with sometimes contrasting results, especially in humans (Havlicek and Roberts 2009).

Surely, the study of the MHC in conservation genetics is of primary importance (Radwan et al. 2010), especially in bottlenecked species or in those that are threatened with extinction.

In the present study, we investigated the variability of three MHC class II *loci* (DRB1, DQB1, and DQA1) in the Italian wolf population, which has been affected by a long-term isolation and bottleneck (Lucchini et al. 2004), reaching a concerning low population size. Nonetheless, it has been recently expanding at a fast-growing pace, re-colonizing many areas of the former distribution range (Fabbri et al. 2007). However, hybridization with feral dogs has been repeatedly documented (Verardi et al. 2006), although probably affecting a limited portion of the population.

Our study shows that a good level of variability at the MHC genes has been retained, respectively showing 9, 7 and 9 alleles at DRB1, DQA1 and DQB1, combined into 20 *multilocus* haplotypes and representing more than 50% of the alleles described in the overall European or North American wolf populations (Seddon and Ellegren 2002). As a comparison, the highly endangered Mexican wolf population only shows 5 DRB1 alleles (Hedrick et al. 2000), whereas the Swedish population, which likely originated from a very limited (<5) number of founders (Seddon and Ellegren 2004), shows 5, 4, and 4 alleles at the same three *loci*, respectively.

Two DRB1 alleles were described here for the first time, but most of them were shared across canids, compatibly with the well known MHC trans-species polymorphism, as emerges from the phylogenetic trees of the three *loci*.

Compared to 12 microsatellite markers, the heterozygosity levels at the MHC are similar or even higher. These results are coherent with previous studies (Aguilar et al. 2004), but not with evolutionary models that have been recently proposed for bottlenecked populations (Ejmond and Radwan 2011). However, when we performed a Ewen-Watterson test of heterozygosity, the observed values were higher than expected for STRs, although their

significance changed according to the used model (TPM vs. IAM). A significant excess would be predictable in the case of a recent population bottleneck, as described for the Italian wolf (Lucchini et al. 2004). However, the same excess was not found at the MHC *loci*. If confirmed, this pattern would suggest that during the bottleneck the heterozygosity levels were similarly affected at the neutral and functional *loci*, but that a higher proportion of alleles was retained in the case of MHC, possibly due to positive or balancing selection.

The difference in the heterozygosity levels between MHC and STR *loci* is more marked in the individuals showing traces of having an admixed wolf*dog origin, indicating a higher level of differentiation at the functional *loci* between the two source populations. Coherently, private alleles have been found in the admixed group, whereas the wolves showing atypical phenotypic traits, such as the dark coat color (Anderson et al. 2009, Randi 2011), only showed alleles common to both the other groups, and none of the alleles private to the admixed wolves. This confirms the assignment based on the neutral *loci*, although a limited number of markers can be inefficient in detecting past hybridization events or gene introgression. However, MHC sequencing did not add significant power to the detection of such events, like in the case of the Maremma Regional Park canids (Caniglia et al. submitted), where the described pack has been founded by third- or fourth-generation admixed wolf*dog individuals, also carrying the black mutation at the *K-locus*: in the present study, they only showed some of the most common alleles of the wild-type wolves, suggesting that the dog contribution to the MHC genes have not been retained, or simply can not be detected given the high level of allele sharing between the two groups.

Investigating the geographic distribution of the allele richness across the Italian peninsula, we can see a pattern of reduced variability at the MHC in the Alpine group, only showing three haplotypes compared with 5 to 7 in the Apennine groups. This is compatible with the recent colonization of the Alps by a limited number of founders (Fabbri et al. 2007), although the amount of analyzed samples is too small to perform a comprehensive model of the colonization by comparing functional and background variability.

The number of haplotypes detected in the Northern Apennine doubled when increasing seven times the sample size, but only adding rare variants and a single additional allele at DQA1 and DQB1, suggesting that most of the variation, at least in terms of alleles, has been sampled.

Overall the population, the *loci* were not deviating from HWE. A single site at DRB1 showed a significant departure from it in wild-type wolves, at position 60. Interestingly, this nucleotide corresponds to the single synonymous mutation differentiating one of the two

newly described alleles from its closest sequence (DRB1*09201 allele), possibly suggesting a recent mutation in the derived state that has not yet reached the equilibrium. The other newly described allele, on the contrary, seems to be basal to its closest neighbor, therefore suggesting a more ancient origin.

The topology of the phylogenetic trees did not show any clustering of the alleles found in the Italian wolf population, which are spread throughout the branches. This is compatible with a general pattern of trans-species polymorphism described for all class-II MHC *loci* (van Oosterhout 2009, Seddon and Ellegren 2002).

The high values in the d_N/d_S ratio are a clear trace of strong historical selection on the MHC, although the departure from neutrality (as from computing Tajima's D in sliding windows) is concentrated in specific portions of the exons, hosting several amino acid position known to act as peptide binding sites (Hedrick et al. 2002).

Contrary to our expectations, the results based on the careful pedigree reconstruction of 19 packs and 26 breeding pairs showed no evidence of MHC-based disassortative mate choice. Conversely, we found traces of an assortative mating behavior by which the reproductive wolves tend to choose mates who share one or both alleles at each MHC *locus*, and generally showing a peptide similarity higher than expected under a random mating scheme. This pattern was also confirmed by looking at trios of alternative partners, where the actual mates turned out to be more similar than alternative, unrelated wolves belonging to the same pack. However, we did not find any significant bias when looking at the background relatedness. This is not completely surprising, since also Geffen et al. (2011), studying the relationship between mate relatedness at neutral *loci* and the probability of kin encounters in four wolf populations, did not find evidence of any inbreeding avoidance strategy, except within natal groups.

We did not find evidence of mating-up processes as in Griggio et al. (2011), nor of higher levels of MHC heterozygosity in breeding vs. non-breeding individuals, as in Thoss et al. (2011). These findings are also in potential contrast with the high levels of variability displayed by the analyzed *loci*.

However, things get clearer when considering the effects of MHC on a panel of fitness traits in the breeding wolves, as deduced from the pedigrees. The number of reproductions, that implies getting and remaining at the top-rank in a given pack, is proportionally correlated to the relatedness of the mates at the MHC, confirming a positive influence of an assortative mating scheme also on this fitness trait. However, this was inversely correlated with the background relatedness of the mates at the neutral *loci*, suggesting that the relationship at the

MHC can be not representative of the general genetic differences, and that a general inbreeding avoidance at other *loci* can be rewarded, even though not actively chosen.

Once the wolves get to reproduce, the total offspring they produce seems to be associated in the same way to the relatedness between mates, but a stronger, positive effect is given by the levels of heterozygosity of the mates, both at the MHC and STRs, confirming results from previous studies (Setchell and Huchard 2010, Thoss et al. 2011). Similarly, the heterozygosity levels are also positively correlated to the average litter size produced per year, but in this case there is no evidence of any direct effect from the relatedness between mates.

Therefore, the genetic combination that maximizes the fitness traits seems to be: 1) being related at the MHC; 2) having a dissimilar genetic background; 3) having high heterozygosity levels both at MHC and STRs.

Consequently, we can deduce that the diversity at the MHC is maintained not by a disassortative mating preference, but rather by a relevant advantage of the heterozygote. However, this only matters once the wolves get to find a mate, since the heterozygosity levels of the breeders are not higher than the ones showed by unrelated, non-breeding individuals - they are actually lower. This leads us to deduce that an intermediate level of heterozygosity (Eizaguirre et al. 2009, Nowak 1992) at the MHC can 1) at first, enhance the probability of finding a partner with a similar MHC panel, compared to extremely high levels of heterozygosity; 2) later, be rewarded by a higher fecundity, compared to lower levels of heterozygosity.

In addition, the need of a disassortative mating scheme based on MHC may not be needed (Jamieson et al. 2009) in an expanding population such as the Italian wolf, where the kin encounter rate outside the natal pack is probably relatively low (Geffen et al. 2011), or can be substituted by other behaviors commonly adopted by wolves, such as long-distance dispersals,

However, the explanation for the benefits to the mates of being more similar at the MHC still needs to be found. Contrary to our findings, several studies indicate a direct reward to a disassortative mating scheme (*i.e.* Setchell et al. 2010), which could start from a lower incidence of abortions (Berger et al. 2010), but also leading to indirect advantages related to an increased offspring heterozygosity, allowing a wider response to pathogens (Doherty and Zinkernagel 1975).

However, Lewis (1998) argued that the preferential association for MHC-similar individuals would decrease the probability of infection with unfamiliar pathogens, and could also be the main mechanism that led to the evolution of kin altruism. If this is true, reproducing with

MHC-similar mates without increasing the general inbreeding levels would be beneficial, and there is little doubt about the highly development of altruism in wolf societies (Geffen et al. 1996).

The real answer, therefore, can probably rely on the almost unique (at least among the species so far investigated) social structure that characterizes wolves: the pack. The pack is the unit that determines the distribution of wolves in the space. It is mainly composed of a familial group, plus additional unrelated wolves, the adoptees, although the pack size can greatly vary with the latitude and the prey dimension and availability. Within a pack, the individuals actively cooperate to hunt, defend their territory and raise the pups. However, a stringent hierarchy maintained by the most-fit, leading individuals allows a single pair of non-related mates to reproduce each year, with few exceptions (Caniglia et al. submitted, Vonholdt et al. 2008). These two factors, subsequential reproduction and territory defense, could partially explain an assortative mating scheme in wolves.

- 1) Since each year only the most-fit individuals get to the higher rank in the pack and reproduce, this allows for a constant adaptation to the environment (both to its resources and to its pathogens). If a given wolf gets to the top rank thanks to its combination of genes, which allowed him to be the most fit in that particular moment in that environment (beside the non-genetic components of the fitness), he can probably benefit by mating with an individual carrying similar functional genes – that is, similarly adapted and best-fit, especially in functionally essential clusters such as the MHC. Therefore, selection could have favored individuals seeking for mates sharing a larger proportion of MHC alleles, since they have already proved to be ‘good genes’ in getting to the higher rank in a pack. We can call this “Top-ranked Allele Sharing” hypothesis, which would provide a tool to constantly keep-up with local environmental changes, including pathogens (Penn and Potts 1999).

However, such an underlying heterogeneity, required to sustain a fluctuating selection hypothesis (Hill 1991, Spurgin and Richardson 2010), is difficult to prove, since most of the territory occupied by a population, at least at a low-scale (hundreds of kilometers), can largely share similar preys, and similar pathogens. Therefore, a second hypothesis could better explain the data:

- 2) Given the highly territorial distribution of packs, wolves actively mark the boundaries and the most important portions of the home ranges, in order to delimitate the borders of the territory exploited by each single pack. Intraspecific strikes commonly occur when those limits are broken by an adjacent pack, often leading to cruel fights that can

represent a relevant cause of mortality among wolves (Ciucci et al. 2007, Lovari et al. 2007). The MHC is well known to affect the discrimination of body odors, and its composition can be easily detected even by quasi-anosmic species such as humans. Therefore, we hypothesize that a breeding pair of wolves can be more effective in marking its territory when sharing a common panel of MHC peptides: given the extreme number of possible allelic combinations, it could represent a specific ‘pack-fingerprint’, producing a univocal signal that is more strongly recognized by adjacent packs, as well as by dispersing wolves. This could lower the need of actively fight for defending the pack’s territory, therefore allowing more energy to be dedicated to food retrieval, to reproduction and to offspring raise. We can define this as the “Marking-homogeneity”, or “Pack MHC fingerprinting”, hypothesis.

However, both models have some implicit limitations, and more research is needed to confirm or contradict these hypotheses.

First of all, by expanding the sample size, although given the elusiveness and low density of wolves in the wild this is a difficult task to accomplish and can require years of work of tens of people on the field.

Secondly, we need to confirm the fitness estimates based on the reconstruction of pedigrees by genetic sampling with stronger, long-term observational data, as well as in other populations, possibly including homogeneous data on parasitic exposure; however, this would only be possible in a very limited number of cases where both genetic and extensive tracking or trapping data are available, such as for the wolves in Yellowstone National Park.

Thirdly, by designing MHC-linked microsatellite, which would allow for a much cheaper genotyping of a larger number of genes in the MHC cluster, and to look for any specific influence of Class I, II, and III genes on fitness traits and mating choice. (Aside of this study, a panel of *ca.* 15 primers to amplify MHC-linked microsatellites has been designed, and is going to be tested shortly afterwards).

Fourthly, the relation between MHC and Olfactory Receptor (OR) genes, which can act as potential mediators of reciprocal recognition (Spehr et al. 2006, Ziegler et al. 2010), should be better investigated. Unfortunately, no olfactory genes have been annotated so far on dog and wolf chromosome 12 -the one hosting the main MHC cluster in canids. However, a series of MHC genes have been recently described in the telomeric end of chromosome 35 (Santos et al. 2010), as the residual portion of a chromosomal split that occurred before the divergence between canids and felids. In this region, also a number of functional OR genes is present, suggesting a potential linkage between the sequences coding for olfaction recognition and for

immune functions. The study of their interaction could disentangle potential confounding effects and spread a clearer light on the role played by both MHC and olfaction in the mating choice.

More generally, a better understanding of the molecular mechanisms of reciprocal recognition through MHC genes, especially between individuals (Ziegler et al. 2010), could greatly improve our ability to link genetic data to the observations on reproductive and social behavior of vertebrates, and to improve our hypotheses.

In the end, this study represents a small but relevant step in the knowledge of the genetic variability of the protected population of the Italian wolf, hopefully leading to better conservation strategies with special regard to its expansion in areas of re-colonization and to the introgression of genes from dogs at a functional level.

Beside this, although with the partial support given by our limited sample size, we hope that this study can raise the attention toward a better understanding of the importance of the MHC in the social life of wolves in the wild, and in their adaptation to the environment; seemingly, with an effect on both mate choice and fitness: Choosy wolves, assortative mating and the heterozygote advantage?

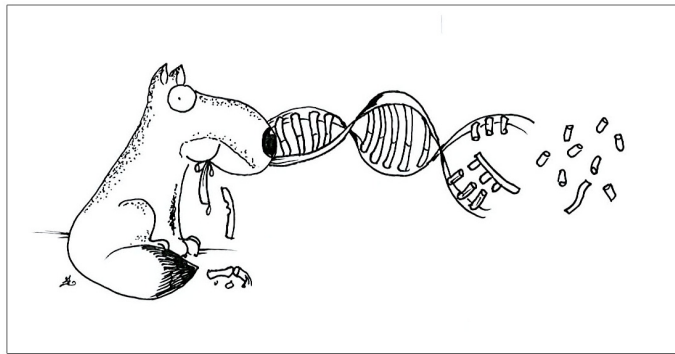
3. The first wolf genome project

What to do with a genome?

Well, this is not a question that people use to ask themselves everyday, but its importance is quickly growing, as fast as the current technological and scientific advance we had an overview on in the first chapter.

However, this question also implies the great opportunity of dealing with the billion-nucleotide sequences coming out from a next-generation sequencing project, such as for a non-model organism like the wolf. Of course it is just a matter of choice, so big is the amount of biological information derivable from it.

Therefore, we will try to figure out what interesting information can be obtained (and how) from the first complete genome draft of *Canis lupus*, especially in relation to the already available dog genome assembly (Lindblad-Toh et al. 2005) and to one of the most interesting topics that can be addressed by this comparison: the mechanisms and effects of dog domestication.



Edoardo Velli: 'The wolf genome' (pencil drawing on paper)

IMPORTANT NOTE:

All this study has been performed at University of California, Los Angeles (UCLA), Department of Ecology and Evolutionary Biology, under the supervision of Prof. R. K. Wayne and of J. Novembre (Principal Investigator), and A.H. Freedman as main author.

Only the general background of the project, and the analytical sections that I collaborated to, will be included in this thesis.

For these, the work has been performed in strict collaboration with R.M. Schweizer, P.M. Silva, and Z.X. Fan, as well -for the splice sites analysis- with M.Roy.

The project is currently in progress; therefore all the methods and results presented in this chapter are partial and subject to changes. They can only be cited and reproduced upon completion from future official publications, or by specific request to the authors.

3.1. Background

3.1.1. Overview on whole-genome studies

Somebody considers it a race. For some others, it is just scientific advance.

However, contrary to the public idea, the genomic age already started some dozens years ago (Turner et al. 2009), with the complete genome of a phage to be sequenced.

But it took decades (and a lot of technological advance) to pass from the small virus genomes to bacteria and then vertebrates.

Among them, the human genome was the first one to be completed, together with the rat, in 2003, shortly followed by the mouse.

In the last few years, also thanks to Next Generation Sequencing and the relative drop in the sequencing costs, the number of complete genome drafts blew up, reaching the number of 67 only considering the animal kingdom (today available via UCSC Genome Browser):

- 23 mammals: 7 primates (human, chimp *Pan troglodytes*, gorilla *Gorilla gorilla*, orangutan *Pongo pygmaeus*, gibbon *Nomascus leucogenys*, macaque *Macaca mulatta*, marmoset *Callithrix jacchus*), 3 rodents (mouse *Mus musculus*, rat *Rattus norvegicus*, guinea pig *Cavia porcellus*), 7 domesticated species (dog, cat, cow, horse, rabbit, sheep, pig), plus the panda *Ailuropoda melanoleuca*, the microbat *Myotis lucifugus*, the elephant *Loxodonta africana*, two marsupials (wallaby *Macropus eugenii* and opossum *Monodelphis domestica*) and a monotreme (platypus *Ornithorhynchus anatinus*);
- 3 birds: Chicken *Gallus gallus*, turkey *Meleagris gallopavo*, zebra finch *Taeniopygia guttata*;
- 1 reptile: anole lizard *Anolis carolinensis*;
- 2 amphibians: western clawed frog *Xenopus tropicalis*;
- 4 fish: puffer *Tetraodon nigroviridis*, zebrafish *Danio rerio*, fugu *Takifugu rubripes*, Stickleback *Gasterosteus aculeatus*, medaka *Oryzias latipes*;
- 1 Petromyzontide: lamprey *Petromyzon marinus*;
- 1 Cephalochordate: lancelet *Branchiostoma floridae*;
- 1 tunicate: sea squirt *Ciona intestinalis*;
- 1 echinoderm: sea urchin *Strongylocentrotus purpuratus*;
- 13 insects (including 11 *Drosophila* species, bee *Apis mellifera*, and *Anopheles gambiae*);
- 1 mollusc: sea hare *Aplysia californica*;
- 6 nematodes (including 5 *Caenorhabditis* spp.);

As we see, this list is clearly mammal-biased, but it is already obvious that, in the next few years, most of the animal *phyla* are likely to be investigated. And, beside catalogue purposes, every genome project contributes with new analytical tools for the scientific community. Of course, almost all the projects are exploiting the opportunities given by NGS, and some of the most recent ones to be concluded include panda, lizard, cod and orangutan genomes: let us have an overview on their methods and findings.

The panda is a seriously endangered species, with less than 2,500 individuals surviving in the wild. Its phylogenetic position within the mammal class has always been controversial, since it shows unique features even among Ursidae, like its highly-specialized diet based on bamboo. A *de novo* assembly of short read data was completed in 2010 (Li et al. 2010), providing a constantly high coverage (>20x). Compared to human and dog, the panda genome shows a lower divergence rate and a smaller proportion of recent segmental duplications. The identified positively-selected genes were mostly deputed to immunity functions. The pseudogenization of one of the five main taste receptors could partly explain the dietary differences with other Ursidae, although retaining the functional genes associated with a carnivorous diet. However, despite the limited current population size, the levels of heterozygosity turned out to be high, showing about 2,7M heterozygous SNVs in the diploid genome.

Also the recent 40x assembly (Star et al. 2011) of the Atlantic cod (*Gadus morhua*) genome revealed some unexpected features. In fact, genes from the MHC II pathway (usually highly conserved throughout vertebrates) seem to be completely lacking. A possible compensatory role can be played by a suite of Toll-like receptor genes and a unique -even among teleosts- expansion of MHC I genes.

The first reptile whose genome was completed (Alföldi et al. 2011) is the green anole lizard (*Anolis carolinensis*). Its unique (highly constant GC content) and shared features (synteny of microchromosome with some of the chicken ones) allowed to better resolve the phylogeny of amniotes. Moreover, it has been shown to possess genetically-determining sex chromosomes, although with mechanisms different from both birds and mammals. Interestingly, it has been also proven that egg-related proteins evolved at a faster rate compared to the other, suggesting important roles in the developments of amniotes.

A Sumatran Orang-Utan (*Pongo abelii*) genome assembly was recently completed (Locke et al. 2011), sided by a number of short reads from five other Sumatran and five Bornean (*Pongo pygmaeus*) individuals. The Orang-Utan genome showed a slower evolution rate compared to other primates, with almost inactive Alu repeats and fewer rearrangements

and duplications. Signals of selection on lipid metabolism pathways could be linked to the uniquely low energy usage of the species. Despite current population sizes, the two species of Orang-Utan show opposite trends in effective population history to what expected.

But NGS methods have also been applied to the study of genomes from extinct species or lineages, such as the Neandertal (Burbano et al. 2010, Green et al. 2010). The challenges of working at a genome-wide scale with paleontological samples dating back some ten thousand years are remarkable, especially when considering that modern human contaminations are hardly distinguishable from the target DNA, which was only between 1% and 5% of the total DNA contained in the samples. Nonetheless, by applying stringent procedures and NGS techniques, the authors managed to extract, enrich and assemble more than 4Gb of sequences from three Neandertal individuals. The average divergence from humans has been computed, although Neandertal falls into the variation of modern humans for most of the genomic regions. Looking at functional sites where the modern humans have a unique derived allele compared to Neandertal and chimpanzee, only 78 non-synonymous substitutions appeared to be fixed and different in humans, several of them falling into genes expressed in the skin. Also when looking at human-accelerated regions, most of the variants were common to both humans and Neandertal; only in a few cases they were unique to modern humans, therefore indicating interesting regions of human-specific selection, and the genes there included. An example is *RUNX2*, which is likely to affect the skull and chest morphology, two of the most discriminating traits between modern humans and Neandertals or other archaic hominins. The population divergence time between Neandertal and human was inferred to date back between 270,000 and 440,000 years ago, with a higher similarity of Neandertals to Eurasian than to African humans, and an evidence for gene flow that occurred between Neandertals and non-Africans ancestors: this could indicate a non negligible, but limited (1-4%) contribution to the genome of some present day populations by Neandertals, although also being compatible with an ancient population structure in the African clade.

Of course, in order to maximally exploit the large amount of data coming from a single individual, new methods have been developed to support the analyses.

To detect older selective sweeps, in the same paper Green et al. (2010) developed a method consisting in: 1) identify ancestral and derived alleles (e.g. in humans and Neandertals) compared to an older ancestor (e.g. the chimpanzee); 2) identify which sites carry high-frequency derived alleles in the reference genome (e.g. humans); 3) compute the expected number of sites showing a derived allele in the target genome (i.e. Neandertal), based on the assumption that for most sites ‘the variation within current humans is old enough to include

Neandertals; 4) identify those regions where the target genome is devoid of derived alleles. This would indicate the presence of a human-specific advantageous mutation, which arose and swept causing its frequency to increase or get to fixation, allowing to identify the genes likely to have been targeted by selection.

Similarly, to identify gene flow from Neandertals to some of the human populations, they hypothesized that regions of Neandertal derivation should show a lower divergence between the two (when considering haplotype data), but a higher divergence between human sequences with Neandertal and non-Neandertal derivation. This would also allow discriminating actual gene flow from regions with a naturally-low mutation rate, since the latter would lead to a limited divergence among human lineages as well.

But the availability of single, whole genome sequences (rather than multiple but single-*locus* sequences) was also exploited in order to infer demographic parameters.

That is the case for the method proposed by Li and Durbin (2011), namely the pairwise sequentially Markovian coalescent model (PSMC). It is based on the fact that the distribution of time since the most recent common ancestor (TMRCA) between two alleles in an individual reveals information about the effective population size (N_e) through time, assuming a given generation time and a scaled mutation rate. This can be done by studying how the local density of heterozygous sites changes across the genome, then reconstructing the TMRCA distribution across the chromosomes, which is likely to reflect segments of constant TMRCA separated by historical recombination events (Fig. 28).

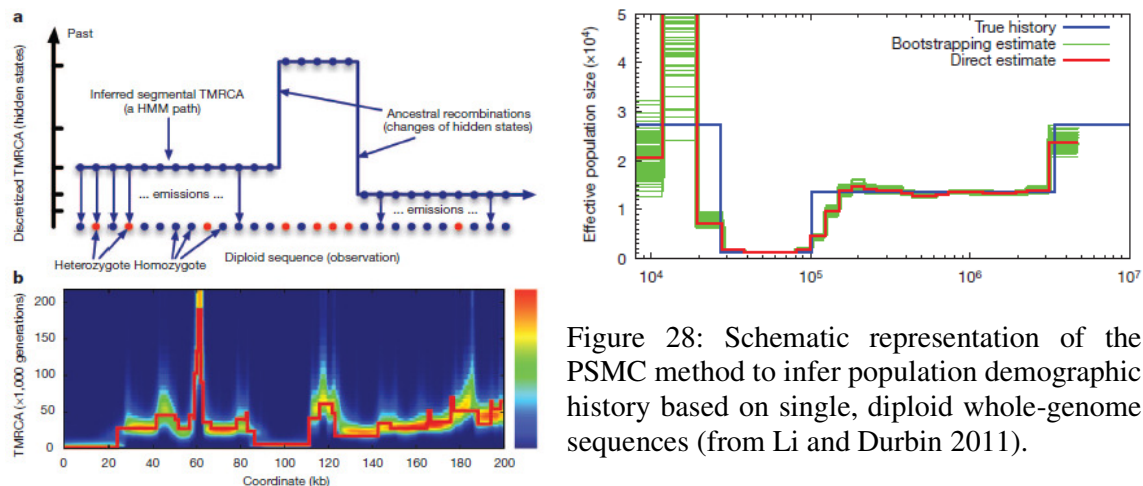


Figure 28: Schematic representation of the PSMC method to infer population demographic history based on single, diploid whole-genome sequences (from Li and Durbin 2011).

However, despite new effective methods are quickly being developed, the need for the availability super-individual information (Ouborg et al. 2010) is leading to the rapid development of population genomics, a quantitative extension of genomics where multiple

genomes from the same species are sequenced and compared in order to gain a better and representative insight to the population-level variation. Of course, this is mainly addressed at the description of the human genomic variation (The 1000 Genomes Project Consortium 2010), and the Genome 10K Project (Genome 10K Community of Scientists 2009), aiming at completing the genome sequences of about 10,000 vertebrate species, is already a reality. Moreover, metagenomics is exploiting the possibilities ensured by NGS in order to reconstruct the unknown species composition of environmental samples (Handelsman 2004), with special regard for bacterial communities.

As we saw in the case of mammals, a great attention has been dedicated to the study of domesticated species, mostly given their importance as food and labor source. However, the study of the complete genomes of their wild relatives, such as wild rice, the bison, or the bighorn, is also of primary importance for understanding the impacts of domestication and artificial selection, identifying which genes contributed the most to the large changes that occurred in a relatively short time, and which ones can be related to diseases and pathologies affecting the domesticated species, but not their wild relatives.

In the next paragraphs, we will see in details what insights we gained from the study of the most ancient species to be domesticated, the dog, and how we will benefit from the comparison with its ancestor, the wolf.

3.1.2. The dog genome

The dog genome was one of the earliest to be sequenced.

A first 1.5x draft was completed in 2003 by Kirkness et al., followed by a 7.5x assembly in 2005 (Lindblad-Toh et al. 2005), both performed by whole-genome shotgun (WGS).

Beside being ‘the man’s best friend’, from an evolutionary perspective (Lindblad-Toh et al. 2005, Wayne and Ostrander 2007) the dog genome was important to understand the phylogenetic relationship with human and mouse, to identify which genomic features are specific to each lineage, and which ones are common to most mammals.

Moreover (Galibert and André 2008), the huge phenotypic variation showed by dog breeds is almost unique in the animal kingdom, but its genetic basis still needs to be uncovered. Additionally, dogs and humans share a large panel of diseases (Boyko 2011), and the availability of the complete dog genome provided the unique opportunity to identify the genes related to common pathologies, thanks to subsequent GWAS based on the polymorphic sites identified in the assembly.

The dog genome (Lindblad-Toh et al. 2005) is organized in 38 autosomes, plus X and Y sexual chromosomes. It is composed of *ca.* 2.41 Gb (billion bases), being therefore smaller than in human (2.9 Gb) and mouse (2.5 Gb).

It shows lineage-specific transposable elements (notably, a highly active carnivore-specific SINE family, SINEC_Cf), although the total amount of repeated regions is lower, partially explaining the smaller genome size. The average G+C content is about 41%, similar to human. The relative rate of nucleotide divergence is higher than in human, but lower than in mouse, possibly reflecting the different generation times. Both divergence rates and G+C content are higher in proximity of telomeres. About 5% of the genome shows a high conservation relative to mouse and humans, suggesting that genetic regions other than coding elements (which account for *ca.* 1% of the genome) constitute conserved functional elements; they are probably related to the regulation of gene expression, and enriched in proximity of genes responsible for development.

The total number of genes predicted in dog is about 19,300, lower than the *ca.* 22,000 identified in human. The level of gene duplication was also showed to be lower than in human, although with some expanded dog-specific families. The selective pressure on the genome appears to be intermediate between human and mouse, coherently with their population sizes, with a few genes showing dog-accelerated evolution, which are mainly related to metabolic functions.

About 2.5 million SNPs were identified, both within the sequenced female boxer (Tasha, 770,000) and between Tasha and the previously sequenced Poodle (1.46 millions), with the remnant ones being identified against a subset of shotgun sequences from other nine dog breeds, four wolves, and one coyote, therefore leading to a general estimate of ~ 1 SNP every 900 bp.

This large amount of variation led to resolve in details the phylogeny within canids (Fig. 2, chapter 1), and subsequently to design SNP chips to perform extended investigations at the population level, such as in the CanMap project (Boyko et al. 2010).

One of its applications, based on data about haplotype and allele sharing, was to better resolve the phylogeny among dog breeds (Vonholdt et al. 2010), in relation to their ancestor, the wolf, showing that most of the morphological and functional clusterization is matched, although with few significant exceptions, such as for the Pekingese (Fig. 29).

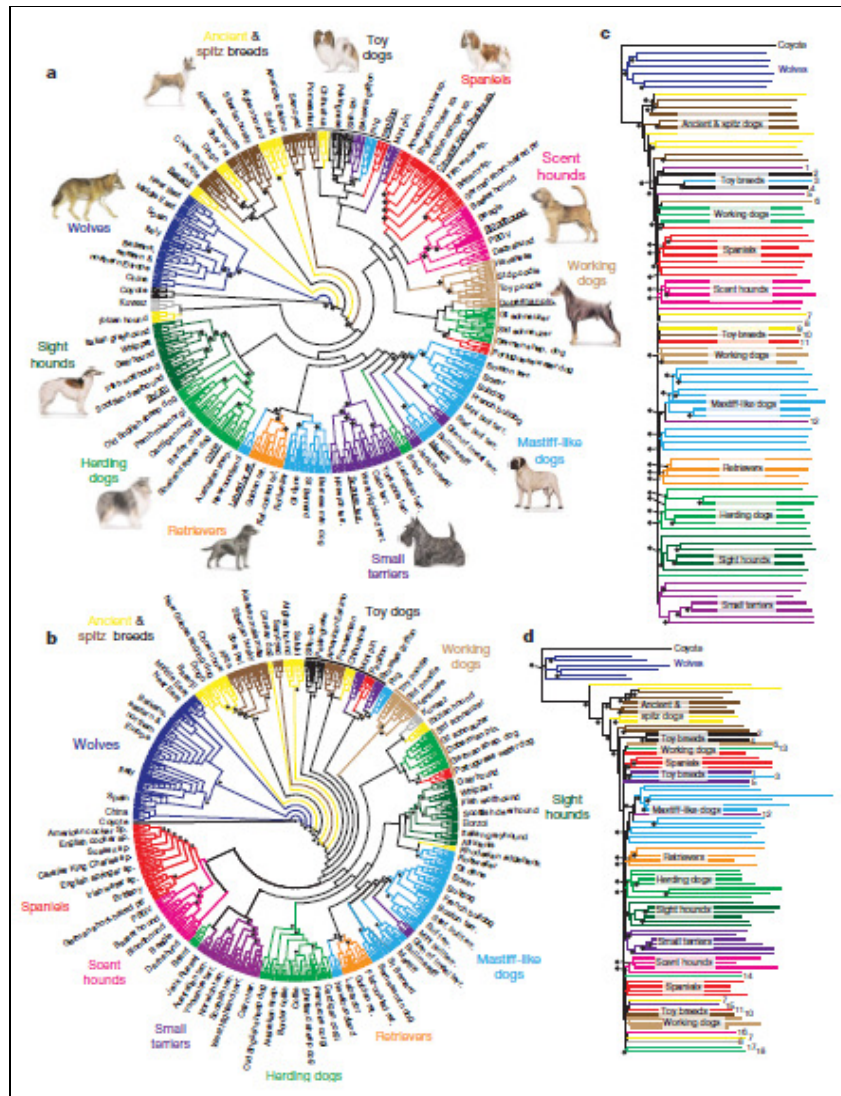


Figure 29: Neighbour-Joining trees of domestic dogs and gray wolves based on haplotype (a-c) and allele (b-d) sharing. From Vonholdt et al. (2010)

Similarly, it was possible to resolve the population differentiation among wolves and coyotes, and to clarify the origin of admixed wolf/coyote populations, such as the Great Lakes and the Red wolves (Vonholdt et al. 2011).

Following studies deepened the knowledge about structural (Chen et al. 2009) and copy-number variation in the dog genome (Nicholas et al. 2011, 2009), as well as about linkage disequilibrium (Wong et al. 2010) and selective sweeps (Quilez et al. 2011).

From a biomedical point of view, the availability of the dog genome allowed to identify the genetic basis of several diseases (Boyko 2011), such as meningoencephalitis (Barber et al. 2011); other 100 genetic diseases, half of which common with humans (reviewed in Giger et al. 2006, Karlsson and Lindblad-Toh 2008) have been identified to date.

Moreover, several phenotypic traits have been linked to their genetic determinants: coat color (Candille et al. 2007, Dreger and Schmutz 2011, Karlsson et al. 2007) and shape (Cadieu et al. 2009), dorsal ridge in the fur (Salmon Hillbertz et al. 2007), hairlessness (Drögemüller et al. 2009) and wrinkled skin (Akey et al. 2010), as well as leg length (Parker et al. 2009), body size (Boyko et al. 2010, Sutter et al. 2007, Vaysse et al. 2011) and shape (Boyko et al. 2010), which are some of the traits showing the widest variation among breeds, but also tail (Vaysse et al. 2011) and ear morphology (Boyko et al. 2010, Vaysse et al. 2011), and even behavioral traits such as sociability (Vaysse et al. 2011).

However, despite the huge advance in our knowledge about dog genetics and evolution, some unresolved questions still remain: how did domestication occur? In which continent? Did it happen only once, or was it a multiple process? What are the genes mainly affected by it?

Although some of these issues have been addressed in specific studies, as we will see in the next paragraph, many of them are still largely unknown.

3.1.3. Dog domestication and breeding

Late Pleistocene human populations were strictly environment-dependent: agriculture and animal breeding were not yet developed, and most human groups had to rely their survival on gathering and hunting.

In these conditions, they also had to share many resources all over the Northern Hemisphere with a strong and flexible competitor: the wolf.

Most theories suggest that dogs evolved through a mutually beneficial relationship with humans, sharing – instead of fighting for – living space and food sources.

The history of the domestic dog started at least 15,000 years ago from its only wild ancestor, the wolf (Lindblad-Toh et al. 2005, Vila et al. 1997, Vonholdt et al. 2010).

For sake of exhaustiveness, we should mention that some eccentric papers still dispute this evidence (Koler-Matznick 2002), on the basis of old beliefs that were common from Darwin to Lorenz, but nowadays confuted (Boyko 2011).

Archaeological remains dating 14,000-10,000 years ago and assigned to dogs have been found throughout Europe, Near East and Russia (ref. in Boyko 2011, Germonpre et al. 2009), indicating that by the same period the population of dogs was already large and that dogs from Eurasia followed humans in their colonization of the American continents (Leonard et al. 2002).

But also the recently discovered fossil of a large canid from Goyet (Belgium), dated *ca.* 31,700 BP (Germonpre et al. 2009), revealed marked differences with wolves, and many more similarities with dogs. Therefore, it has been considered as a ‘Palaeolithic dog’, suggesting that domestication already occurred in Europe during the Aurignacian (~47,000-31,000 BP).

An even more ancient dog-like canid (Ovodov et al. 2011) has been documented in the Razboinichya Cave (Altai Mountains, southern Siberia), dating *ca.* 33,000 BP. However, the authors suggest that this lineage, probably disrupted by the climatic and cultural changes, was independent from the ones leading to post-Glacial, early-Holocene dogs.

These findings are not surprising, since mitochondrial DNA data (Vila et al. 1997) possibly anticipate the timing of the first domestication up to 100,000 years ago, although estimates based on different models (Savolainen et al. 2002) predict a range of 5,400-16,300 years ago. However, also the location where domestication occurred is still disputed.

One of the first hypotheses, based on mtDNA data, traced the center of domestication in East Asia (Savolainen et al. 2002), where the variability among dog sequences was found to be maximum, although also showing that several maternal wolf lineages contributed to the extant dog gene pool.

But a similar level of diversity was found when analyzing a large sample of African village dogs (Boyko et al. 2009), therefore questioning the evidences for an East Asian derivation.

A third hypothesis supports a Middle Eastern origin of dogs, as evinced from the study of the IGF1 gene region (the one mainly responsible for body size in canids), where the dog haplotypes are more closely related to the ones from Middle Eastern wolves (Gray et al. 2010). These findings were confirmed by a genome-wide haplotype comparison (Vonholdt et al. 2010), showing that most dog breeds are currently sharing a larger proportion of haplotypes with wolves from Middle East than with any other wolf population, with a few local exceptions indicative of subsequent crossings.

In any case, more than one domestication event could have occurred independently (Vila et al. 1997), since during most of the late Pleistocene humans and wolves coexisted over a wide geographic area, providing ample opportunity for independent domestication events, as suggested by Ovodov et al. (2011) after the findings in Altai Mountains and Goyet.

Interesting information about the history of domestication comes from the analysis of haplotypes and linkage disequilibrium (Gray et al. 2009, Lindblad-Toh et al. 2005) in the dog genome. The presence of both large (Mb-sized) and short (kb-sized) blocks suggests that domestication occurred in two main distinct phases (Fig. 30), through important bottlenecks

whose traces are still visible in the dog genome: one during the initial domestication from wild wolves (~27,000 years ago), the second one in correspondence with the breed formation, which mainly occurred as late as in the 19th century.

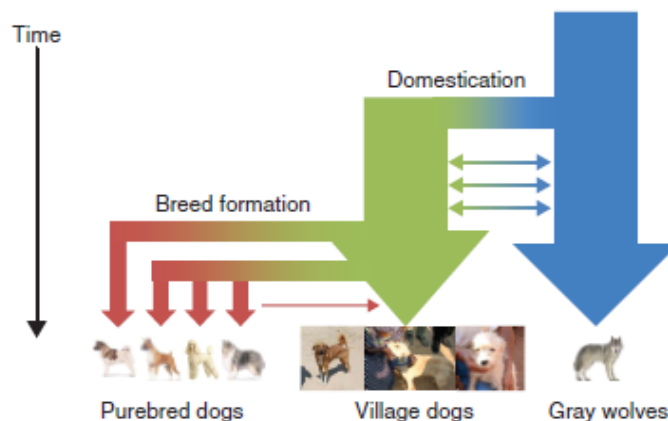


Figure 30: schematic representation of the two-step process leading to modern dogs. After the first domestication from wild wolves, a strong artificial selection was operated during the modern breed formation, especially in the 1800's (from Boyko 2011).

Since then, contrary to what happened for millennia, with semi-feral 'village' dogs only showing general domestic features (Fig. 30), humans started to strictly select dogs that specifically excelled at herding, hunting and obedience; in this process, they created breeds rich in phenotypes that both mimic human behaviors and support our needs. Dogs have also been bred for desired physical characteristics (for 'sport') such as size, skull shape, coat colour and texture, producing breeds with closely delineated morphologies. This evolutionary experiment produced diverse domestic breeds, overall harbouring more morphological diversity than what exists within the remainder of the *Canidae* family, but with reduced variation within single breeds. Therefore, the extreme phenotypic diversity of dogs, even during the early stages of domestication, clearly implies a varied genetic heritage. Meanwhile, backcrossing with wolves could have provided part of the raw material for artificial selection and for the extraordinary degree of phenotypic diversity in the domestic dog (Vila et al. 1997), although a few genes are now known to be responsible for a large variation in morphology, as we saw in the previous paragraph.

Studies on mitochondrial and Y chromosome markers (Sundqvist et al. 2006) also showed a sex bias in the origin of breeds, with the contribution of a lower number of males than females, as well as a lower genetic exchange between breeds, with paternal lineages being more differentiated than the maternal ones.

But the change or relaxation of natural selective pressures following domestication allowed the faster accumulation of likely deleterious, non-synonymous mutations in the dog genome, both at the mitochondrial (Björnerfeldt et al. 2006) and nuclear level (Cruz et al. 2008). On the other side, this would permit the faster rise of new genetic diversity, potentially underlying some of the unique features of dogs.

What is sure is that, despite the relatively recent divergence time between domestic dogs and gray wolves, the two *taxa* show remarkable morphological and behavioural differences. Morphologically (Nowak 2003), dog skull shows different proportions, being usually shorter and with a larger palate, as well as with a steeper forehead (named ‘stop’) and a different orbital angle. A vestigial fifth digit in the rear paws, named dewclaw (Galis et al. 2001) is relatively common in dog, but almost absent in wolves (Ciucci et al. 2003). Moreover, the dogs can show a large variety of coat patterns, also including spots, whereas wolves show the typically wild-type coat pattern shading from white to black. The ears in most dog breeds are droopy, contrary to the wolves, in which they are always erected. This is one of the main features generally described as ‘domestication syndrome’ (reviewed in Trut et al. 2009), which includes a series of other traits common to several domesticated species, such as tameness, curly tail, attention to visual clues, enhanced memory, and faster response to behavioral conditioning.

Impressive correspondences at morphological and behavioral traits came from the experiments on Belyaev’s silver foxes (*Vulpes vulpes*, Kukekova et al. 2011) that were artificially raised and selected for tameness: after few generations, the foxes started to show floppy ears and curly tail, and to use vocal communication to attract human attention (Gogoleva et al. 2011), as well as a higher tolerance to human presence and resistance to stress.

These traits are probably connected from a functional point of view, or are linked at the chromosomal level, and some of the genes regulating these traits are starting to be identified (Kukekova et al. 2011, Vonholdt et al. 2010).

Compared with wolves (Topal et al. 2005), dogs have a preferential looking at humans in problem-solving situations and their superior performance in using human directional gestures supports the existence of strong genetic modifications related to the domestication process in the emergence of social cognitive abilities (Hare et al. 2010).

Significant differences also emerge in reproduction: dogs can commonly reach sexual maturity at 6-12 months of age, whereas it usually occurs at 2 years in wolves (Fuller et al. 2003), although this difference can be largely due to ‘nurture’, as it happens in human

populations, since captive wolves have been recorded to breed as early as 10 months of age. In addition, dogs can reach two oestrus periods per year, in any season, instead of a single oestrus in late-winter as in wolves, highlighting significant differences in the development timing. These differences can also influence the dog generation time, that is shortened than in wolves, therefore possibly contributing to an accelerated accumulation of mutations, as shown in several studies (Björnerfeldt et al. 2006, Cruz et al. 2008).

However, to explain these notable differences that arose in a relatively short time span, Saetre et al. (2004) suggested that the two groups may mostly differ in patterns of gene expression. Comparing with microarray technology the expression patterns in three areas of the brain (hypothalamus, amygdala and frontal cortex), they identified genes with region-specific expression patterns. In wolves, like in *C. latrans*, the hypothalamus showed a highly conserved expression profile, contrary to domestic dog, suggesting that selection on dogs for behavioural traits may have resulted in modifications of mRNA expression patterns in a few hypothalamic genes with multiple functions. This could indicate that rapid changes in brain gene expression may provide a mechanism for rapid adaptive changes during differentiation, particularly on behavioural characters. Similarly, significant variation in gene expression levels was also found in the pre-frontal cortex of tame vs. aggressive silver foxes (Kukekova et al. 2011). However, also hormonal regulation can play a significant role in shaping the phenotype of domesticated species (Trut et al. 2009).

Whatever the molecular mechanisms underlying dog domestication, several theories (reviewed in Boyko 2011) have been proposed to explain how it could have happened that wolves started to run with humans thousands of years ago, before any other species.

One of the first theories, or the ‘pet-keeping’ model, is the one proposed by Sir Francis Galton back in 1856, and later supported by Zeuner (1963). It hypothesizes that young or pup wolves could have been kept as pets for human companionship, some of them getting tamer and remaining with humans until adulthood. Probably, food supply by humans could have played an important role for the development of this association, later resulting in domestication. Surely, the early sensitive windows are of primary importance for the development of interactive skills (Udell and Wynne 2010), and selection for tameness by humans could have led to enhanced abilities in the comprehension of human gestures, accordingly to the ‘domestication hypothesis’ (Hare et al. 2010).

A second model, or the ‘self-domestication’, is proposed by Coppinger and Coppinger (2001). Starting from the fact that often humans create food dump areas, he hypothesized that some wolves could have been attracted and starting scavenging on the food therein available.

The access to this additional food source could have provided a selective advantage for those individuals, especially when they also got more tolerant to the human presence, progressively leading to isolation from the original wolf population and to a closer relationship with humans.

However, in our opinion, the assumption of the model might not be realistic, especially if the time of the first domestication goes back to the late Pleistocene. Hunter-gatherer human societies, besides having probably limited areas dedicated to waste disposal, took great care in using most of the animal body parts for the most diverse functions beside consumption, such as crafting of clothes and weapons. Therefore, it is not likely that the limited food waste remaining from humans could have sustained wolves for a protracted time, enough to develop stable relationships, whereas it could have greatly supported the early dog populations after the advent of agriculture and stable human settlements (Leonard et al. 2005).

A third model ('Classic domestication') was suggested by Crockford (2006): his hypothesis considers that individual wolves may vary in physiology (similarly to what happens for coat color), in particular for some important regulatory paths related to the response to stress (i.e. thyroid hormones). Therefore, a subset of wolves could have withstood anthropogenic environment better than others, allowing them to tolerate a progressive association with humans, eventually leading to domestication.

An alternative model is indicated as human-wolf co-evolution, or symbiosis (Schleidt and Shalter 2003), and it is based on the observation that humans and wolves were largely sharing their ecological niches, hunting common prey (and maybe trying to take over each other's kills). This overlap would ensure that the interactions between the two species were common, at different times and places, possibly leading to a closer relationship in which both had to gain something thanks to their different skills (e.g. superior olfaction and higher speed for wolves; more subtle vocal communication, and usage of weapons or traps for humans).

However, domestication (Topal et al. 2005) is generally thought as an evolutionary process controlled by human influence, while recent studies have suggested an unusual competence of dogs in social interactions with humans, like in cooperation, social learning and communication (although not necessarily outperforming wolves in all human-driven tasks -Udell et al. 2010, but see Hare et al. 2010), giving more support for a 'self-domestication' hypothesis.

Preliminary evidence for such a directionality of the first contacts also came up during a recent monitoring project on the Italian wolf (Galaverni et al. 2012); during the study, performed with camera traps in the territory of a known pack, we recorded an unusual

behavior of wolves who, rather than keeping a distance from human drive hunts (Ruth et al. 2003), seemed to actively follow the hunts, probably for catching wounded prey not reached by the human hunters. This would provide a clear adaptive benefit to wolves tolerating human proximity (at least, when also the contrary happened), suggesting that an early-stage contact possibly started by wolves, rather than by humans.

Probably, a mix of several models is the best explanation of the domestication process, which could have also occurred several times and in different ways.

However, with the complete genome of both dog and wolf available, we hope to better understand which genes were primarily affected by domestication, and whether more than one wolf lineage significantly contributed to the early dog genome, therefore spreading more light on these unresolved questions.

3.2. Aims

What makes a dog, a dog, and a wolf, a wolf?

This is the main question we asked ourselves at the beginning of the project.

Of course, many people are perfectly conscious of what makes their own pet unique all over the animal kingdom: full membership in the family, perfect communication skills and understanding of the owner's will, cooperation, probably also compassion and sense of morality (Bekoff and Pierce 2009).

However, without desire to hurt anyone's sensibility, these features are actually very common in most of the pets living in our neighborhoods, and wolves share many of them with their domesticated descendents.

Therefore, is there any trait that we can really find in dogs, but not in wolves, and *vice versa*?

As we saw in the previous paragraph, several features are common to most domesticated species, and a few genes have been found to be correlated with important morphological traits responsible for marked differences among dogs, and between dogs and wolves.

Nonetheless, the mentioned studies have been performed on a large number of markers, but still representing a fraction of the total genomic variability.

Therefore, by sequencing the complete genome of a wolf, we have the unprecedented opportunity to identify most of the genetic features potentially discriminating the two *taxa*, providing the basis for future population-wide studies.

But first of all, the level of genetic variability in wolves should be assessed. However, even the complete sequencing of a genome would only provide information on the intra-individual variability, failing to highlight what could represent lineage-specific differences and the actual level of differentiation between individuals and populations, which is crucial to identify wolf- vs. dog-specific features. Therefore, the samples from two different wolf individuals were included in the present study: one from Croatia, as representative of the European population, and one from Israel, since Middle East is thought to be a good candidate region for the center of domestication (Vonholdt et al. 2011).

On this basis, one of the main questions we would like to ask is whether the variability that we currently see in dogs is only a portion of the standing genetic variation hosted in the wolf genome. This pattern would be coherent with the two bottlenecks that affected the dog populations at the time of first domestication and, more recently, in correspondence with breed formation (Lindblad-Toh et al. 2005). On the other side, the relaxation of selective constraints in dogs, as we saw (Björnerfeldt et al. 2006, Cruz et al. 2008), could provide a mechanism to explain the large variability that we observe today between different dog lineages.

Modern breeds are known to have been strongly selected for specific traits, both behavioral and phenotypic; accordingly, a reduced variability within breed has already been showed, but the differences among breeds are still impressive. On the contrary, the so-called village dogs did not undergo such a stringent selection, mostly retaining general dog features and not showing specific 'sport' traits beside local variation. However, the two dog genome assemblies available to-date belong to highly selected breeds (poodle and boxer), which are known to show high levels of homozygosity.

Therefore, we also included in the present study the DNAs from two of the most ancient dog lineages in the world: dingo and basenji.

The former is the only placental mammal living in the wild in Australia; although showing most of the dog morphological features, several of its traits are intermediate between dogs and wolves. Its origin goes probably back to the time of the first human colonization of Oceania from Southeast Asia, about 5,000 years ago (Savolainen et al. 2004). Other intermediate features are the reproductive cycles, generally with a single pregnancy per year, seasonality in the estrus period and sexual maturity mostly reached only at two years of age (Jones and Stevens 1988). From a management perspective, its conservation is threatened by extensive hybridization with domestic dogs (Savolainen et al. 2004).

Beside the recent selection into a defined hound breed, basenji is an ancient African dog lineage original of the Congo basin, sharing several features with African pariah dogs (Adam R Boyko et al. 2009). Genome-wide studies (Vonholdt et al. 2011) indicate that basenji is probably one of the most ancestral dog lineages of all. Similarly to dingoes, they also show a single reproduction a year, and they typically lack extensive barking vocalizations.

These two ancient dog lineages could therefore represent optimal candidates to investigate the early steps of domestication, and to identify the genes correlated to their intermediate characters between wolves and modern bred dogs.

The detection of traces of selection is one of the subsequent aims of the study. Whereas most of the genome could still include largely shared features, specific regions linked to genes hardly selected during domestication should retain a lower diversity in dogs than in wolves, caused by consequent genetic sweeps, and a higher number of non-synonymous mutations compared to wolves.

However, to better trace the directionality of the mutations, and to try to replicate recently developed methods, such as in Green et al. (2010), the additional sample from a golden jackal (*Canis aureus*) was included in the study and used as a reference for most of the analyses.

With this multi-species approach, all the main genomic features will be investigated as of interest to pinpoint the *taxon*-specific elements that were more likely affected by domestication: single nucleotide (SNVs), structural (SVs) and copy number variants (CNVs). Finally, the time of domestication is one of the most discussed questions that the project will try to answer, also considering the case of possible post-divergence gene flow between the sequenced lineages and their population size changes occurred through time.

In the next paragraph, we will see in details the general methods employed in the project, and the ones specifically applied to the study of transcript-level variation, including splice sites and regulatory elements.

3.3.Methods

As we saw, genomics is a fast-evolving field, in which the available methods and resources are constantly changing. This implies that continuous improvements come out every year, often every month, but also that it is increasingly difficult to keep up with these advances.

In our case, the availability of a reference genome (canFam2, Lindblad-Toh et al. 2005) is a great opportunity to overcome the need for a self-assembly.

However, an improved assembly (canFam3) was recently completed, but not yet made publicly available. This means that a number of resources developed for the previous assembly (most of them available as genomic ‘tracks’ in the main databases, UCSC and NCBI) have not been updated yet, requiring additional steps to be included in the study.

In some cases, we chose to develop them *de novo*, such as for the identification of repeated elements; in others, we decided to simply translate the extant resources into the new genomic coordinates, such as for most of the gene annotations.

Similarly, all the methods described in these paragraphs are the ones we applied until the time of writing, but they could change significantly until the completion of the project.

3.3.1. Sequencing and mapping methods

A priori information

In order to determine the amount of variation expected in the mitochondrial chromosome, we downloaded whole mtDNA sequences available at NCBI (as in July 11, 2011) for the following canids: 2 poodles (DQ480494; AY565739), 1 basenji (AY656737), 1 Russian wolf (DQ480503), 2 Saudi Arabian wolves (DQ480506; DQ480507), 3 coyotes (*Canis latrans*, DQ480509; DQ480510; DQ480511), plus 1 golden jackal (*Canis aureus*, unpublished). The sequences were manually aligned in BIOEDIT 7.0 (Hall 2004) and the average number of observed nucleotide differences for each *taxon* compared to the boxer sequence was computed in DNASP v5.10 (Librado and Rozas 2009).

The same procedure was used to determine the level of nucleotide differences between *taxa* at the nuclear level, using as source 8080 bp sequences from 12 neutrally-behaving coding genes from Gray et al. (2009), available in a number of canid species.

All the samples whose sex was unknown were tested using the DBXIG and DBY7 markers from Seddon (2005), and the one with known sex was also included as a control (Basenji, Male).

To further match molecular sexing with genomic data, a pairwise comparison of the depth-of-coverage (DOC) computed in sliding windows along the chromosome X was performed for all the possible pairs of samples; their regression coefficient was then computed, expecting a double coverage, on average, along the sexual chromosome for female (XX) compared to male (XY) individuals.

Platforms and sample preparation

In order to differentiate the NGS approaches, two different platforms were used, with different libraries: for each sample, a paired-end library of 100+100 bp with a 400bp insert was run on a Illumina HiSeq machine; additionally, a simple fragment (50-75 bp) plus a long mate-pair libraries (50+50bp, with an insert of 1.5kb) were run on ABI SOLiD 4 instrument. The number of runs for each library and sample are indicated in Tab. 21.

Sample	SOLiD ABI Long mate-pair (50bp/50bp, 1.5kb insert)	SOLiD ABI Fragment (50-75bp)	HiSeq Illumina Paired-end (400bp insert 100/100bp)
Basenji	1 slide	Excluded	1 lane
Dingo	1 slide	2 slides	1 lane
Israeli wolf	1 slide	1 slide	1 lane
Croatian wolf	1 slide	1 slide	1 lane
Golden jackal	3 slides	1 slide + 3/4 slide	1 lane

Table 21: Type of libraries and amount of sequencing effort per platform, for each sample. The units (slides, lanes) correspond to the partitions available on each machine (whose details are indicated in Tab. 1, chapter 1).

Several computational stations were used throughout the analysis process (Tab. 22). Two servers were used during the first steps of alignment, when the required computational power and the size of the files are larger. Downstream analyses were also performed on local multi-core computers, whereas laptops were mostly used to access the servers in order to start and control the desired processes remotely.

Machine	No. of cores	RAM (total)	Operating System
'Hoffmann' (cluster)	12	48 Gb	CentOS
'Panga' (server)	12	32 Gb	CentOS
'Sisyphus' (computer)	8	21 Gb	Linus Ubuntu
PCs (e.g. laptop)	2	2.5 Gb	Windows XP, MacOS

Table 22: Example of computational resources used during the project, and their characteristics: number of cores, total RAM and operating systems.

Read alignment and validation

The reads were aligned to the reference genome assembly canFam3 using the software BIOSCOPE TM and NOVOALIGN (<http://www.novocraft.com/main/downloadpage.php>), respectively for SOLiD and HiSeq data.

Artificial duplicates, likely to have occurred during the amplification steps, were identified and removed with PICARD (<http://picard.sourceforge.net/index.shtml>).

GATK (http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit) was subsequently used in order to proceed with local realignments around variable sites, or short insertions or deletions (InDels), and then to recalibrate the quality scores assigned to each nucleotide in the alignment.

Before the final acceptance of the aligned sequences, several filters were applied in order to further reduce the possible sources of error: for every position to be validated, a maximum coverage of two times the average value across the genome should be respected, as well as a minimum genotype quality (GQ) of 20, as called from GATK. This allows excluding both the sites showing a low quality, but also the ones with excessive local coverage, which could be indicative of possible deletions or mismatches of similar, but non-homologous, sequences. Additionally, a given SNV should not set less than 5bp apart from another one, since the presence of multiple variable sites within a limited region is more likely to be due to misalignment problems rather than to a true series of polymorphisms. Moreover, regions rich in repeats of Guanine followed by Cytosine (C post G, or 'CpG', islands), have been masked out except for the transcript-level analyses, since these regions are supposedly very rare in the genome (given a natural deamination tendency that, on a long time scale, turns Cytosine into Thymine), except in regulatory regions, where they have a functional role.

To estimate an overall concordance with other methods, the variant sites were then compared with the alleles found by analyzing the same samples on a canine Illumina 172K SNP bead chip.

The total coverage for each sample was plotted as a cumulative distribution.

Whenever needed, custom scripts were written in the programming language PYTHON (<http://python.org/>), also exploiting the functionalities of the plug-in packages BIOPYTHON (<http://biopython.org/wiki/Download>) and EGGLIB (<http://sourceforge.net/projects/egglib/>).

3.3.2. Analyses workflow

The analyses proceeded in a step-by-step approach:

- we first checked for genomic features in the reference assembly that could influence the amplification and mapping success of our reads;
- second, we tried to assemble a reliable and exhaustive gene annotation database;
- in the end, we proceeded with a series of functional analyses in protein coding (Fig. 33) and regulatory regions (Fig. 34).

Genomic features in the reference assembly

Several genetic features can affect the final sequence data in different ways: some of them by altering the homogeneity of the library preparation, others by making harder or easier to map the sequenced reads to the reference genome.

The relative content of Gs (Guanine) and Cs (Cytosine) in a sequence ('GC content') is known to potentially affect the efficiency of amplification reactions, given the different number and strength of their molecular bonds compared to Adenine and Thymine. Therefore, we tested whether the GC content locally affected the depth of coverage (DOC) of our samples. For each chromosome, DOC and GC content values were calculated within 100 randomly chosen windows of 5Kb, and they were plotted using the R software package (<http://www.r-project.org/>).

Another source of bias in the possibility of uniquely map the sequenced reads to the reference assembly is given by the presence of repeated elements (LTR, SINE, LINE, STR, retrotransposons) along the genome. When a duplication event is relatively recent, since there is limited time for new mutations to occur, the two or more copies hardly differentiate one from the other, making it impossible for the mapping software to uniquely assign the reads to a specific position.

Therefore, the software REPEATMASKER (<http://www.repeatmasker.org/RMDownload.html>) was downloaded and run on the reference canFam3 genome assembly to identify the regions matching the repeated elements contained in RepBase (<http://www.girinst.org/server/RepBase/index.php>). The [xsmall] option was used, allowing the software to represent the sequence of repeated elements in lowercase letters.

Similarly, even outside known repeated elements, it is possible that throughout the genome a number of k-mers (sequences of k nucleotides) randomly match one another. If a given k-mer with size larger than the read length is present several times in the genome, it can introduce an additional source of bias during the mapping step, whereas -if it is unique- it will be easier for the software to map the reads against it. This issue is usually indicated as ‘mappability’.

Therefore, in order to account for biases in the mappability of the reads along the reference genome, TALLYMER (GENOMETOOLS, <http://genometools.org/pub>) was used to verify how many times every possible 50mer (sequence of 50 nucleotides), corresponding to our shorter read length, was contained in the reference genome. This value was retrieved and plotted for every non-overlapping 50 bp window. Regions showing values consistently bigger than one were therefore non unique, and considered sensible for possible mismapping problems.

A list of known Copy Number Variants (CNVs) in dogs was retrieved from (Nicholas et al. 2011). In their study, they used a high density genome wide tiling array to discover CNVs in modern dog breeds and gray wolf, and also summarized all the previously reported CNV regions discovered in modern dogs. All the locations of the known CNV regions were transposed from canFam2 into canFam3 assembly thanks to the LIFTOVER TOOL (<http://hgdownload.cse.ucsc.edu/admin/exe/>) and considered as a further potential source of mismapping problems, although specific investigation about their presence in our target genomes will be carried out during other steps of the project.

As it is well known, only a limited portion of the eukaryote genomes is constituted by functional elements. By aligning the genomes of several species, it is possible to identify which regions are more conserved across *taxa*, probably implying functional constraints, and which ones are free to more rapidly evolve and diverge. The conservation scores computed across four mammal genomes were thus downloaded from UCSC genome browser (<http://hgdownload.cse.ucsc.edu/goldenPath/canFam2/phastCons4way/>) and transposed to canFam3 assembly by LIFTOVER, expecting a higher proportion of mutations in the less conserved regions.

Gene annotation

A key step in every genomic study is the availability of a good set of annotated genes, for whose transcripts the starting and ending coordinates of each exon (and, if present, also of untranslated terminal regions, UTRs) are known by gene expression studies or inferred by computational mining procedures. However, except rare cases, reliable and complete gene annotation sources are not available for the target organism, making every assumption about mutations in coding regions harder.

Nonetheless, at least in the case of the dog, a number of different databases provide information about gene coordinates and functions, although they can largely vary and the accuracy of the provided data can be not extensively verified.

In order to build the most comprehensive gene annotation set, we combined the available information from three different sources:

- UCSC RefGene (<ftp://hgdownload.cse.ucsc.edu/goldenPath/canFam2/database/>);
- ENSEMBL via Biomart (<http://uswest.ensembl.org/biomart/martview>);
- NCBI Seq_gene (ftp.ncbi/genomes/Canis_familiaris/mapview).

The retrieved information included gene names, symbols and genomic coordinates of coding exons and UTRs. Additionally, ENSEMBL and NCBI databases showed information about alternative transcripts, at least for some of the genes. The UCSC RefGene file contained information on 1,131 genes and 1,168 transcripts, ENSEMBL 24,600 genes and 30,915 transcripts, NCBI Seq_gene, 19,767 genes and 33,653 transcripts.

All the transcripts were tested for having an apparently functional coding sequence (CDS) by retrieving the corresponding sequence from the canFam3 genome and looking for: i) the presence of start and stop codons, ii) transcript length to be multiple of 3bp, and iii) absence of premature stop codons. Approximately 74% of transcripts from UCSC RefGene, only 35% from Ensembl, and 96% from NCBI Seq_gene satisfied these conditions (Fig. 31), and were grouped to build our final annotation set ('CDS-OK transcripts', Tab. 23).

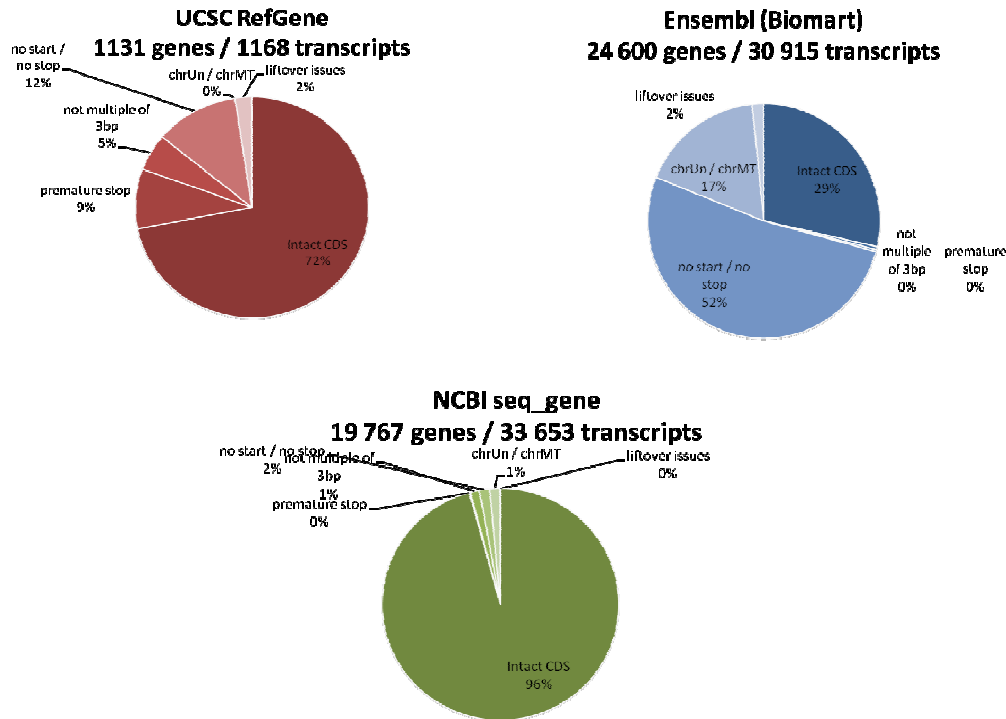


Figure 31: Pie charts indicating the total number of annotated transcripts included in each source. The slices represent what proportion did show to have an intact CDS (having start and stop codon, being multiple of 3bp, not showing premature stops in the boxer genome), how many did not meet these criteria, were mapped to mitochondrial or unknown chromosomes, or were not successfully transposed from canFam2 to canFam3 dog genome assembly.

Transcript Count	RefGene	Ensembl	NCBI
On known chr in canFam3	1135	25146	33158
With perfect CDS	959	8944	32280
Unique to each source	221	5267	28311
Total unique transcripts		37810	

Table 23: Summary of the number of annotated transcripts retrieved from each source, being mapped on a known chromosome on canFam3 assembly, having an intact CDS, and being unique to each source. The total number of transcripts that we included in functional analyses is also indicated.

The ‘CDS-OK transcripts’ set is the non-redundant intersection of NCBI Seq_gene (32,280 transcripts – 19,494 genes), ENSEMBL/Biomart (8,944 unique transcripts – 4,954 genes) and UCSC RefGene (959 unique transcripts – 145 genes). When a transcript was

present in more than one annotation set, the priority was given to the one coming from NCBI Seq_gene, then from ENSEMBL/Biomart, finally from UCSC RefGene, for a total of 37,810 unique transcripts and 18,782 genes (Fig. 32).

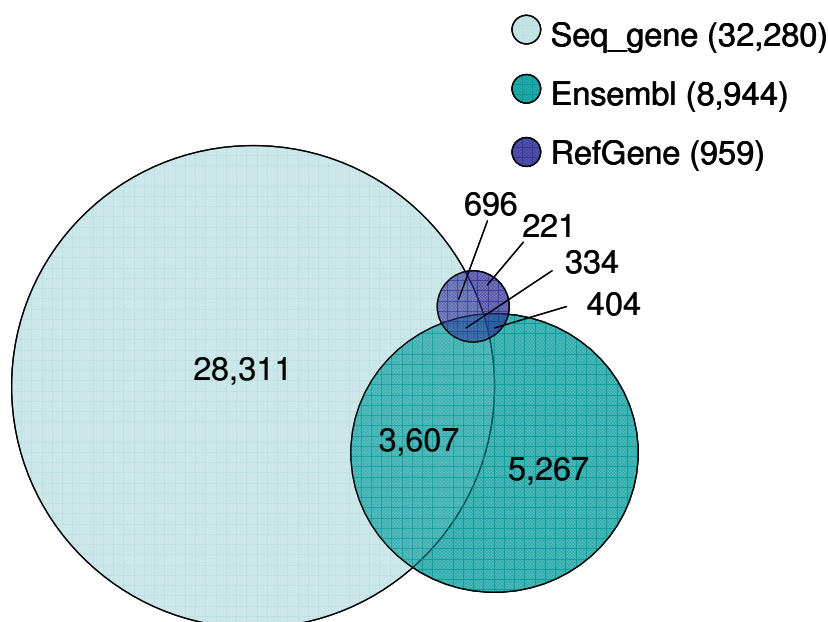


Figure 32: Venn-Euler diagram representing the number of transcripts available from each source meeting our functionality criteria, and their levels of overlap. If a transcript was present in more than one database, it was included only once, the priority given to the ones in Seq_gene, then in Ensembl, lastly in RefGene.

The 17,112 transcripts that did not pass the CDS filters in the boxer genome were retained and grouped in an additional annotation dataset ('CDS-fail transcripts' set), to examine whether the CDS could be intact in other *taxa*, but not in boxer (Fig. 33).

Coding regions analyses: DOC, SNVs, splice sites and functionality in transcripts

First, we calculated the bases passing quality controls in all the transcripts in terms of 1) average percent of bases passing quality control for each *taxon* and 2) the frequency distribution of bases passing quality control filters.

Next, we used linear regression to test for correlation of transcript-specific coverage among each pair of *taxa*, and to test for correlation between percentage of bases passing quality controls and GC content.

For transcripts with either excessive (higher than three times the average DOC) or zero pre-filter DOC, the DOC was calculated in sliding windows within the region surrounding the transcript ($\pm 1\text{Mb}$) to assess whether the reduction or increase in DOC was transcript-specific,

therefore suggesting a duplication or deletion event, or the result of the overall coverage variation in a wider region. Regions of zero DOC were experimentally verified using Polymerase Chain Reaction (PCR) followed by gel run and/or Sanger sequencing.

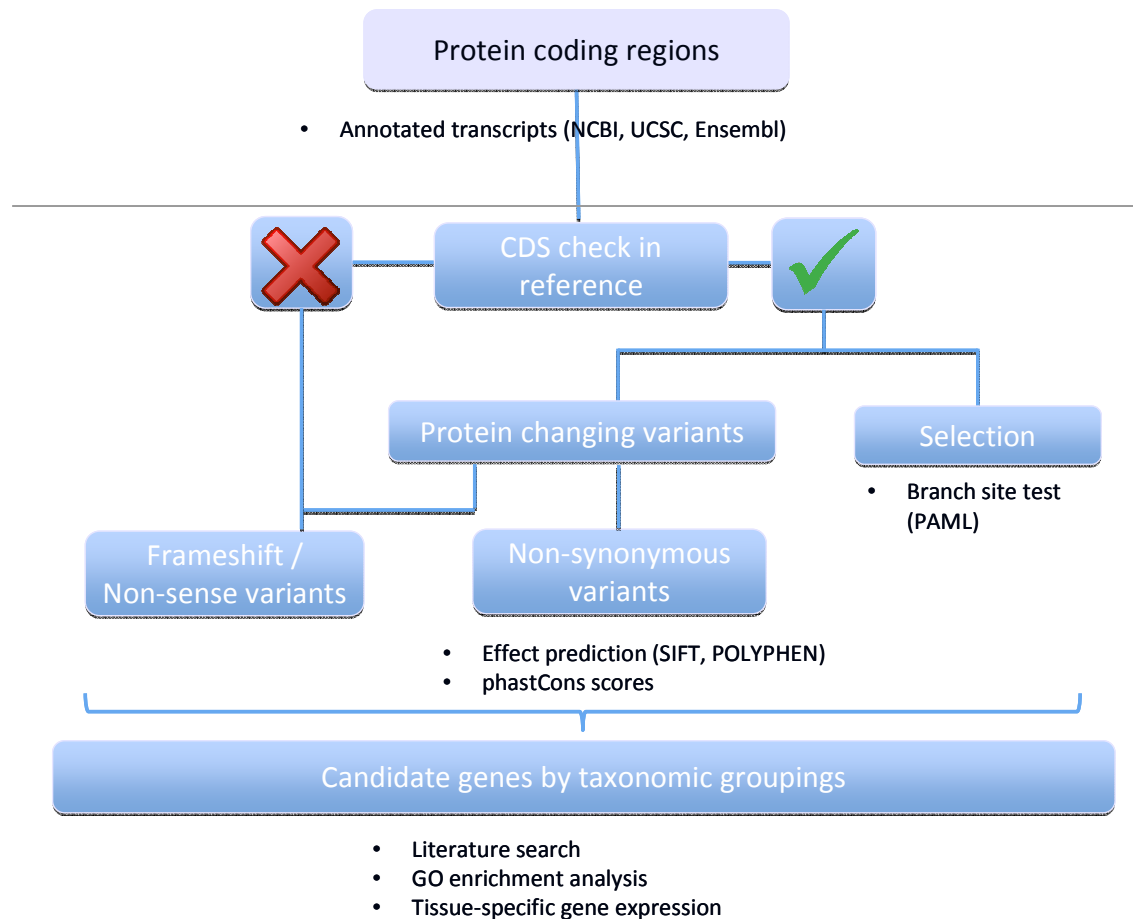


Figure 33: Schematic workflow of the analyses performed on protein-coding regions, starting from the known annotated transcripts and ending with a list of candidate genes hosting differential variants, potentially associated with domestication

Each of the annotated transcripts within our CDS-OK set was tested for an intact CDS (see above) in each of the five sequenced *taxa*. For all the transcripts that in one or more *taxa* were missing a start or a stop codon in the positions predicted by the annotation, we checked for an alternative open reading frame (ORF) within 1000 bp from the start or the end of the transcript. If a transcript in one or more *taxa* still failed to meet the above conditions, the transcript was considered as potentially pseudogenized.

For each transcript sequence in our *taxa*, we identified SNV polymorphisms that satisfied our filtering conditions (see paragraph 3.3.1) and distinguished between synonymous

and non-synonymous mutations by translating the transcripts via BIOPYTHON based on the universal genetic code. Sites with non-synonymous mutations were grouped according to their distribution among groups of *taxa*, or as being specific to a given *taxon*, and were retained for further analyses. Most attention was dedicated to the mutations clustering according to the dogs (boxer, dingo, and basenji) *vs.* wild canids (wolves and jackal) partition. The transcripts hosting one or more SNVs were considered as candidates genes for subsequent analyses.

The effects of SNV mutations on protein structure and functionality were predicted using POLYPHEN-2 (Adzhubei et al. 2010). The boxer protein sequences were used as reference, but in order to obtain a prediction in accord to the evolutionary direction of the mutations (from wolf-jackal ancestor to the boxer), for each mutation we considered the boxer as being the mutated state, and the alternative allele found in one or more of the other *taxa* (jackal, wolves or ancient dogs) as the reference state. The sites were then ranked according to the predicted qualitative effects of the amino acid substitution ('benign', 'possibly damaging' or 'probably damaging'). However, the effects of the mutations were not considered in their directionality, but for the amount of effects, with 'benign' mutations causing little or no effect, and 'damaging' mutations affecting more severely the protein functionality.

Although able to take into account several structural parameters, POLYPHEN prediction are only based on a positive BLAST of the submitted sequences to described human proteins.

Therefore, we performed additional analyses in SIFT (<http://sift.jcvi.org/>), based on sequence homology and multiple alignments with all proteins from NCBI_nr database (<ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz>). The basic assumption of the software is that functionally important positions should be more conserved in an alignment of the protein family, and functionally less important positions can be more variable. Therefore, it predicts that positions with normalized probabilities less than 0.05 are deleterious, whereas those greater than (or equal to) 0.05 are tolerated.

For each position, the predictions from both POLYPHEN and SIFT were compared, and only the concordant ('possibly damaging' or 'probably damaging', and 'deleterious') sites were considered as having the largest effects.

Subsequently, in order to detect in which tissues the candidate genes resulting from the previous analyses are expressed, we followed the procedure in Li et al. (submitted) and downloaded the following gene expression databases from <http://biogps.org/downloads/> (Wu et al. 2009): Human U133A/GNF1H Gene Atlas (GEO code GSE1133), Human GNF1H chip annotation, Human U133A chip annotation. For each gene of interest, the five tissues where it was most highly expressed were considered.

However, beside non-synonymous single nucleotide variants, important differences in the protein functions can be determined by the arrangement of the exons that will be actually translated. This process is regulated by the splice sites, which determine the inclusion of a given exon in the final mRNA product of the gene, its complete skip or the exclusion of part of its sequence, either at the 5' or at the 3' end.

Active splice sites are canonically composed by AG or GT nucleotides at the first two bases upstream and downstream of the exon, respectively. Therefore, we looked for their potential differences among *taxa* by analyzing two different, and partly complementary, sets of splice sites:

- a) All the splice sites adjacent to the dog exon boundaries as from our final annotation set;
- b) Splice sites adjacent to all human (hg19 assembly) exons from KnownGene database with any syntenic correspondence (UCSC genome browser) with the dog genome (canFam2), with the following conditions: i. being conserved in human and mouse but not in dog; ii. being canonical (AG, GT); iii. being maximally conserved among other species in a 46way multi-genome alignment (UCSC genome browser). The coordinates of these sites were then transposed into canFam3 assembly by LIFTOVER.

The first set is more likely to identify genes with splice variants that are active in dogs (whose genome has been used to develop the annotation databases we based on), but potentially not in the other *taxa*, whereas the second one is targeted to find splice sites that are not conserved in the boxer genome (therefore less likely to be included in the annotated transcripts), but which could still be conserved in the ancestral *taxa* as they are in a number of other mammals. For all the sites, we retrieved the corresponding dinucleotide sequences from our genomes and we identified all the splice variants meeting our quality criteria, which were then tagged according to their partitions among *taxa*.

Starting from sites that were alternatively active in dogs vs. wild canids and having complete information in all the *taxa*, we looked at the molecular function of the genes they belonged to and we identified a list of candidate hits potentially relevant for dog domestication; for those, specific primers were designed (PRIMER3 at NCBI, <http://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi>) and the amplified DNA was Sanger re-sequenced to validate the genomic data in the analyzed samples. If the alternative variant was confirmed, the same splice site was further sequenced in a panel of additional dogs, wolves and jackals.

Whenever their functions are known, the genes can be grouped into categories reflecting their roles, activities, and molecular pathways. Several databases provide access to these categorizations, and sets of candidate genes can be tested for being enriched in

particular categories. Therefore, genes with either non-synonymous mutations, variant splice sites, or disrupted open reading frame (ORF) were grouped as: i) boxer-specific; ii) dogs (boxer, dingo, basenji) vs. wild canids (golden jackal, wolves), iii) wolves-specific, and iv) jackal-specific, with particular focus on the second group, and tested for enrichment in Gene Ontology (GO) categories, KEGG/REACTOME pathways and Human Phenotype Ontology using G-PROFILER (<http://biit.cs.ut.ee/gprofiler/>). All the dog (*Canis familiaris*) genes annotated in ENSEMBL were used as the reference set, and a Benjamini-Hochberg correction was applied to control for false discovery rate (FDR).

Olfactory Receptor (OR) genes represent the largest gene family in most mammalian genomes, with more than a thousand genes described in mouse and rat (reviewed in Rouquier and Giorgi 2007). Also humans, despite a significant reduction in their olfaction skills that is common in primate species, still retain about 960 OR genes, most of which (*ca.* 60%) are potentially functional. Intermediate number have been documented in dog, with more than a thousand genes identified, *ca.* 20% of which are pseudogenized (Robin et al. 2009, Tacher et al. 2005), but little is known about their functionality in wolves (Quignon et al. 2011, Zhang et al. 2011). However, many of them seem not to be included in the available gene annotation datasets. Therefore, to test for evidence of selection or pseudogenization in this important gene family, we obtained the list of .fasta sequences of the 1121 genes previously described in dogs (Robin et al. 2009). A local BLAST (>90% identity) of these sequences to the boxer genome (canFam3) retrieved 954 hits matching their expected chromosomal location in the genome assembly. This subset of sequences was analyzed for DOC, pseudogenization (via our coding sequence check), and SNVs, as described above.

Regulatory regions: promoters and UTRs

Given the possible large-scale effects of gene regulation processes on phenotypic traits that differentiate dogs from their wild ancestors, we investigated the presence of mutations in the main regulatory regions flanking the coding sequences of the annotated genes of the reference canFam3 genome: 5'UTRs, 3'UTRs, and promoter regions (Fig. 34). Promoters and 5'UTR host the sites bound by the elements that regulate the transcription of mRNAs, whereas 3'UTRs can be targeted during post-transcriptional regulation.

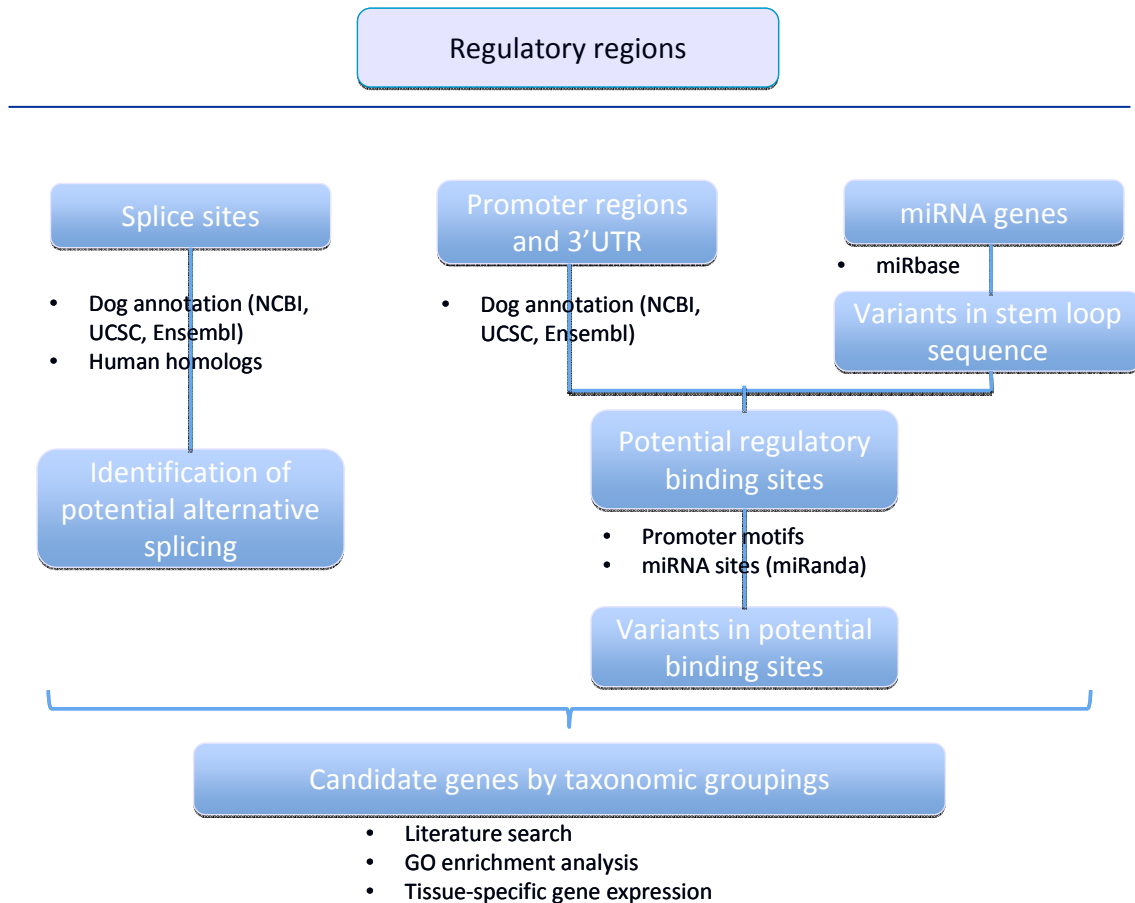


Figure 34: Schematic workflow of the analyses performed on regulatory regions, splice sites and microRNA-coding genes, ending in a list of candidate genes potentially associated with domestication

For each transcript, whenever the coordinates of the UTRs were available in our annotation sources they were taken into account, otherwise the transcript was considered without UTRs. The promoter regions, instead, were considered as being the sequences including 500bp upstream and downstream of the transcription start site (TSS) of each transcript. Although promoter sites can reside up to thousand bases upstream of the TSS, most of them are in this proximal interval (Xie et al. 2005). If the beginning of the CDS of the transcript (the first exon) was within 500bp from the TSS, the promoter end was considered as the nucleotide preceding the CDS start. If a known 5'UTR was present, its sequence was fully included in the promoter region; if more than one 5'UTR was present, different promoter regions were defined in order to include a progressive number of 5'UTRs.

We then matched a set of mammal-specific known regulatory motifs described in Xie et al. (2005) to every promoter and 3'UTR region in order to detect the actual binding sites for regulatory elements; to do that, we used the package EGGLIB in PYTHON, checking for

matches on both strands and allowing for one mismatch from the known motif. If a hit was found, it was retained only if the average PhastCons score (a conservation value computed base by base on a 4-genome alignment (<http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=245824499&c=chr14&g=multiz4way>, Siepel et al. 2005) of its sequence was higher than 0.8, since most of the actual binding sites are supposedly highly conserved across species. Then, the presence of mutations in our sequences was evaluated, and the genes were considered as potential candidates for differential regulation levels.

Additionally, micro RNAs (miRNA) are short (*ca.* 23bp) RNA sequences potentially responsible for post-transcriptional gene regulation processes (Alvarez-Garcia and Miska 2005, Bushati and Cohen 2007) However, genes coding for miRNAs are largely missing from the annotation sources. Therefore, in order to look for variants in their sequences among our canid genomes (Zhou et al. 2008), we downloaded and converted from canFam2 to canFam3 the genomic location of known miRNA genes available in miRbase (<ftp://mirbase.org/pub/mirbase/CURRENT/genomes/cfa.gff>), as well as the sequences of the stem loops and of the mature sequences from the same source (http://www.mirbase.org/cgi-bin/mirna_summary.pl?org=cfa). We then aligned the mature sequences to the respective stem loop sequences and calculated the genomic positions of the mature sequences start and end; finally, we verified the presence of mutations in the miRNA sequence in our *taxa* and their location (within or outside the mature sequence). To predict specific miRNA target sites, we ran MIRANDA (http://cbio.mskcc.org/microrna_data/miRanda-aug2010.tar.gz) on the previously identified 3'UTR sequences (score ≥ 120 , energy ≤ -20) and calculated the genomic positions of the predicted target sites.

Genome browser tracks

As a tool for visualizing the main genomic features, we utilized JBROWSE genome browser (<http://jbrowse.org/>) that allows to efficiently display and track genome-wide data by selecting any genomic interval, which we based on the most recent dog genome assembly (canFam3).

We included both quantitative (e.g. GC content and missing positions, mappability and conservation scores) and structural data (such as gene regions -divided in exons, introns and UTRs), and the information resulting from previous analyses (e.g. the variable positions described in dogs, known CNVs, the values of selective pressure computed by F_{ST} and $F_{ST}/XP-EHH$ in Vonholdt et al. 2010). For these data, specific custom tracks have been

produced and uploaded through a dedicated server, and the browser will be open to the whole scientific community at the end of the project.

The GC content and the rate of missing positions (Ns) in the reference genome were calculated in non-overlapping sliding windows of 100bp. Mappability scores were computed with TALLYMER along 50bp windows, as described above, and visualized. Repetitive elements detected by REPEATMASKER were indicated in the corresponding positions, as well as the known CNV regions. Four-way (human, mouse, rat and dog) conservation scores (PhastCons) were downloaded from <ftp://hgdownload.cse.ucsc.edu/goldenPath/canFam2/phastCons4way> and transposed from canFam2 to canFam3 assembly using LIFTOVER. A database containing the known single nucleotide polymorphisms in dog (dbSNP) was downloaded from <http://genome-preview.ucsc.edu/cgi-bin/hgTables> and transposed to canFam3 as described above.

For each of the sequenced genomes, the average DOC and genotype quality (GQ) scores computed across 5bp windows were visualized. For each SNV passing filtering criteria in some of the *taxa*, its position was visualized; for mutations occurring in a known coding sequence, their effect on the amino acid sequence (e.g. Leucine to Serine) was also displayed when accessed. An example screenshot is given in Fig. 35.

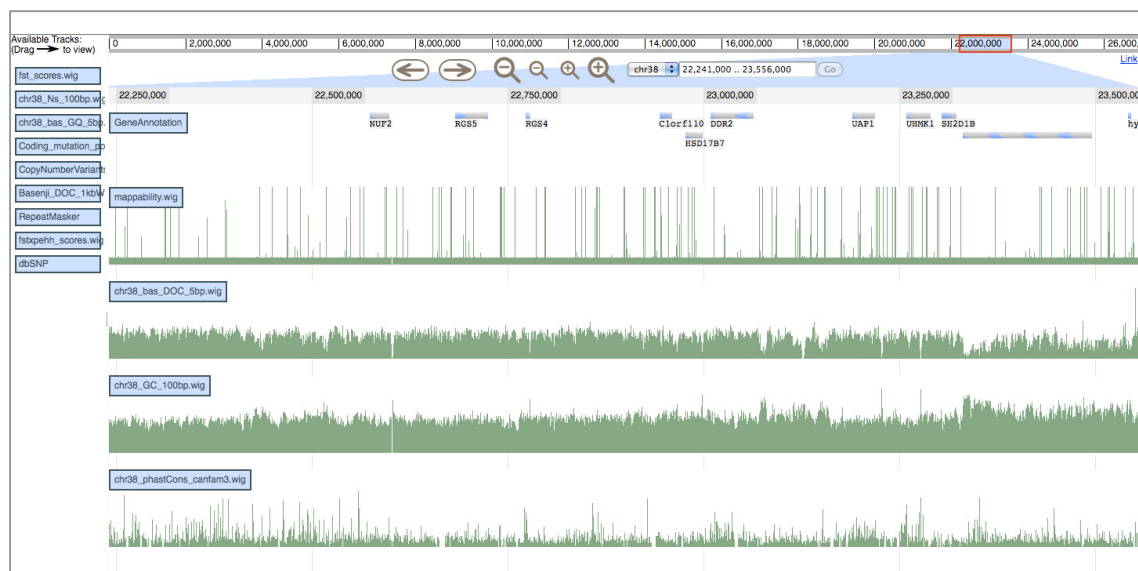


Figure 35: Example of visualization in JBrowse Genome Browser. In the top bar, the selected chromosome for the reference genome is displayed, whereas the toolbar allows to move along it, regulate the zoom level, or select any specific chromosome and location. On the left, the available track names are displayed (e.g. dbSNP, RepeatMasker). They only need to be dragged on the central screen in order to be visualized, in addition to the ones already present (e.g. gene annotation, mappability, etc.). Extra custom tracks can be uploaded.

3.4. Preliminary results

The sequencing effort yielded a minimum of 500 million reads per sample, most of which (>90%) were uniquely aligned; a variable percentage of duplicated reads was removed, resulting in a coverage ranging from 12x for the Basenji to about 25x for the Croatian wolf (Tab. 24).

	Basenji	Dingo	Israeli Wolf	Croatian Wolf	Golden Jackal
Total Reads (*)	535.8	755.9	1,281.3	1,534.5	2,468.5
Reads Aligned (*)	500.9	712.8	1,249.5	1,464.5	2,437.8
% Reads Aligned	93.5%	94.3%	97.5%	95.4%	98.8%
% PCR Duplicates	13.1%	24.0%	20.2%	16.0%	51.5%
Unique Aligned Reads (*)	435.2	541.6	997.3	1,229.8	1,181.8
Unique Aligned Bases	36.9 Gb	50.3 Gb	52.7 Gb	62.4 Gb	59.5 Gb
Average Coverage	14.59x	19.87x	20.83x	24.67x	23.52x

Table 24: Sequencing output, number and proportion of mapped and duplicated reads, and effective average coverage per sample. (* = millions of reads). Courtesy of Adam Freedman.

The cumulative distribution of the minimum coverage across the genome (Fig. 36), revealed that more than 80% of it was covered at a minimum of 10x in all *taxa* (and about 60% at 20x), except for the basenji, whose average coverage was lower given that one of its libraries had to be excluded, since not meeting minimum quality criteria.

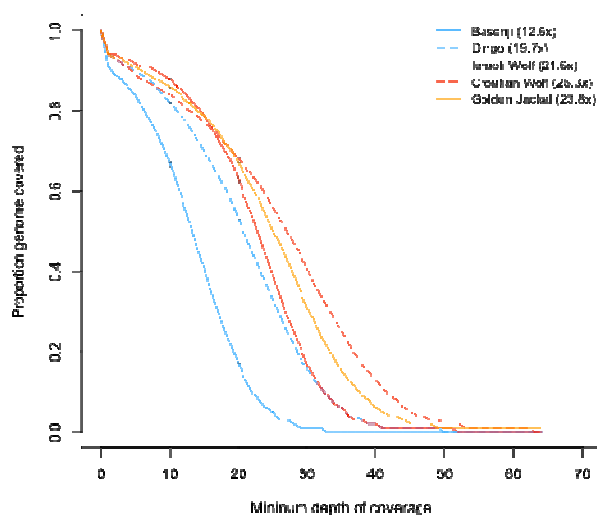


Figure 36: Distribution of the coverage across the genome, with the cumulative proportion of genome covered at a minimum depth is showed for each sample.

When computed across sliding windows (100 kb), the coverage distribution turned out to be strongly correlated between *taxa* (Fig. 37), with a common decrease in the telomeric regions of most chromosomes.

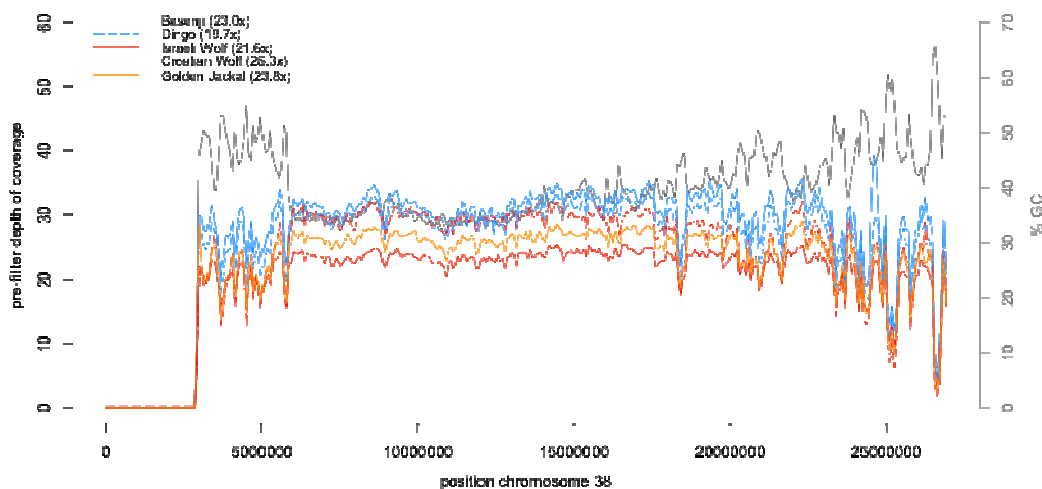


Figure 37: Example of the variation in the pre-filter depth of coverage for each sample along chromosome 38, as calculated across sliding windows of 100 kb (left axis). The percent GC content is indicated by the gray line (right axis).

However, the fluctuations in the coverage were almost perfectly explained by the variation in the GC content along the same regions (Fig. 37), with a strong negative correlation (Fig. 38). This implies that can be difficult to successfully analyze regions rich in GC, such as telomeres, but also that a position covered in a given *taxon* will be more probably covered also in the other ones, increasing the regions where a comparison is possible.

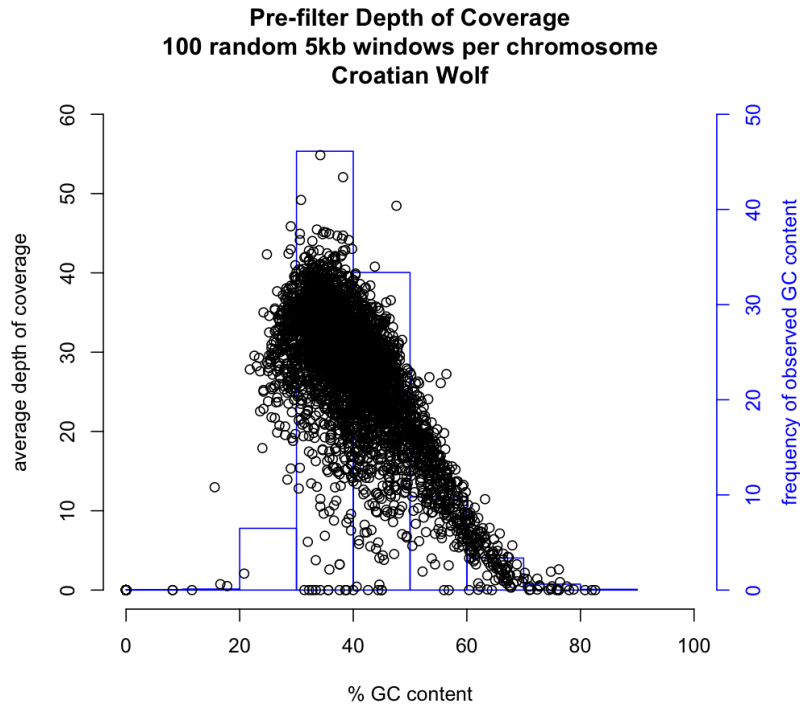


Figure 38: Correlation between depth of coverage and GC content (%), as calculated within 100 random 5 kb windows per chromosome (left axis). The frequency distribution of GC content classes in the windows is indicated by the blue histogram (right axis).

However, some regions showed a constant coverage close to zero in all *taxa*, the largest of which (1.8 Mb) was situated on chromosome 31. In this case, the decrease in the coverage was not explained by any corresponding increase in GC content. This would be explained by two reasons: a large duplication event that occurred in the boxer lineage, but not in the other *taxa*; the presence of some genomic features that did not allow the reads to be mapped in this region, but maybe elsewhere. To explain it, we blasted the whole region against canFam2 assembly through NCBI to look for similar regions in the genome, without finding any significant hit other than the region itself. Therefore, we divided the whole region into 5000 bp segments and systematically blasted them, using a local installation of BLAST on canFam3 assembly. This time, most of the segments returned a double match, not outside, but within the region itself, and some of them (both around and within the region) matched a common LINE element. These findings (Fig. 39) indicate that a duplication event, possibly facilitated by the presence of LINE elements, actually occurred in the dog genome, but was not resolved in the previous genome assembly. However, it could still represent either a new feature specific to the boxer, not present in the other *taxa*, or simply being due to the impossibility for the software to successfully map the reads within the region, that is also duplicated in the

other *taxa*. To verify these alternative possibilities, a couple of primers was designed on the flanking regions, and at both ends of each duplicated segment; then the DNAs from the same animals were amplified: in case the reactions only worked at the flanking sites, this would mean that the duplication is not present in the genomes other than the boxer, whereas if it yielded a product also between the two repeated regions, the presence of the duplication would be confirmed also in the other *taxa*.

The latter hypothesis turned out to be true, implying the uniform presence of the duplication across all *taxa*, and the inability for the alignment software to uniquely map the reads in such a duplicated region. This pattern was later confirmed by calculating the mappability scores across the genome, which showed a constant value of 2 in the entire region, except for a limited portion around position 31.5 Mb, where a short deletion in only one of the duplicated segments made the mapping of some reads possible (Fig. 39).

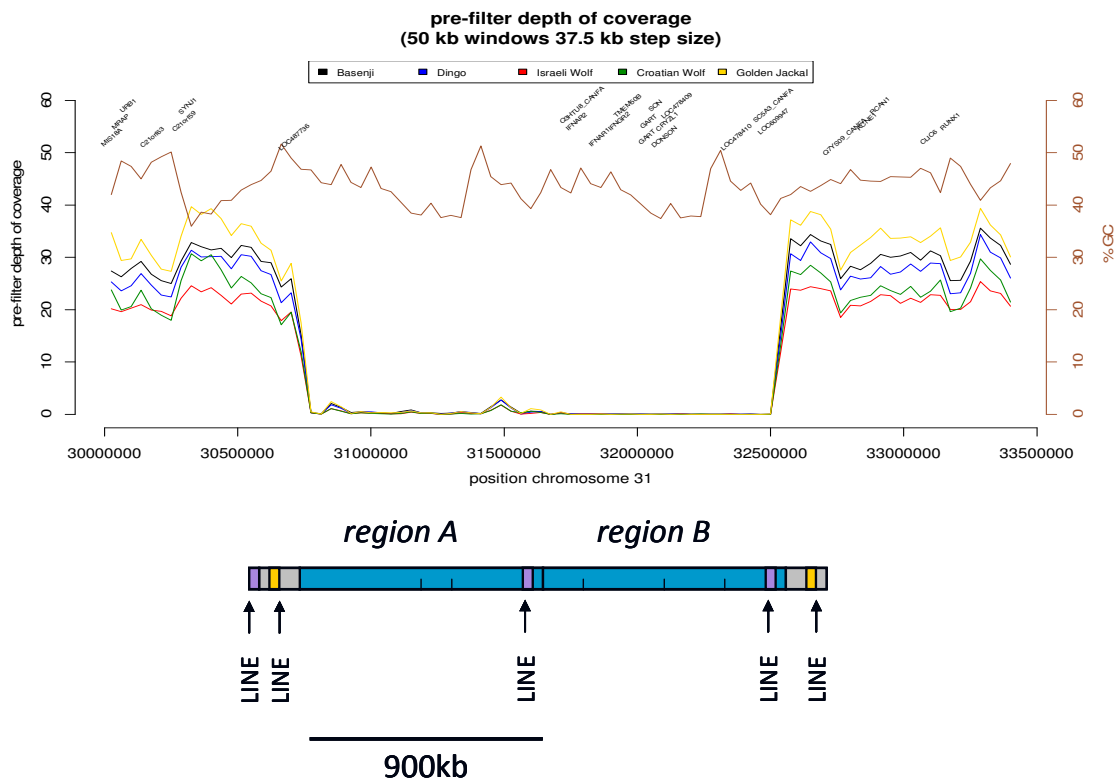


Figure 39: Distribution of the DOC (left axis) and GC content (right axis), along a portion of chromosome 31, where a large region (1.8 Mb) appears to have a constant drop in the coverage for all *taxa*, due to a large duplication within the region that does not allow a unique mapping of the reads. A schematic representation of the chromosome features is showed below the corresponding coordinates.

The analysis of SNVs within functional regions was possible for a high number of transcripts, promoter regions and UTRs (Tab. 25), yielding more than 100,000 variable sites that matched reliability criteria (Tab. 26).

Class	Aim	Number
Transcripts CDS OK	Functional analysis	37,810
Transcripts CDS fail	Functional analysis	17,112
Olfactory receptor genes	Functional analysis	952
miRNA genes (from miRbase)	Variation detection	323
Promoter regions	Regulatory motif search (promoters)	40,402
5' UTR	Regulatory motif search (promoters)	27,403
3' UTR	Regulatory motif search (miRNAs & other)	25,494

Table 25: Total number of genetic features included in the analyses of functional regions. Olfactory receptors and micro RNA genes have been identified independently from the main annotation sources.

The mutations resulted in a disruption of the ORF in *ca.* 1000 cases, in most of which the presence of a premature stop codon in one or more *taxa* other than the boxer was revealed. More than 17K out of 37K transcripts with a correct CDS hosted one or more non-synonymous substitutions.

However, when grouping them according to the distribution of alleles across *taxa*, most of them turned out to be specific to the golden jackal sample, as expected from its more distant phylogeny. Only 140 out of more than 50,000 differentiated dogs *vs.* wild canids without any missing data, therefore significantly lowering the number of candidate genes associated with domestication.

Moreover, after performing a prediction on the effects of the mutations by POLYPHEN, only 3 out of more than 9,000 possibly or probably damaging mutations perfectly differentiated dogs from wild canids without missing data, whereas in other 100 cases the allele information for one or more *taxa* was missing or not reliable, but still compatible with this partition.

Total number polymorphisms in coding regions	136,459
-Transcripts with synonymous mutations	25,397
-Transcripts with non-synonymous mutations	17,326
-Premature stop codons in ≥ 1 <i>taxa</i>	874
-Extended CDS in ≥ 1 <i>taxa</i>	134
-Synonymous mutations	85,103
-Non-synonymous mutations	51,356
-boxer-specific (no missing data)	988
-dogs vs. wild canids (no missing data)	140
-wolf-specific (no missing data)	62
-jackal-specific (no missing data)	4,525
-PolyPhen: possibly or probably damaging mutations	9,333
-dogs vs. wild canids (with missing data)	100
-dogs vs. wild canids (no missing data)	3

Table 26: Total number of single nucleotide variants discovered, their effects on the protein level (synonymous or non-synonymous, non-sense or CDS-extending), and the allelic partition across *taxa*. For non-synonymous mutations, the number of them leading to an extreme change in the protein functionality as predicted by POLYPHEN is indicated.

The three genes identified in this manner (Tab. 27) were further investigated in the literature. The first of them, SLK, has been associated with uterine fibrosis (Cha *et al* 2011), and known to be implicated in oocyte meiosis, apoptosis, nucleotide-excision repair, and protein amino acid phosphorylation. It is thought to interact at the protein level with other known genes, such as CASP3, CLSTN1, KEAP1, PDZK1.

RNF12 is known as an X-Encoded Dose-Dependent Activator of X Chromosome Inactivation (Jonkers *et al.* 2009). Beside the pun, it also interacts with Estrogen receptor alpha, therefore showing an important role both in the X-chromosome dose control and in hormonal regulation.

DLGAP5 (also known as DAP-5, HURP or DLG7) stabilizes microtubules in vicinity of chromosomes, controls the spindle dynamics, promotes the interkinetochore tension and an efficient kinetochore capture (Wilde 2006). It has been found to be expressed also in cancer and stem cells (Gudmundsson *et al.* 2007, Sanderson and Clarke 2006).

However, when a gene enrichment analysis was performed on all the genes hosting non-synonymous mutations discriminating dogs from wild canids, other interesting genes

turned out to be related to functional categories significantly more represented than expected (Fig. 40).

Gene and description (transcript)	Chromosome Location	Mutation type	Polarity change	Charge change	PolyPhen prediction
SLK serine/threonine kinase 2 (XM_544006.2)	Chr28:19,286,345 - 19,345,508	D475V	polar --> nonpolar	negative -> neutral	Possibly damaging
RNF12 similar to ring finger protein 12 (XM_849915.1)	ChrX:29,454,754 - 29,457,935	W119L	same	same	Probably damaging
DLGAP5 discs, large (Drosophila) homolog-associated protein 5 XM_537454.2	Chr08:33,998,363 - 34,035,858	R225G	same	positive -> neutral	Possibly damaging

Table 27: List of the only three genes hosting non-synonymous mutations discriminating dogs (boxer, dingo and basenji) from wild canids (wolves and jackal), whose effects are predicted to be significantly affecting the protein functionality by POLYPHEN. The mutation type and its changes in amino acid charge and polarity are indicated.







P-value	T	Q	Q&T	Q&T/Q	Q&T/T	term ID	term domain and name
4.05e-04	13	44	2	0.056	0.154	 GO:0070325	MF lipoprotein particle receptor binding (1)
2.35e-04	10	44	2	0.056	0.200	 GO:0050750	MF low-density lipoprotein particle receptor binding (2)
P-value	T	Q	Q&T	Q&T/Q	Q&T/T	term ID	term domain and name
1.39e-03	13	36	2	0.250	0.154	 HP:0010979	hp Abnormality of the level of lipoprotein cholesterol (1)
1.18e-03	12	36	2	0.250	0.167	 HP:0010981	hp Hypolipoproteinemia (2)
2.42e-03	145	36	4	0.500	0.028	HP:0005918	hp Abnormality of the phalanges of the hand (1)
4.35e-03	80	36	3	0.375	0.037	HP:0003119	hp Abnormality of lipid metabolism (1)
4.03e-03	22	36	2	0.250	0.091	 HP:0000908	hp Hypoplastic ribs (1)
2.40e-03	17	36	2	0.250	0.118	 HP:0000773	hp Short ribs (2)
9.84e-04	11	36	2	0.250	0.182	HP:0006477	hp Abnormality of the alveolar ridges (1)
P-value	T	Q	Q&T	Q&T/Q	Q&T/T	term ID	term domain and name
4.70e-02	16	41	1	0.067	0.062	KEGG:00450	ke Selenocompound metabolism (1)
2.96e-02	10	41	1	0.067	0.100	KEGG:00290	ke Valine, leucine and isoleucine biosynthesis (1)
5.55e-02	19	41	1	0.067	0.053	KEGG:04977	ke Vitamin digestion and absorption (1)

Figure 40: Output of the enrichment analysis performed in G-PROFILER, on the non-synonymous mutations discriminating dogs (boxer, dingo and basenji) from wolves and golden jackal, for different functional category types and databases (MF=Molecular Function; GO=Gene Ontology; hp=Human Phenotype; ke=KEGG pathways). Enrichment p-values are indicated in the first column, the associated role in the last one.

In particular, APOB is the main constituent of LDL, and affects the levels of hypercholesterolemia. The gene FANCD2 is associated to Fanconi Anemia – a recessive disorder causing chromosomal instability, breakage and defective DNA repair. PPP1R3A is a subunit of protein phosphatase1; it binds to muscle glycogen and is possibly involved in

obesity and diabetes. COL17A1 (collagen XVII) is known to be related to a diminished epidermal adhesion and skin blistering. EVC2 is associated to Ellis–van Creveld Syndrome, which causes polydactyly, congenital heart defects, short-limbed dwarfism, cleft palate, and malformation of the wrist bones. Lastly, GNPAB (N-acetylglucosamine-1-phosphate transferase) is involved in mucopolidosis, a metabolic disease that causes mental and developmental problems.

Similarly, other interesting mutations discriminating domestic from wild canids emerged in the UTR regions (42 different genes at the 3'UTR, 84 genes at the 5'UTR). Among them, CLOCK influences circadian rhythms; STARD6 is part of the cholesterol homeostasis pathway, and GCG is a glucagon precursor; LRRN3 and LRRN6 (also known as LINGO2) are highly expressed in brain, whereas NLGN1 encodes neuronal cell surface proteins, involved in the formation and remodeling of central nervous system synapses. Other brain-related genes are SLC6A15 (a possible neurotransmitter transporter, highly expressed in brain, whose variants have been linked to depression) and CA10 (which may play a role in brain development). MYH8 (Maccatrozzo et al. 2007) and TTN are related to skeletal muscle contraction, showing the highest expression in cardiac and skeletal muscle tissue, affecting the muscular resting tension and being under regulated (MYH8) in dystrophic dog muscles (Guevel et al. 2011).

Among the genes showing a different open reading frame between domestic and wild canids, an interesting hit comes from transcripts having a premature stop in wild canids, therefore -from an evolutionary point of view- meaning an extended frame in dogs. The LOC612984 gene corresponds to Harakiri, BCL2 interacting protein. The activator of apoptosis harakiri (in humans known as *HRK* gene) regulates apoptosis through interaction with death-repressor proteins BCL-2 and BCL-X(L). Also named DP5, Harakiri is induced during neuronal apoptosis following exposure to amyloid beta protein.

The analysis of the splice sites revealed a limited number of them to be polymorphic in our *taxa* (Tab. 28), with only about one thousand variant sites summing both intron start and end. Even less of them showed a change from canonical to non-canonical dinucleotides, and only one (Tab. 29) had a perfect partition between dogs (GT) and wild canids (AT), without missing data.

This gene is called TRPS1, and codes for a zinc finger transcription factor that is associated with the Trichorhinophalangeal syndrome, which causes unique a series of facial features and skeletal abnormalities: bulbous nose, elongated philtrum, sparse hair, cone-shaped epiphyses and mild growth retardation.

Category /taxa grouping	Intron end	Intron start
Total number of sites analyzed	359,431	
Total number of variant sites	446	626
Boxer-specific	19	24
Dogs vs. wild canids	20	28
Wolf-specific	101	114
Golden jackal-specific	142	214

Table 28: Total number of splice sites identified from genes annotated in dogs, the ones showing variants in the dinucleotide sequence flanking the exons (intron end, intron start) and their allelic partitions across *taxa*. Only in 20+28 cases there was a change discriminating dogs from wild canids, and suggesting the presence of a potential alternatively spliced exon.

Transcript ID	box	bas	din	isw	crw	jac	grouping	strand	gene
XM_853200	GC	TC	TC	TC	TC	TC	boxer/others	+	POLR2A
XM_548912	GC	GG	GG	GG	GG	GG	boxer/others	+	DCAF8L2
XM_847325	GT	AT	AT	AT	AT	AT	boxer/others	+	LOC609966
XM_539698	AC	CC	CC	CC	CC	CC	boxer/others	+	LOC482581
XM_858274	GT	GC	GC	GC	GC	GC	boxer/others	-	HEATR5B
XM_862638	GT	GC	GC	GC	GC	GC	boxer/others	-	PGM3
XM_857538	GT	GC	GC	GC	GC	GC	boxer/others	-	XDH
XM_534142	CA	CG	CG	CG	CG	CG	boxer/others	-	PCDH9
XM_546769	AC	GC	GC	??	??	??	ancient_dogs/other	+	DBNDD1
XM_848093	GT	GT/GG	GG	GT	GT	GT	ancient_dogs/other	+	LOC610566
XM_846490	CA	TA	CA/TA	CA	CA	CA	ancient_dogs/other	+	LOC609263
XM_533391	GT	AT/GT	AT	GT	GT	GT	ancient_dogs/other	-	ZNF532
XM_844149	GT	GC/GT	GC/GT	GT	GT	??	ancient_dogs/other	-	
XM_857104	TC	CC/TC	CC	TC	TC	TC	ancient_dogs/other	-	ETV1
XM_534593	CC	CC	CC	CT/CC	CT	CC	wolves/others	+	WDR69
XM_852823	GT	?T	GT	AT	AT/GT	GT	wolves/others	+	AP2M1
XM_845217	GT	GT	GT	GT/GC	GC	??	wolves/others	-	PLEKHB1
XM_861283	GT	GT	GT	AT/GT	AT	GT	wolves/others	-	SPTB
XM_545728	TT	TT	TT	TC	TC/TT	TC	wild/others	+	OR10T2
XM_855643	GT	GT	GT	AT	AT	AT	wild/others	-	TRPS1
XM_547809	GC	?C	GC	AC	AC/GC	GC/AC	wild/others	-	TRIM9
XM_861669	GT	GT	GT	GT/CT	CT/GT	CT	wild/others	-	DST

Table 29: Example table of variable splice sites situated at the intron ends, and their partition across *taxa* (box=boxer, din=dingo, bas=basenji, isw=Israeli wolf, crw=Croatian wolf, jac=golden jackal). In a single case (gene TRPS1) the dinucleotides were fixed and split between dogs (where the splice site is in the canonical form GT, therefore more likely to be active) and wild canids, suggesting the presence of a potential alternative splicing form (“?” Symbol denotes bases without data or that did not pass quality filters).

Among the set of splice sites identified in humans, and with a correspondence in the dog genome, another single gene shows fixed and differential splice sites between dogs and wild canids (the latter group showing the canonical form): the ANLN (anillin, actin binding protein) gene. This gene is required for cytokinesis, and is essential for the structural integrity of the cleavage furrow and for the completion of cleavage furrow ingression.

Another gene, SNX19 (sorting nexin 19) is also compatible with this pattern, although with one uncalled nucleotide in jackal. SNX19 is a member of a large group of proteins localized in the cytoplasm showing a phospholipid-binding motif. Some members of this family have been shown to facilitate the protein targeting process.

The remaining results, concerning the other analyses presented in the methods section, at the time of writing are currently being processed and validated; therefore they can not be included in this thesis. The reported ones, although preliminary, will be discussed in the next paragraph.

3.5. Discussion and implications

The availability of the complete dog genome draft (Kirkness et al. 2003, Lindblad-Toh et al. 2005) was a powerful source of information. It allowed to investigate the levels of variability within and between dog breeds, to compare the genomic evolution with humans and other mammal species under their respective selective forces (Lindblad-Toh et al. 2005), to identify a number of genes related to phenotypic traits of primary importance in shaping the dog morphology (Boyko et al. 2010, Vaysse et al. 2011), as well as hundreds of diseases seriously affecting the life of the man's best friend.

It also showed how the domestication process, supposedly occurred between 15,000 and 27,000 years ago, occurred in at least two separated steps, the original domestication from wolves and the strong artificial selection that led to the creation of modern breeds. The number of polymorphic markers identified in the dog genome also allowed to carefully reconstruct the phylogeny among canid lineages (Vonholdt et al. 2010), ruling out any *taxon* other than wolf as the ancestor of the domestic dog, and contributing to resolve the assignment of highly debated *taxa*, such as red (*Canis rufus*) and Great Lakes wolves (*Canis lycaon*), showed to be admixed forms between wolves and coyotes.

Now, the completion of the first wolf genome draft will shed light on other important, unresolved questions. Primarily, it could provide more specific answers on where and when the first domestication event(s) occurred, and possibly from which source population of wolves. Parallel, it will give us the opportunity of identifying at a genome-wide scale which

genetic features (Single Nucleotide Variants in functional regions, regulatory elements, Copy Number Variants, etc.) are differentially present in the dog and in the wolf genome, and possibly being associated with domestication.

Here, we present the preliminary results from a 5-genome sequencing project conducted by an international team based at University of California, Los Angeles. Using different Next-Gen platforms (ABI SOLiD and Illumina HiSeq, with multiple libraries of single reads and mate-pair or paired-end reads), we completed the sequencing of five canid genomes, including samples from two ancient dog breeds (dingo and basenji), from two wolf populations (Europe and Israel), and a golden jackal (also from Israel), which have been mapped against the most recent dog assembly (canFam3).

The inclusion of the two dogs is of primary importance to discriminate patterns of artificial selection in modern dog breeds (like the ones whose genomes are currently available: poodle and boxer) from the ones associated with the original domestication from wolves, since dingo and basenji are thought to be two of the most ancient extant dog lineages in the world. Moreover, the comparison of their genome with the ones of two representative wolf specimen will help resolving the uncertainty about the location of the center(s) of origin of domestic dogs, since different hypotheses have been proposed, suggesting that domestication could have first occurred in South-eastern Asia (Savolainen et al. 2002), therefore close to the distribution of modern dingoes, in Africa (Boyko et al. 2009), where basenji originated and currently live, or in the Middle East (Vonholdt et al. 2010), where one of the two wolf samples comes from.

The complete sequencing of a golden jackal will be also helpful in resolving the directionality of the mutations, therefore to better understand which mutations likely arose and were selected in the dog genome, and which alleles have been randomly inherited from an ancient polymorphism, as already applied to the study of *Hominidae* (Green et al. 2010).

However, the knowledge of the main genomic features that can enhance or affect the ability to amplify and map the sequenced reads is of primary importance for obtaining a reliable and high-coverage assembly. The relative frequency of Gs and Cs (GC content) in a given genomic region is known to potentially affect the success in the amplification of reads. In our genomic assemblies, it confirmed to be the main feature influencing the coverage of our samples along the genome, with a strong negative correlation. On the one side, this reduces the proportion of genome that can be covered at a sufficiently high depth to generate good quality nucleotide calls, e.g. in subtelomeric regions); on the other side, since the GC content is highly conserved in similar species, it can help in obtaining a larger proportion of

genome that can be successfully compared between *taxa* than what would be expected under random fluctuations of the coverage, therefore increasing the number of high quality markers that can be used in downstream analyses.

Still, other genomic features can seriously affect the ability to uniquely map the sequenced read to the reference genome. Relatively recent large duplications, such as the one we described on chromosome 31 (two segments of *ca.* 900 kb each), make it difficult to assign the reads to a single location, since the time from the duplication event was not enough for a number of new mutations to occur and significantly distinguish the two copies. Nonetheless, traditional sequencing techniques can be applied to verify the presence of such elements in the draft genomes, ruling out the possibility that they be *taxon*-specific, as we demonstrated by simple targeted PCR amplification.

More generally, the possibility of uniquely mapping reads along the genome ('mappability') strongly influences the final coverage, and can be determined by casual (a given stretch of nucleotides randomly being present multiple times within a genome) or structural events (such as duplications or copy number variants). The predetermination of mappability scores, computed throughout the genome, can then be of great help in predicting our ability to obtain a low or zero coverage in given regions, as well as to avoid incorrect calls in positions where non-homologous reads are assigned because of high mappability scores.

Nonetheless, even after considering all the parameters that can influence our assemblies, most of the downstream analyses rely on the knowledge of the gene locations and functions. Several databases provide extended availability of gene annotation datasets. However, two limitations can seriously affect their usefulness: the number and the quality of the annotated transcripts. In fact, beside a few well studied or model organisms, a good annotation set could be totally lacking for the target species, requiring time-expensive computational reconstructions or extensive mRNA sequencing. In our case, three different sources were available through common databases. However, some of them revealed a very limited number of transcripts (as in the case of RefGene dataset) or an extremely high number of transcripts not meeting basic conditions, such as having a start and a stop codon, or being multiple of 3 bp (e.g. for the Ensembl gene set). Of course, most of the annotation pipelines rely on multi-way comparisons across genomes (Derrien et al. 2011) to identify highly conserved regions that could underlie the presence of functional elements, as well as on the presence of known gene-delimiting sequences (TATA boxes, splice sites, starting and ending codons, etc.). However, the high number of transcripts make it almost impossible (or at least time-expensive) to manually curate the databases, opening the way to incorrect coordinate

assignments, even for well studied genes, such as for OPCML (Fig. 41) in Ensembl, which does not have any start codon in the first exon, conversely showing premature stop codons along its sequence. Several studies can be (and have been) published based on these annotation sources, implying incorrect results and misleading conclusions.

Therefore, a careful verification should be performed when using large annotation datasets, even by simple controls as in our case. In the end, we joined from three different sources a large assembly of presumably well-annotated transcripts, corresponding to a total number of genes (18,782) lower but comparable with what described in humans (*ca.* 22,000, EnSEMBL build 26), and not taking into account the more than 17,000 transcripts that could be missing our criteria in the boxer, but being anyway correctly annotated and still functional in the other *taxa*.

Nonetheless, the constant update of genomic information by the community (e.g. the recent canFam3 assembly) makes it difficult to keep up the pace with the new data, requiring additional efforts in order to access and work on the state-of-the-art resources.

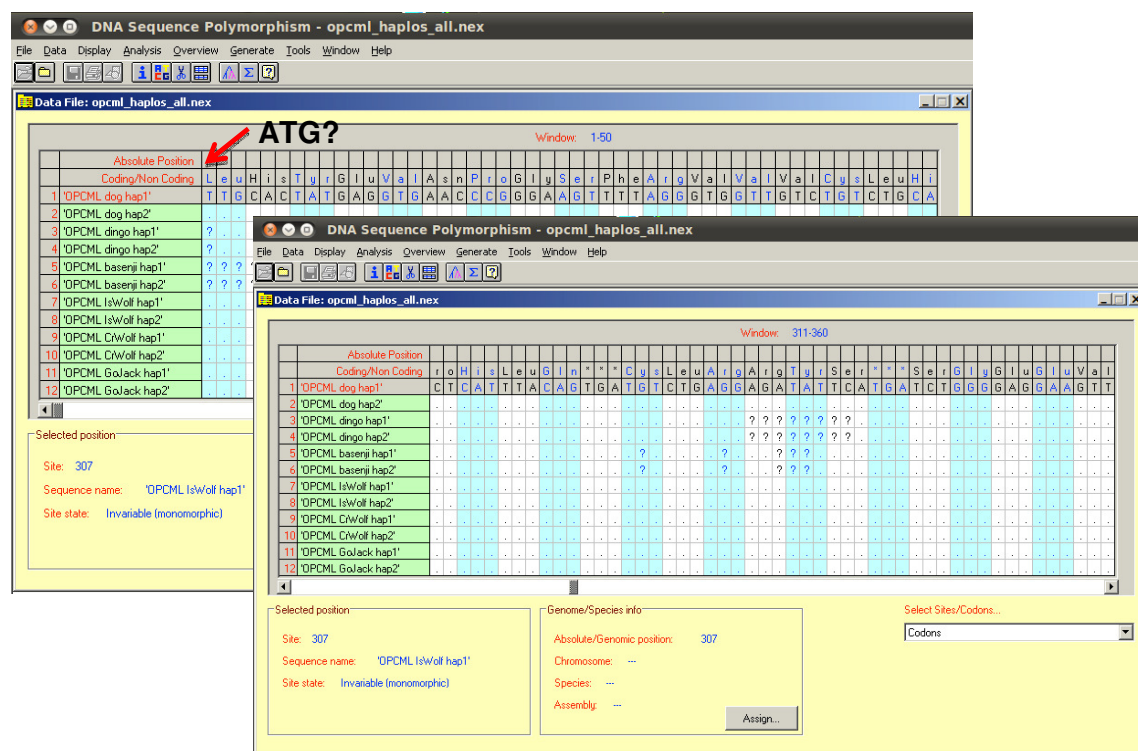


Figure 41: Nucleotide and amino acid sequence of the OPCML gene, as retrieved by the boxer genome based on annotation information available in Ensembl, and visualized in DNASP. The absence of a start codon and the presence of multiple premature stop codons suggest errors in the coordinate assignment.

Despite being only a portion of all the possible investigations to be performed on a whole genome study, analyses on the functional regions are of key interest in order to understand important differences between domestic and wild canids, which can directly affect the protein sequence and functionality, or their level of expression.

From our different pipelines, we were able to identify a number of candidate hits that could be discriminatory between the two groups, thus possibly related to domestication.

What is worth noticing is, first of all, the relatively small number of candidates we ended up identifying, even starting from an impressive amount of sequence data. This means that, on the one side, applying strict filtering conditions in order to retain only high quality data can limit the quantity of information that can be successfully analyzed; on the other side, this can also mean that a multi-genome approach, including both two ancient dog breeds and an ancestral outgroup, the jackal, probably helped in ruling out *taxon*-specific variants, therefore adding support to the variants identified. Anyhow, the numbers of dog-specific hits we found are in the same order of magnitude as the ones detected from the comparison of modern human and Neandertal genomes (Green et al. 2010), although in that case the population divergence time is about ten times what expected between dogs and wolves.

Among the non-synonymous mutations showing the largest predicted effects at the protein level, RNF12 is the most interesting one.

In the somatic cells of female placental mammals (Barakat et al. 2011, Jonkers et al. 2009), one of the two X chromosomes (but not the other one) is randomly inactivated to minimize sex-related dosage differences of *ca.* 1000 X-encoded genes, after a temporary imprinting-dependent inactivation step. The X chromosome inactivation (XCI) is mainly operated by RNF12, and triggered by the nuclear concentration of one or more X-encoded XCI-activators, such as the Xist gene product, and the removal of the gene can result lethal in female offspring (Shin et al. 2010). Additionally, RNF12 is also linked to a network regulating the cellular pluripotency (Navarro et al. 2011).

Acting as a regulator of the X chromosome dose inactivation, and interacting with developmental networks and estrogen receptors, mutations on RNF12 sequence are likely to produce primary effects, although expression studies on the two allelic variants identified in this study (in domestic *vs.* wild canids) should be performed. Interestingly, it has been shown that a larger number of females compared to males (Sundqvist et al. 2006) contributed to the dog gene pool, potentially allowing for a faster spread of X-carried mutations in the dog population during domestication.

However, other genes emerging from the enrichment analysis on non-synonymous mutations discriminating domestic *vs.* wild canids can lead to interesting implications.

APOB and PPP1R3A, which respectively affect the levels of hypercholesterolemia and are involved in obesity and diabetes, can have played a role in the change of dietary habits connected with the shift from living in the wild to the association with human settlements - and possibly feeding on leftovers: the alleles that may have provided an initial adaptive benefit in exploiting new food sources, could have also turned into risky variants when the food intake in pet dogs recently blasted, leading to diseases that are common in modern breeds, such as diabetes.

Similarly, mammary tumors are the most common tumor type in female dogs (Yoshikawa et al. 2008), constituting approximately 40% to 50% of all tumors in female dogs. FANCD2 is one of the genes responsible for Fanconi anemia, a rare autosomal disorder also linked to tumor or leukemia susceptibility, and cellular hypersensitivity to DNA cross-linking agents. Therefore, mutations in this gene could potentially enhance the predisposition for a number of tumors in dogs, whose incidence of mammary tumors is higher than in any other species (Yoshikawa et al. 2008). Parallel, ANLN gene, which showed a non-canonical splice site in domestic but not in wild canids, is also known to be implicated in the onset of a number of tumors (Shimizu et al. 2007, Suzuki et al. 2005, Tamura et al. 2007).

However, from a morphological point of view, a few traits can uniquely differentiate dogs and wolves. One of them is the rear limb dewclaw, an example of hind-limb-specific preaxial polydactyly that is common in dogs (Ruvinsky and Sampson 2001), but not in wolves (Ciucci et al. 2003). Interestingly, EVC2 is associated to Ellis-van Creveld Syndrome, which causes polydactyly, heart diseases, short-limbed dwarfism, cleft palate, and malformation of the wrist bones, most of which are common in large dog breeds such as Bernese and St. Bernard (Galis et al. 2001), presumably with pleiotropic effects; therefore, although not being the causal mutation, the variant we identified could represent an intermediate state in the development of these conditions in dogs, but not in wolves. However, other genes have been demonstrated to be implicated in polydactyly in dogs, such as LMBR1 (Park et al. 2008), therefore follow-up studies should be performed to evaluate the specific role of EVC2 in dogs.

Other behavioral differences between wolves and dogs are related to daily patterns of activity. Whereas in most parts of the world (Packard 2003) wolves are mainly nocturnal, dogs are most active during the day, adhering to human activity patterns. From the analysis of mutations in UTRs, we found a differential SNV in the CLOCK gene (King et al. 1997),

which has been demonstrated to strongly influence the circadian rhythms in a number of species. Of course, a single variant in the UTR sequence does not necessarily imply a different level of expression of the coded protein, but signals of selection in these untranslated regions (data not shown) appear to be particularly strong compared to both exonic and intronic mutations, suggesting their effective role in the dog differentiation from wolves.

Several other genes with differences in the UTR sequences are related to brain functions.

LRRN3 is highly expressed in brain and in immune system cells; in humans, it has been shown to be the gene whose expression is mostly correlated with aging at all (Harries et al. 2011). It is true that in captivity wolves can live almost the double than in the wild, reaching up to 12-14 years of age. However, what is rare in nature is common in the case of pet dogs, especially for small breeds, which can often live up to 15 years or more. Therefore, the maintenance of good levels of cerebral and immune functions with age could have been positively selected during domestication. In the same gene family, LRRN6 (also known as LINGO2) has been deeply studied for its role in neuronal activity, ganglia development (Bryan et al. 2008) and Parkinson disease (Vilariño-Güell et al. 2010). In particular, a sister gene, LINGO1, has been strongly associated to the efficiency of myelination of the axons by the oligodendrocytes in the central nervous system, therefore playing a key role in the speed on signal transduction in the brain (Mi et al. 2005). Parallel, NLGN1 encodes neuronal cell-adhesion proteins, involved in formation and remodeling of central nervous system synapses, which play a key role in regulating and refining nervous signals (Sudhof 2008); it has been associated to autism disorders (Millson et al. 2011) and anxiety behaviors following early life stress (Benekareddy et al. 2011). An additional brain-related genes with a differential allele in UTRs is SLC6A15, a neurotransmitter transporter (Takanaga et al. 2005) highly expressed in brain, whose variants have been linked to major depression (Kohli et al. 2011), although its knocking-out does not strongly affect the behavior and functions in mice (Drgonova et al. 2007).

Such a number of candidate genes related to neuronal functions is not surprising. Gene expression studies (Saetre et al. 2004) revealed a strong difference in the expression levels of a series of brain-related genes, with particular differences in the amygdala (where the expression levels are highly conserved in wild canids, but widely variable in dogs) and hypothalamus. The authors suggest that a limited number of genes with multiple functions related to cognition and behavior could have been strongly selected at the expression level during domestication.

Beside having a role during neuronal apoptosis (Imaizumi et al. 1999), the Harakiri (or DP5) gene, whose CDS was differentially delimited in domestic *vs.* wild canids, also has an important role in embryo development (Jurisicova et al. 2003); mice knocked out for this gene (Imaizumi et al. 2004) were viable, but showed delayed neuronal cell death, therefore potentially implying large effects in brain cell survival and regulation.

Lastly, the top difference at the splice site level falls in the TRPS1 gene, which codes for a transcription factor associated to the Trichorhinophalangeal syndrome type1 (Kunath et al. 2002, Malik et al. 2001, Malik et al. 2002); this syndrome is denoted by a series of facial features and skeletal abnormalities, such as sparse scalp hair, a bulbous tip of the nose, a long philtrum, a thin upper vermilion border, protruding ears, short stature, brachydactyly, cone-shaped epiphyses in the phalanges and hip malformations (Fig. 42).

Interestingly, the shape of the facial portion of the skull, although largely variable in dogs, is one of the diagnostic traits differentiating dogs and wolves: mainly, by the presence of the ‘stop’ in dogs, as well as a shorter and larger palate causing a different teeth placement (Germonpre et al. 2009, Nowak 2003, Ruvinsky and Sampson 2001).

Also hair differences, although their genetic determinants have been well described (Cadieu et al. 2009), are another trait distinguishing dogs and wolves, and dissimilarities in the ear shape are one of the main characters differentiating most of the domestic breeds from the wild species (Trut et al. 2009).

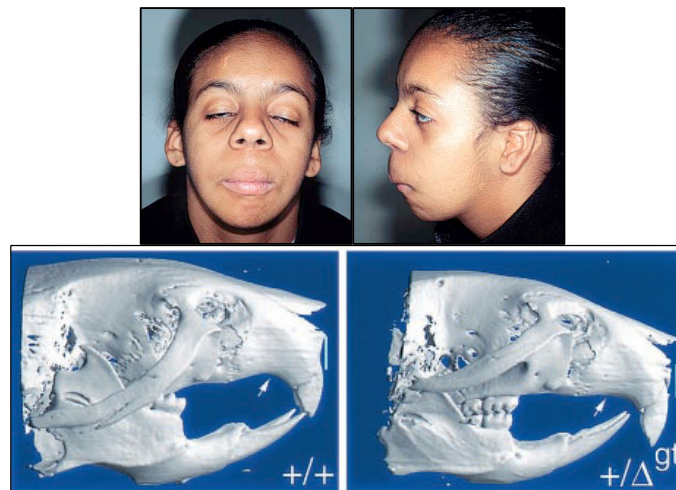


Figure 42: Above, phenotypic effects associated to the Trichorhinophalangeal syndrome in an affected patient. The abnormally long distance between nose and lips, and large bulbous nose, are apparent (from Shin and Chang 2001). Below, differences in skull shape between a wild-type and a mutant mouse where TRPS1 gene is altered, showing a severely different morphology of the palatal arch (from Malik et al. 2002). Resembling discrepancies are visible between wolf and dog skulls.

Although several of these traits could be determined by the multiple action of several genes (Boyko et al. 2010), it is intriguing to look for follow-up studies on TRSP1, since a different exon combination in its protein, as possible from the presence of a mutated splice site, would potentially be linked to many of the traits that differentiate wolves and dogs from a morphological point of view.

Of course, most of the points hereby discussed can only be speculative, as long as complete results will be obtained and follow-up studies will be performed:

- 1) first of all, by confirming the presence of the mutations discovered through NGS in candidate genes by Sanger re-sequencing;
- 2) by enlarging the panel of wolves and dogs to be analyzed, in order to differentiate random clustering of the alleles among our *taxa* from real differences in the allele frequencies or fixation;
- 3) by performing transcriptome and gene expression studies, in order to actually demonstrate, at the protein level, the link between the discovered variants and their phenotypic (*latu sensu*) correlates.

However, the ones just described remain nonetheless interesting findings to be validated, and also to be compared to their homologous counterparts in other domesticated species whose genomes are already available.

In any case, the implications of this multi-genome project are apparent: unravel the genetic basis of dog-specific traits, and some of the diseases correlated to them, is of primary importance both for a better understanding of the effects of domestication and, in the next future, also for ensuring a better quality of life to our 'best friends'.

Further data coming from selection scans, comparisons of dog/wolf diversity and divergence across the genome, and demographic models combined to post-diverge gene flow (all of them being performed at the time of writing), will hopefully help to obtain a more complete overview on the whole dog domestication process, including its timing and location. The identification of hundreds of thousands of genetic markers specific for wolves (and related species such as the golden jackal), both at the functional and the neutral level, can be soon applied to the study of entire wild populations, helping to resolve questions such as gene introgression and hybridization; likewise, it will be possible to better investigate the population size and history of endangered populations, such as the Mexican and -at a lower extent- the Italian wolf, reconstructing past bottlenecks or expansions. Similarly, it will be easier to identify the traces of local adaptations and selective pressures acting on the genomes

of endangered species, helping to develop targeted strategies for the preservation of the diversity at the most important genes targeted by selection, as in the case of the MHC.

However, beside scientific and conservation purposes, the project will represent for every person interested in this topic a further step in understanding how, some thousand years ago, two of the most successful species on the planet -*Homo sapiens* and *Canis lupus*-, decided to cross and link each other's path, possibly coevolving and changing forever, at least to a certain extent, their own evolutionary destiny.

4. Conclusions

The scientist's work is surely a privilege. Having the possibility to study so closely some of the most interesting topics in biology (but it would be the same in any other field, from philosophy to archaeology), and sometimes being able to add a little brick to the extant knowledge, is a reward *per se*.

However, if it is true that we are standing on the giants' shoulders, nowadays we seriously run the risk to feel dizzy.

On the one side, because the findings and the cultural level of many scientists of the past, that with a simple pencil were able to trace down the paths of selection and adaptation like Darwin, or to flawlessly model the frameworks of evolution like Fisher, are almost impossible to parallel or even to come close.

On the other side, because the incredibly fast scientific advances of the current era can easily lead a 'simple' naturalist to get lost in a world of essential bioinformatics, fundamental statistics and revolutionary technology.

Therefore, educational programs should make an effort to update their ways of conveying a growing quantity and quality of information, by providing both the practical tools and theoretical background to allow us to ask the right questions - that should be the most relevant for the 21st century society – and to know where to look for the solutions.

So that, even from the height of a giant, the landscape of science can still appear to us as wide, pristine and beautiful.

5. Bibliography

- Adzhubei, I. A, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev. 2010. A method and server for predicting damaging missense mutations. *Nature methods* 7:248-9.
- Aguilar, A., G. Roemer, S. Debenham, M. Binns, D. Garcelon, and R. K. Wayne. 2004. High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. *Proceedings of the National Academy of Sciences of the United States of America* 101:3490-4.
- Akey, J.M., A. L. Ruhe, D. T. Akey, A. K. Wong, C.F. Connelly, et al. 2010. Tracking footprints of artificial selection in the dog genome. *PNAS USA* 107: 1160–1165.
- Alföldi, J., F. Di Palma, M. Grabherr, C. Williams, L. Kong, E. Mauceli, P. Russell, C. B. Lowe, R. E. Glor, J. D. Jaffe, D. a. Ray, S. Boissinot, A. M. Shedlock, C. Botka, T. A. Castoe, J. K. Colbourne, M. K. Fujita, R. G. Moreno, B. F. ten Hallers, D. Haussler, A. Heger, D. Heiman, D. E. Janes, J. Johnson, P. J. de Jong, M. Y. Koriabine, M. Lara, P. A. Novick, C. L. Organ, S. E. Peach, S. Poe, D. D. Pollock, K. de Queiroz, T. Sanger, S. Searle, J. D. Smith, Z. Smith, R. Swofford, J. Turner-Maier, J. Wade, S. Young, A. Zadissa, S. V. Edwards, T. C. Glenn, C. J. Schneider, J. B. Losos, E. S. Lander, M. Breen, C. P. Ponting, and K. Lindblad-Toh. 2011. The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* 1-53.
- Allendorf, F. W., P. Hohenlohe, and G. Luikart. 2010. Genomics and the future of conservation genetics. *Nature reviews. Genetics* 11:697-709.
- Altobello, G. 1921. Mammiferi IV, Carnivori. In: *Fauna d’Abruzzo e Molise*. Campobasso, Italy, pp 38–45.
- Altshuler, D., V. J. Pollara, C. R. Cowles, W. J. van Etten, J. Baldwin, L. Linton, and E. S. Lander. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407:513-516.
- Alvarez-Garcia, I., and E. A. Miska. 2005. MicroRNA functions in animal development and human disease. *Development (Cambridge, England)* 132:4653-62.
- Anderson, T. M., B. M. vonHoldt, S. I. Candille, M. Musiani, C. Greco, D. R. Stahler, D. W. Smith, B. Padhukasahasram, E. Randi, J. A. Leonard, C. D. Bustamante, E. A. Ostrander, H. Tang, R. K. Wayne, and G. S. Barsh. 2009. Molecular and evolutionary history of melanism in North American gray wolves. *Science (New York, N.Y.)* 323:1339-43.
- Angles, J M, L J Kennedy, and N. C. Pedersen. 2005. Frequency and distribution of alleles of canine MHC-II DLA-DQB1, DLA-DQA1 and DLA-DRB1 in 25 representative American Kennel Club breeds. *Tissue antigens* 66:173-84.
- Apollonio, M, L. Mattioli, M. Scandura, L. Mauri, A. Gazzola, and E. Avanzinelli. 2004. Wolves in the Casentinesi Forests: insights for wolf conservation in Italy from a protected area with a rich wild prey community. *Biological Conservation* 120:249-260.

- Babik, W., P. Taberlet, M. J. Ejsmond, and J. Radwan. 2009. New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. *Molecular ecology resources* 9:713-9.
- Barakat, T. S., N. Gunhanlar, C. G. Pardo, E. M. Achame, M. Ghazvini, R. Boers, A. Kenter, E. Rentmeester, J. A. Grootegoed, and J. Gribnau. 2011. RNF12 activates Xist and is essential for X chromosome inactivation. *PLoS genetics* 7:e1002001.
- Barber, R. M., S. J. Schatzberg, J. J. Corneveaux, A. N. Allen, B. F. Porter, J. J. Pruzin, S. R. Platt, M. Kent, and M. J. Huentelman. 2011. Identification of risk Loci for necrotizing meningoencephalitis in pug dogs. *The Journal of heredity* 102 Suppl :S40-6.
- Barja, I., F. J. de Miguel, and F. Bárcena. 2004. The importance of crossroads in faecal marking behaviour of the wolves (*Canis lupus*). *Die Naturwissenschaften* 91:489-92.
- Bekoff, M., and J. Pierce. 2009. *Wild justice. The moral lives of animals*. Chicago University Press.
- Belov, K., J. E. Deakin, A. T. Papenfuss, M. L. Baker, S. D. Melman, et al. 2006. Reconstructing an Ancestral Mammalian Immune Supercomplex from a Marsupial Major Histocompatibility Complex. *PLoS Biol* 4:e46. doi:10.1371/journal.pbio.0040046.
- Benekareddy, M., K. C. Vadodaria, A. R. Nair, and V. A. Vaidya. 2011. Postnatal Serotonin Type 2 Receptor Blockade Prevents the Emergence of Anxiety Behavior, Dysregulated Stress-Induced Immediate Early Gene Responses, and Specific Transcriptional Changes that Arise Following Early Life Stress, *Biological Psychiatry* 70:1024-1032.
- Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. 2011. GenBank. *Nucleic acids research* 39:D32-7.
- Berger, D. S., W. A. Hogge, M. M. Barmada, and R. E. Ferrell. 2010. Comprehensive analysis of HLA-G: implications for recurrent spontaneous abortion. *Reproductive Science* 17:331-8.
- Berggren, K. T., and J. M Seddon. 2005. MHC promoter polymorphism in grey wolves and domestic dogs. *Immunogenetics* 57:267-72.
- Berggren, K. T., and J. M. Seddon. 2008. Allelic combinations of promoter and exon 2 in DQB1 in dogs and wolves. *Journal of molecular evolution* 67:76-84.
- Bernatchez, L, and C. Landry. 2003. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *Journal of evolutionary biology* 16:363-77.
- Björnerfeldt, Susanne, M. T. Webster, and C. Vilà. 2006. Relaxation of selective constraint on dog mitochondrial DNA following domestication. *Genome research* 16:990-4.
- Boitani, L. 1984. Genetic considerations on wolf conservation in Italy. *Bollettino di Zoologia* 51:367-373.
- Boitani, L. 2000. Action plan for the conservation of wolves (*Canis lupus*) in Europe. *Nature and environment*.

- Boitani, L. 2003. Wolf conservation and recovery. In: Mech L. D. and L. Boitani (Eds.), *Wolves: behavior, ecology and conservation*. The University of Chicago Press, pp 317-340.
- Bos, D. H., S. M. Turner, and J. A. Dewoody. 2007. Haplotype inference from diploid sequence data: evaluating performance using non- neutral MHC sequences. *Hereditas* 144:228–234.
- Bossdorf, O., C. L. Richards, and M. Pigliucci. 2008. Epigenetics for ecologists. *Ecology letters* 11:106-15.
- Boyko, A. R, R. H. Boyko, C. M. Boyko, H. G. Parker, M. Castelhana, L. Corey, J. D. Degenhardt, A. Auton, M. Hedimbi, R. Kityo, E. A. Ostrander, J. Schoenebeck, R. J. Todhunter, P. Jones, and C. D. Bustamante. 2009. Complex population structure in African village dogs and its implications for inferring dog domestication history. *PNAS USA* 106:13903-8.
- Boyko, A. R, P. Quignon, Lin Li, J. J. Schoenebeck, J. D. Degenhardt, K. E. Lohmueller, K. Zhao, A. Brisbin, H. G. Parker, B. M. vonHoldt, M. Cargill, A. Auton, A. Reynolds, A. G. Elkhoun, M. Castelhana, D. S. Mosher, N. B. Sutter, G. S. Johnson, J. Novembre, M. J. Hubisz, A. Siepel, Robert K Wayne, C. D. Bustamante, and E. A. Ostrander. 2010. A simple genetic architecture underlies morphological variation in dogs. *PLoS biology* 8:e1000451.
- Boyko, A. R. 2011. The domestic dog: man’s best friend in the genomic era. *Genome biology* 12:216.
- Braslavsky, I., B. Hebert, E. Kartalov, and S. R. Quake. 2003. Sequence information can be obtained from single DNA molecules. *PNAS USA* 100:3960–3964
- Breitenmoser, U. 1998. Large predators in the Alps: The fall and rise of man’s competitors. *Biological Conservation* 83:279-289.
- Brouwer, L., I. Barr, M. van de Pol, T. Burke, J. Komdeur, and D. S. Richardson. 2010. MHC-dependent survival in a wild population: evidence for hidden genetic benefits gained through extra-pair fertilizations. *Molecular ecology* 19:3444-55.
- Burbano, H. A., E. Hodges, R. E. Green, A. W. Briggs, J. Krause, M. Meyer, J. M. Good, T. Maricic, P. L. F. Johnson, Z. Xuan, M. Rooks, A. Bhattacharjee, L. Brizuela, F. W. Albert, M. de la Rasilla, J. Fortea, A. Rosas, M. Lachmann, G. J. Hannon, and S. Pääbo. 2010. Targeted investigation of the Neandertal genome by array-based sequence capture. *Science (New York, N.Y.)* 328:723-5.
- Bushati, N., and S. M. Cohen. 2007. microRNA functions. *Annual review of cell and developmental biology* 23:175-205.
- Cadiou, E., M. W. Neff, P. Quignon, K. Walsh, K. Chase, H. G. Parker, B. M. Vonholdt, A. Rhue, A. Boyko, A. Byers, A. Wong, D. S. Mosher, A. G. Elkhoun, T. C. Spady, C. André, K Gordon Lark, M. Cargill, C. D. Bustamante, R. K. Wayne, and E. A. Ostrander. 2009. Coat variation in the domestic dog is governed by variants in three genes. *Science (New York, N.Y.)* 326:150-3.

- Candille, S. I., C. B. Kaelin, B. M. Cattanach, B. Yu, D. A. Thompson, M. A. Nix, J. A. Kerns, S. M. Schmutz, G. L. Millhauser, and G. S. Barsh. 2007. A b-defensin mutation causes black coat color in domestic dogs. *Science* 318:1418-23.
- Caniglia, R., E. Fabbri, M. Galaverni, and E. Randi. In prep. Genetic structure and pack dynamics in an expanding wolf population.
- Caniglia, R., E. Fabbri, C. Greco, M. Galaverni, and E. Randi. 2010. Forensic DNA against wildlife poaching: identification of a serial wolf killing in Italy. *Forensic science international. Genetics* 4:334-8.
- Caniglia, R., E. Fabbri, C. Greco, M. Galaverni, L. Manghi, L. Boitani, A. Sforzi, and E. Randi. Submitted. Black coats in an admixed wolf x dog pack. Is melanism an indicator of hybridisation?
- Caniglia, R., E. Fabbri, C. Greco, and E. Randi. 2010. Non-invasive genetic monitoring of the wolf (*Canis lupus*) population in Emilia-Romagna, Proceeding of the Conference: Scientific Research and Management for Wolf Conservation in Italy, Ministry of Environment.
- Cercueil, A., E. Bellemain, and S. Manel. 2002. Parente: computer program for parentage analysis. *Journal of Heredity* 93:458-459.
- Cha, P. C., A. Takahashi, N. Hosono, S. K. Low, N. Kamatani, M. Kubo, and Y. Nakamura. 2011. A genome-wide association study identifies three loci associated with susceptibility to uterine fibroids. *Nature Genetics* 43:447-450.
- Chen, W.K., J. D. Swartz, L. J. Rush, and C. E. Alvarez. 2009. Mapping DNA structural variation in dogs. *Genome research* 19:500-9.
- Ciucci, P., and L. Boitani. 1991. Viability assessment of the Italian wolf and guidelines for the management of the wild and a captive population. *Ricerche di Biologia della Selvaggina. Istituto Nazionale di Biologia della Selvaggina* 89:1-58.
- Ciucci, P., and L. Boitani. 1998. *Il Lupo. Elementi di biologia, gestione e ricerca. Istituto Nazionale della Fauna Selvatica, Documenti Tecnici n. 23.*
- Ciucci, P., G. Chapron, V. Guberti, and L. Boitani. 2007. Estimation of mortality parameters from (biased) samples at death: are we getting the basics right in wildlife field studies? A response to Lovari et al. (2007). *Journal of Zoology* 273:125-127.
- Ciucci, P., V. Lucchini, L. Boitani, and E. Randi. 2003. Dewclaws in wolves as evidence of admixed ancestry with dogs. *Canadian journal of zoology* 81:2077-2081.
- Ciucci, P., W. Reggioni, L. Maiorano, and L. Boitani. 2009. Long-Distance Dispersal of a Rescued Wolf From the Northern Apennines to the Western Alps. *The Journal of Wildlife Management* 73:1300-1306. doi:10.2193/2008-510
- Clarke, B. C., and D. R. S. Kirby. 1966. Maintenance of histocompatibility polymorphism. *Nature* 211:999-1000.

- Cokus, S. 2011. Introduction to new generation sequencing. Oral presentation at UCLA conference series, Los Angeles, USA.
- Consuegra, S., and C. G. de Leaniz. 2008. MHC-mediated mate choice increases parasite resistance in salmon. *Proceedings. Biological sciences / The Royal Society* 275:1397-403.
- Coppinger, R., and L. Coppinger. 2001. *Dogs: A Startling New Understanding of Canine Origin, Behavior & Evolution*. New York, NY: Scribner.
- Cornuet, J. M., and G. Luikart. 1996. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* 144:2001–2014.
- Crockford, S. J. 2006. *Rhythms of Life: Thyroid Hormone and the Origin of Species*. Victoria, BC: Trafford.
- Cruz, F., C. Vilà, and M. T. Webster. 2008. The legacy of domestication: accumulation of deleterious mutations in the dog genome. *Molecular biology and evolution* 25:2331-6.
- Dawson, A.G. 1996. *Ice Age Earth. Late Quaternary Geology and Climate*. New York, NY: Routledge.
- Delibes, M. 1990. Status and conservation needs of the wolf in the Council of Europe member States. 47:1–46, *Nature and Environment Series*, Strasbourg.
- Derrien, T., A. Vaysse, C. André, and C. Hitte. 2011. Annotation of the domestic dog genome sequence: finding the missing genes. *Mammalian genome* doi:10.1007/s00335-011-9372-0.
- Dionne, M., K. M. Miller, J. J. Dodson, F. Caron, and L. Bernatchez. 2007. Clinal variation in MHC diversity with temperature: evidence for the role of host-pathogen interaction on local adaptation in Atlantic salmon. *Evolution* 61:2154-64.
- Dionne, M., K. M. Miller, J. J. Dodson, and L. Bernatchez. 2009. MHC standing genetic variation and pathogen resistance in wild Atlantic salmon. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 364:1555-65.
- Ditchkoff, S. S., R. L. Lochmiller, R. E. Masters, S. R. Hooper, and R.A. Van Den Bussche. 2001. Major-histocompatibility-complex-associated variation in secondary sexual traits of white-tailed deer (*Odocoileus virginianus*): evidence for good-genes advertisement. *Evolution* 55: 616–625.
- Doherty, P. C., and R. M. Zinkernagel. 1975. Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. *Nature* 256:50–52.
- Dreger, D. L., and S. M. Schmutz. 2011. A SINE insertion causes the black-and-tan and saddle tan phenotypes in domestic dogs. *The Journal of heredity* 102 Suppl :S11-8.
- Drgonova, J., Q. R. Liu, F. S. Hall, R. M. Krieger, and G. R. Uhl. 2007. Deletion of v7-3 (SLC6A15) transporter allows assessment of its roles in synaptosomal proline uptake,

leucine uptake and behaviors, *Brain Research* 1183:10-20, doi:10.1016/j.brainres.2007.09.001.

- Drögemüller, C., E. K. Karlsson, M. K. Hytonen, M. Perloski, G. Dolf, et al. 2008. A mutation in hairless dogs implicates FOXI3 in ectodermal development. *Science* 321:1462.
- Drummond, A. J., B. Ashton, S. Buxton, M. Cheung, A. Cooper, C. Duran, M. Field, J. Heled, M. Kearse, S. Markowitz, R. Moir, S. Stones-Havas, S. Sturrock, T. Thierer, and A. Wilson. 2011. Geneious v5.4, available from <http://www.geneious.com/>
- Duchesne, P., C. Étienne, and L. Bernatchez. 2006. PERM: a computer program to detect structuring factors in social units. *Molecular ecology notes* 6:965-967.
- Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–138.
- Eizaguirre, C., S. E. Yeates, T. L. Lenz, M. Kalbe, and M. Milinski. 2009. MHC-based mate choice combines good genes and maintenance of MHC polymorphism. *Molecular ecology* 18:3316-29.
- Ejsmond, M. J., and J. Radwan. 2011. MHC diversity in bottlenecked populations: a simulation model. *Conservation Genetics* 12:129-137.
- Ekblom, R., S. A. Saether, P. Fiske, J. A. Kålås, and J. Höglund. 2010. Balancing selection, sexual selection and geographic structure in MHC genes of Great Snipe. *Genetica* 138:453-61.
- Ellis, S. A., R. E. Bontrop, D. F. Antczak, K. Ballingall, C. J. Davies, J. Kaufman, L. J. Kennedy, J. Robinson, D. M. Smith, M. J. Stear, R. J. M. Stet, M. J. Waller, L. Walter, and S. G. E. Marsh. 2006. ISAG/IUIS-VIC Comparative MHC Nomenclature Committee report, 2005. *Immunogenetics* 57:953-8.
- Evans, M. L., M. Dionne, K. M. Miller, and L. Bernatchez. 2012. Mate choice for major histocompatibility complex genetic divergence as a bet-hedging strategy in the Atlantic salmon (*Salmo salar*). *Proceedings of the Royal Society B: Biology* 279:379-386.
- Fabbri, E., C. Miquel, V. Lucchini, A. Santini, R. Caniglia, C. Duchamp, J. M. Weber, B. Lequette, F. Marucco, L. Boitani, L. Fumagalli, P. Taberlet, and E. Randi. 2007. From the Apennines to the Alps: colonization genetics of the naturally expanding Italian wolf (*Canis lupus*) population. *Molecular ecology* 16:1661-71.
- Falush, D., M. Stephens, and J. K. Pritchard. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164:1567–1587.
- Feder, M. E., and T. Mitchell-Olds. 2003. Evolutionary and ecological functional genomics. *Nature reviews. Genetics* 4:651-7.
- Flicek, P., and E. Birney. 2010. Sense from sequence reads: methods for alignment and assembly. *Nature Methods Supplement* 6.

- Fliegner, R. A., S. A. Holloway, S. Lester, C. a McLure, and R. L. Dawkins. 2008. Evaluation of the class II region of the major histocompatibility complex of the greyhound with the genomic matching technique and sequence-based typing. *Tissue antigens* 72:131-6.
- Fox, J. 2005. The R Commander: A basic-statistics graphical user interface to R. *Journal of Statistical Software*, 19:1-42.
- Francino, O, M. Amills, and a Sánchez. 1997. Canine Mhc DRB1 genotyping by PCR-RFLP analysis. *Animal genetics* 28:41-5.
- Francisco, L.V., A. A. Langston, C. S. Mellersh, C. L. Neal, and E. A. Ostrander. 1996. A class of highly polymorphic tetranucleotide repeats for canine genetic mapping. *Mammalian Genome* 7:359–362.
- Frankham, R. 2010. Challenges and opportunities of genetic approaches to biological conservation. *Biological conservation* 143:1919-1927.
- Fredholm, M., and A. K. Winterø. 1995. Variation of short tandem repeats within and between species belonging to the Canidae family. *Mammalian genome* 6:11–18.
- Fujita, P. A., B. Rhead, A. S. Zweig, A. S. Hinrichs, D. Karolchik, M. S. Cline, M. Goldman, G. P. Barber, H. Clawson, A. Coelho, M. Diekhans, T. R. Dreszer, B. M. Gardine, R. a Harte, J. Hillman-Jackson, F. Hsu, V. Kirkup, R. M. Kuhn, K. Learned, C. H. Li, L. R. Meyer, A. Pohl, B. J. Raney, K. R. Rosenbloom, K. E. Smith, D. Haussler, and W. J. Kent. 2011. The UCSC Genome Browser database: update 2011. *Nucleic acids research* 39:D876-82.
- Fuller, T. K., L. D. Mech, and J. F. Cochrane. 2003. Wolf population dynamics. In: Mech L. D. and L. Boitani (Eds.), *Wolves: Behaviour, Ecology and Conservation*. The University of Chicago Press, pp. 161-191.
- Galaverni, M., D. Palumbo, E. Fabbri, R. Caniglia, C. Greco, and E. Randi. 2012. Monitoring wolves (*Canis lupus*) by non-invasive genetics and camera trapping: a small-scale pilot study. *European Journal of Wildlife Research* 58:47-58.
- Galibert, F., and C. André. 2008. The dog: A powerful model for studying genotype-phenotype relationships. *Comparative biochemistry and physiology. Part D, Genomics & proteomics* 3:67-77.
- Galis, F., J. J. M. van Alphen, and J. A. J. Metz. 2001. Why five fingers? Evolutionary constraints on digit numbers. *Trends in Ecology & Evolution* 16:637-646.
- Galton, F. 1865. The first steps towards the domestication of animals. *Transactions Ethnological Society London* 3:122–138.
- Garcia-Muro, E., M. P. Aznar, C. Rodellar, and P. Zaragoza. 1997. Sex specific PCR/RFLPs in the canine ZFX/ZFY loci, *Animal Genetics* 28:156.
- Geffen, E., M. E. Gompper, J. L. Gittleman, H. K. Luh, D. W. MacDonald, and R. K. Wayne. 1996. Size, life-history traits, and social organization in the canidae: a reevaluation. *American Naturalist* 147:140–160.

- Geffen, E., M. Kam, R. Hefner, P. Hersteinsson, A. Angerbjörn, L. Dalèn, E. Fuglei, K. Norèn, J. R. Adams, J. Vucetich, T. J. Meier, L. D. Mech, B. M. Vonholdt, D. R. Stahler, and R. K. Wayne. 2011. Kin encounter rate and inbreeding avoidance in canids. *Molecular ecology* 5348-5358.
- Genome 10K Community of Scientists. 2009. Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10,000 Vertebrate Species. *Journal of Heredity* 100: 659-674.
- Genovesi, P. 2002. Piano d'azione nazionale per la conservazione del Lupo (*Canis lupus*).
- Genovesi, P., and E. Dupré. 2000. Strategia nazionale di conservazione del lupo (*Canis lupus*): indagine sulla presenza e la gestione di cani vaganti in Italia. *Biologia e Conservazione della Fauna* 104:1-36.
- Germonpre, M., M. Sablin, R. Stevens, R. Hedges, M. Hofreiter, M. Stiller, and V. Despres. 2009. Fossil dogs and wolves from Palaeolithic sites in Belgium, the Ukraine and Russia: osteometry, ancient DNA and stable isotopes. *Journal of Archaeological Science* 36:473-490.
- Giger, U., D. R. Sargan, E. and A. McNeil. 2006. Breed-specific hereditary diseases and genetic screening. In: Ostrander, E.A., U. Giger, K. Lindbladh-Toh (Eds.) *The Dog and its Genome*. Monograph, vol. 44. Cold Spring Harbor Laboratory Press, pp. 249-290.
- Glenn, T. C. 2011. Field guide to next-generation DNA sequencers. *Molecular ecology resources* 759-769.
- Gogoleva, S. S., I. A. Volodin, E. V. Volodina, A. V. Kharlamova, and L. N. Trut. 2011. Explosive vocal activity for attracting human attention is related to domestication in silver fox. *Behavioural processes* 86:216-21.
- Gray, M. M., N. B. Sutter, E. A. Ostrander, and R. K. Wayne. 2010. The IGF1 small dog haplotype is derived from Middle Eastern grey wolves. *BMC biology* 8:16.
- Gray, M. M., J. M. Granka, C. D. Bustamante, N. B. Sutter, A. R. Boyko, L. Zhu, E. A. Ostrander, and R. K. Wayne. 2009. Linkage disequilibrium and demographic history of wild and domestic canids. *Genetics* 181:1493.
- Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, et al. 2010. A draft sequence of the Neandertal genome. *Science (New York, N.Y.)* 328:710-22.
- Griggio, M., C. Biard, D. J. Penn, and H. Hoi. 2011. Female house sparrows “count on” male genes: experimental evidence for MHC-dependent mate preference in birds. *BMC evolutionary biology* 11:44.
- Gudmundsson, K. O., L. Thorsteinsson, O. E. Sigurjonsson, J. R. Keller, K. Olafsson, T. Egeland, S. Gudmundsson, and T. Rafnar. 2007. Gene expression analysis of hematopoietic progenitor cells identifies *Dlg7* as a potential stem cell gene. *Stem Cells* 25:1498-1506.

- Guevel, L., J. R. Lavoie, C. Perez-Iratxeta, K. Rouger, L. Dubreil, M. Feron, S. Talon, M. Brand, and L. A. Megeney. 2011. Quantitative proteomic analysis of dystrophic dog muscle. *Journal of Proteome Research* 10:2465-2478.
- Haines, B. P., and P. W. J. Rigby. 2008. Expression of the Lingo/LERN gene family during mouse embryogenesis. *Gene Expression Patterns* 8:79-86, doi:10.1016/j.modgep.2007.10.003.
- Handelsman, J. 2004. Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews* 68:669-685.
- Hare, B., A. Rosati, J. Kaminski, J. Bräuer, J. Call, and M. Tomasello. 2010. The domestication hypothesis for dogs' skills with human communication: a response to Udell et al. (2008) and Wynne et al. (2008). *Animal Behaviour* 79:e1-e6.
- Harries, L. W., D. Hernandez, W. Henley, A. R. Wood, A. C. Holly, R. M. Bradley-Smith, H. Yaghootkar, A. Dutta, A. Murray, T. M. Frayling, J. M. Guralnik, S. Bandinelli, A. Singleton, L. Ferrucci, and D. Melzer. 2011. Human aging is characterized by focused changes in gene expression and deregulation of alternative splicing. *Aging cell* 10:868-78.
- Harrington, F. H., and C. S. Asa. 2003. Wolf communication. In: Mech L. D. and L. Boitani (Eds.), *Wolves: behavior, ecology and conservation*. The University of Chicago Press, pp. 66-103.
- Harrington, F. H., and H. Fred. 2000. What is a howl?, NOVA Online, PBS.
- Havlicek, J., and S. C. Roberts. 2009. MHC-correlated mate choice in humans: a review. *Psychoneuroendocrinology* 34:497-512.
- Hedrick, P. W., and F. L. Black. 1997. HLA and mate selection: no evidence in South Amerindians. *American Journal of Human Genetics* 61: 505-511.
- Hedrick, P. W., R. N. Lee, and D. Garrigan. 2002. Major histocompatibility complex variation in red wolves: evidence for common ancestry with coyotes and balancing selection. *Molecular ecology* 11:1905-13.
- Hedrick, P. W., R. N. Lee, and K. M. Parker. 2000. Major histocompatibility complex (MHC) variation in the endangered Mexican wolf and related canids. *Heredity* 85:617-24.
- Hedrick, P. W., R. N. Lee, and C. Buchanan. 2003. Canine parvovirus enteritis, canine distemper, and major histocompatibility complex genetic variation in Mexican wolves. *Journal of wildlife diseases* 39:909-13.
- Hill, A. V. S. 1991. HLA associations with malaria in Africa: some implications for MHC evolution. In: Klein J. and D. Klein (Eds.), *Molecular evolution of the major histocompatibility complex*. Berlin, Germany: Springer, pp. 403-419.
- Hodges, E., M. Rooks, Z. Xuan, A. Bhattacharjee, D. B. Gordon, L. Brizuela, W. Richard McCombie, and G. J. Hannon. 2009. Hybrid selection of discrete genomic intervals on

- custom-designed microarrays for massively parallel sequencing. *Nature protocols* 4:960-74.
- Horton, R., L. Wilming, V. Rand, R. C. Lovering, E. a Bruford, V. K. Khodiyar, M. J. Lush, S. Povey, C. C. Talbot, M. W. Wright, H. M. Wain, J. Trowsdale, A. Ziegler, and S. Beck. 2004. Gene map of the extended human MHC. *Nature reviews. Genetics* 5:889-99.
- Huber, Đ., J. Kusak, A. Frković, and G. Gužvica. 2002. Causes of wolf mortality in Croatia in the period 1986-2001. *Veterinarski Arhiv* 72:131-139.
- Iacolina, L., M. Scandura, A. Gazzola, N. Cappai, C. Capitani, L. Mattioli, F. Vercillo, and M. Apollonio. 2010. Y-chromosome microsatellite variation in Italian wolves: A contribution to the study of wolf-dog hybridization patterns. *Mammalian Biology - Zeitschrift fur Säugetierkunde* 75:341-347.
- Imaizumi, K., T. Morihara, Y. Mori, T. Katayama, M. Tsuda, T. Furuyama, A. Wanaka, M. Takeda, and M. Tohyama. 1999. The cell death-promoting gene DP5, which interacts with the BCL2 family, is induced during neuronal apoptosis following exposure to amyloid β protein. *Journal of Biological Chemistry* 274:7975-7981. doi:10.1074/jbc.274.12.7975.
- It, V., L. Barrientos, J. López Gappa, D. Posik, S. Díaz, C. Golijow, and G. Giovambattista. 2010. Association of canine juvenile generalized demodicosis with the dog leukocyte antigen system. *Tissue antigens* 76:67-70.
- Jamieson, I. G., S. S. Taylor, L. N. Tracy, H. Kokko, and D. P. Armstrong. 2009. Why some species of birds do not avoid inbreeding: insights from New Zealand robins and saddlebacks. *Behavioral Ecology* 20:575–584.
- Janeš, D., I. Klun, B. Vidan-Jeras, M. Jeras, and S. Kreft. 2010. Influence of MHC on odour perception of 43 chemicals and body odour. *Central European Journal of Biology* 5:324-330.
- Johnson, M., I. Zaretskaya, Y. Raytselis, Y. Merezuk, S. McGinnis, and T. L. Madden. 2008. NCBI BLAST: a better web interface. *Nucleic Acids Research*. 36:W5-W9.
- Jokinen, P., E. M. Rusanen, L. J. Kennedy, and H. Lohi. 2011. MHC class II risk haplotype associated with canine chronic superficial keratitis in German Shepherd dogs. *Veterinary immunology and immunopathology* 140:37-41.
- Jones, E., and P. L. Stevens. 1988. Reproduction in wild canids, *Canis familiaris*, from the Eastern highlands of Victoria. *Australian Wildlife Research* 15:385–397.
- Jonkers, I., T. S. Barakat, E. M. Achame, K. Monkhorst, A. Kenter, E. Rentmeester, F. Grosveld, J. A. Grootegoed, and J. Gribnau. 2009. RNF12 is an X-Encoded dose-dependent activator of X chromosome inactivation. *Cell* 139:999-1011.
- Joslin, P.W.B. 1967. Movements and home site of timber wolves in Algonquin Park. Department of Zoology, University of Toronto, Canada.

- Juriscicova, A., M. Antenos, S. Varmuza, J. L. Tilly, and R. F. Casper. 2003. Expression of apoptosis-related genes during human preimplantation embryo development: potential roles for the Harakiri gene product and Caspase-3 in blastomere fragmentation *Molecular Human Reproduction* 9:133-141, doi:10.1093/molehr/gag016
- Karlsson, E. K., I. Baranowska, C. M. Wade, N. H. C. Salmon Hillbertz, M. C. Zody, N. Anderson, T. M. Biagi, N. Patterson, G. R. Pielberg, E. J. Kulbokas, K. E. Comstock, E. T. Keller, J. P. Mesirov, H. von Euler, O. Kämpe, A. Hedhammar, E. S. Lander, G. Andersson, L. Andersson, and K. Lindblad-Toh. 2007. Efficient mapping of mendelian traits in dogs through genome-wide association. *Nature genetics* 39:1321-8.
- Karlsson, E. K., and K. Lindblad-Toh. 2008. Leader of the pack: gene mapping in dogs and other model organisms. *Nature reviews. Genetics* 9:713-25.
- Kaufman, J., S. Milne, T. W. Gobel, B. A. Walker, J. P. Jacob, C. Auffray, R. Zoorob, and S. Beck. 1999. The chicken B locus is a minimal essential major histocompatibility complex. *Nature* 401:923-925.
- Kennedy, L. J., J. M. Angles, A. Barnes, S. D. Carter, O. Francino, J. A. Gerlach, G. M. Happ, W. E. Ollier, W. Thomson, and J. L. Wagner. 2001. Nomenclature for factors of the dog major histocompatibility system (DLA), 2000: second report of the ISAG DLA Nomenclature Committee. *Animal genetics* 32:193-9.
- Kennedy, L. J., S. D. Carter, A. Barnes, S. Bell, D. Bennett, B. Ollier, and W. Thomson. 1999a. DLA-DRB1 polymorphisms in dogs defined by sequence-specific oligonucleotide probes (SSOP). *Tissue antigens* 53:184-9.
- Kennedy, L. J., S. D. Carter, A. Barnes, S. Bell, D. Bennett, W. E. Ollier, and W. Thomson. 1998. Nine new dog DLA-DRB1 alleles identified by sequence-based typing. *Immunogenetics* 48:296-301.
- Kennedy, L. J., H. J. Huson, J. Leonard, J. M. Angles, L. E. Fox, J. W. Wojciechowski, C. Yuncker, and G. M. Happ. 2006. Association of hypothyroid disease in Doberman Pinscher dogs with a rare major histocompatibility complex DLA class II haplotype. *Tissue antigens* 67:53-6.
- Kennedy, L. J., D. A. Randall, D. Knobel, J. J. Brown, A. R. Fooks, K. Argaw, F. Shiferaw, W. E. R. Ollier, C. Sillero-Zubiri, D. W. Macdonald, and M. K. Laurensen. 2011. Major histocompatibility complex diversity in the endangered Ethiopian wolf (*Canis simensis*). *Tissue antigens* 77:118-25.
- Kennedy, L. J., A. Barnes, A. Short, J. J. Brown, S. Lester, J. Seddon, L. Fleeman, O. Francino, et al. 2007. Canine DLA diversity: 1. New alleles and haplotypes. *Tissue Antigens* 42:272-288.
- Kennedy, L. J., S. D. Carter, A. Barnes, S. Bell, D. Bennett, B. Ollier, and W. Thomson. 1999b. Interbreed variation of DLA-DRB1, DQA1 alleles and haplotypes in the dog. *Veterinary immunology and immunopathology* 69:101-11.

- Kennedy, L. J., S. Quarmby, N Fretwell, A. J. Martin, P. G. Jones, C. A. Jones, and W E R Ollier. 2005. High-resolution characterization of the canine DLA-DRB1 locus using reference strand-mediated conformational analysis. *The Journal of heredity* 96:836-42.
- Kennedy, L. J., J. M. Angles, A. Barnes, L. E. Carmichael, A. D. Radford, W. E. R. Ollier, and G. M Happ. 2007. DLA-DRB1, DQA1, and DQB1 alleles and haplotypes in North American Gray Wolves. *The Journal of heredity* 98:491-9.
- King, D. P., Y. Zhao, A. M. Sangoram, L. D. Wilsbacher, M. Tanaka, M. P. Antoch, T. D. Steeves, M. H. Vitaterna, J. M. Kornhauser, P. L. Lowrey, F. W. Turek, and J. S. Takahashi. 1997. Positional cloning of the mouse circadian clock gene. *Cell* 89:641-53.
- Kirkman, T. W. 1996. Statistics to use. <http://www.physics.csbsju.edu/stats/>, accessed on December 21st, 2011.
- Kirkness, E. F., V. Bafna, A. L. Halpern, S. Levy, K. Remington, D. B. Rusch, A. L. Delcher, M. Pop, W. Wang, C. M. Fraser, and J. C. Venter. 2003. The dog genome: survey sequencing and comparative analysis. *Science* 301:1898-903.
- Kohli, M. A., S. Lucae, P. G. Saemann, M. V. Schmidt, et al. 2011. The neuronal transporter gene SLC6A15 confers risk to major depression. *Neuron* 70:252-265, doi:10.1016/j.neuron.2011.04.005.
- Koler-Matznick, J. 2002. The Origin of the Dog Revisited. *Anthrozoos* 15:98-118.
- Kreeger, T.J. 2003. The internal wolf: Physiology, pathology, and pharmacology. In: L. D. Mech and L. Boitani (Eds.), *Wolves: Behavior, Ecology, and Conservation*. The University of Chicago press, pp. 192-217.
- Kukekova, A. V., J. L. Johnson, C. Teiling, Lewyn Li, I. N. Oskina, A. V. Kharlamova, R. G. Gulevich, R. Padte, M. M. Dubreuil, A. V. Vladimirova, D. V. Shepeleva, S. G. Shikhevich, Q. Sun, L. Ponnala, S. V. Temnykh, L. N. Trut, and G. M. Acland. 2011. Sequence comparison of prefrontal cortical brain transcriptome from a tame and an aggressive silver fox (*Vulpes vulpes*). *BMC genomics* 12:482.
- Kukekova, A. V., L. N. Trut, K. Chase, A. V. Kharlamova, J. L. Johnson, S. V. Temnykh, I. N. Oskina, R. G. Gulevich, A. V. Vladimirova, S. Klebanov, D. V. Shepeleva, S. G. Shikhevich, G. M. Acland, and Karl G Lark. 2011. Mapping Loci for fox domestication: deconstruction/reconstruction of a behavioral phenotype. *Behavior genetics* 41:593-606.
- Kunath, M., H. Ludecke, and A. Vortkamp. 2002. Expression of *Trps1* during mouse embryonic development. *Mechanisms of development* 5:117-120.
- Landry, C., D. Garant, P. Duchesne, and L. Bernatchez. 2001. 'Good genes as heterozygosity': the major histocompatibility complex and mate choice in Atlantic salmon (*Salmo salar*). *Proceedings of the Royal Society of London B: Biological Sciences* 268:1279-1285.
- Leonard, J. A., C. Vilà, and R. K. Wayne. 2005. From wild wolf to domestic dog. In: Ostrander E. A., U. Giger, K. Lindblad-Toh (Eds.), *The dog and its genome*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

- Leonard, J. A., R. K. Wayne, J. Wheeler, R. Valadez, S. Guillén, and C. Vilà. 2002. Ancient DNA evidence for Old World origin of New World dogs. *Science (New York, N.Y.)* 298:1613-6.
- Lewis, K. 1998. Pathogen resistance as the origin of kin altruism. *Journal of Theoretical Biology*. 193:359–363.
- Li, H., and R. Durbin. 2011. Inference of human population history from individual whole-genome sequences. *Nature* 1-5.
- Li, R., W. Fan, G. Tian, H. Zhu, L. He, et al. 2010. The sequence and de novo assembly of the giant panda genome. *Nature* 463:311-7.
- Li et al. submitted. Positive natural selection and artificial selection on brain expressed genes during the evolution of dog.
- Librado, P., and J. Rozas. 2009. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451-1452, doi:10.1093/bioinformatics/btp187.
- Lindblad-Toh, K., C. M. Wade, T. S. Mikkelsen, E. K. Karlsson, D. B. Jaffe, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438:803-19.
- Locke, D. P., L. W. Hillier, W. C. Warren, K. C. Worley, L. V. Nazareth, et al. 2011. Comparative and demographic analysis of orang-utan genomes. *Nature* 469:529-33.
- Lovari, S., A. Sforzi, C. Scala, and R. Fico. 2007. Mortality parameters of the wolf in Italy: does the wolf keep himself from the door? *Journal of Zoology* 272:117-124.
- Lucchini, V., E. Fabbri, F. Marucco, S. Ricci, L. Boitani, and E. Randi. 2002. Noninvasive molecular tracking of colonizing wolf (*Canis lupus*) packs in the western Italian Alps. *Molecular ecology* 11:857-68.
- Lucchini, V., A. Galov, and E. Randi. 2004. Evidence of genetic distinction and long-term population decline in wolves (*Canis lupus*) in the Italian Apennines. *Molecular Ecology* 13:523-536.
- Maccatrozzo, L., F. Caliaro, L. Toniolo, M. Patruno, C. Reggiani, and F. Mascarello. 2007. The sarcomeric myosin heavy chain gene family in the dog: Analysis of isoform diversity and comparison with other mammalian species. *Genomics* 89:224-236, doi:10.1016/j.ygeno.2006.08.004.
- Maglia, G., M. R. Restrepo, E. Mikhailova, and H. Bayley. 2008. Enhanced translocation of single DNA molecules through alpha-hemolysin nanopores by manipulation of internal charge. *PNAS USA* 105:19720–19725.
- Malik, T. H., S. A. Shoichet, P. Latham, T. G. Kroll, L. L. Peters, and R. A. Shivdasani. 2001. Transcriptional repression and developmental functions of the atypical vertebrate GATA protein TRPS1. *The EMBO journal* 20:1715-25.

- Malik, T.H., D. Von Stechow, R. T. Bronson, and R. A. Shivdasani. 2002. Deletion of the GATA domain of TRPS1 causes an absence of facial hair and provides new insights into the bone disorder in inherited tricho-rhino-phalangeal syndromes. *Molecular and Cellular Biology* 22:8592.
- Mardis, E. R. 2008. The impact of next-generation sequencing technology on genetics. *Trends in genetics: TIG* 24:133-41.
- Marucco, F., D. H. Pletscher, L. Boitani, M. K. Schwartz, K. L. Pilgrim, and J.D. Lebreton. 2009. Wolf survival and population trend using non-invasive capture-recapture techniques in the Western Alps. *Journal of Applied Ecology* 46:1003-1010.
- Mech, L. D. 1970. *The wolf: the ecology and behavior of an endangered species*. University of Minnesota Press, Minneapolis, MN.
- Mech, L.D. 1974. Current techniques in the study of elusive wilderness carnivores. In: I. Kjerner and P. Bjurholm (Eds.), *Proceedings of 11th Congress of the International Union of Game Biologists*, Stockholm, Sweden, 3–7 September 1973. Swedish National Environment Protection Board, Stockholm, pp. 315–322.
- Mech, L. D. 1999. Alpha status, dominance, and division of labour in wolf packs, *Canadian Journal of Zoology* 77:1196-1203.
- Mech, L. D., and L. Boitani (Eds.). 2003. *Wolves: Behaviour, Ecology and Conservation*. The University of Chicago Press.
- MHC Sequencing Consortium. 1999. Complete sequence and gene map of a human major histocompatibility complex. *Nature* 401:921–923.
- Mi, S., R. H. Miller, X. Lee, M. L. Scott, S. Shulag-Morskaya, Z. Shao, J. Chang, G. Thill, M. Levesque, M. Zhang, C. Hession, D. Sah, B. Trapp, Z. He, V. Jung, J. M. McCoy, and R. B. Pepinsky. 2005. LINGO-1 negatively regulates myelination by oligodendrocytes. *Nature neuroscience* 8:745-51.
- Millson, A., D. Lagrave, M. J. H. Willis, L. R. Rowe, E. Lyon, and S. T. South. 2011. Chromosomal loss of 3q26.3-3q26.32, involving a partial neuroligin 1 deletion, identified by genomic microarray in a child with microcephaly, seizure disorder, and severe intellectual disability. *American journal of medical genetics. Part A* .
- Murgia, C., J. K. Pritchard, S. Y. Kim, A. Fassati, and R. A. Weiss. 2006. Clonal origin and evolution of a transmissible cancer. *Cell* 126:477-87.
- Musiani, M., J. A. Leonard, H. D. Cluff, C. C. Gates, S. Mariani, P. C. Paquet, C. Vilà, and R. K. Wayne. 2007. Differentiation of tundra/taiga and boreal coniferous forest wolves: genetics, coat colour and association with migratory caribou. *Molecular ecology* 16:4149-70.
- Navarro, P., M. Moffat, N. P. Mullin, and I. Chambers. 2011. The X-inactivation trans-activator Rnf12 is negatively regulated by pluripotency factors in embryonic stem cells. *Human genetics* 130:255-64.

- Neff, B. D., S. R. Garner, J. W. Heath, and D. D. Heath. 2008. The MHC and non-random mating in a captive population of Chinook salmon. *Heredity* 101:175-85.
- Ng, S. B., E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. E. Eichler, M. Bamshad, D. A. Nickerson, and J. Shendure. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461:272-6.
- Nicholas, T. J., C. Baker, E. E. Eichler, and J. M. Akey. 2011. A high-resolution integrated map of copy number polymorphisms within and between breeds of the modern domesticated dog. *BMC genomics* 12:414.
- Nicholas, T. J., Z. Cheng, M. Ventura, K. Mealey, E. E. Eichler, and J. M. Akey. 2009. The genomic architecture of segmental duplications and associated copy number variants in dogs. *Genome research* 19:491-9.
- Nowak, R. M. 2003. Wolf evolution and taxonomy. In: Mech, L. D. and L. Boitani (Eds.), *Wolves: Behaviour, Ecology and Conservation*, The University of Chicago Press, pp. 239-258.
- Nowak, M. A., K. Tarczyhornocho, and J. M. Austyn. 1992. The optimal number of major histocompatibility complex molecules in an individual. *PNAS USA* 89:10896–10899.
- van Oosterhout, C. 2009. A new theory of MHC evolution: beyond selection on the immune genes. *Proceedings. Biological sciences / The Royal Society* 276:657-65.
- van Oosterhout, C., D. A. Joyce, and S. M. Cummings. 2006. Evolution of MHC class IIB in the genome of wild and ornamental guppies, *Poecilia reticulata*. *Heredity* 97:111–118.
- van Orsouw, N. J., R. C. J. Hogers, A. Janssen, F. Yalcin, S. Snoeijers, E. Verstege, H. Schneiders, H. van der Poel, J. van Oeveren, H. Versteegen, and M. J. T. van Eijk. 2007. Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. *PloS one* 2:e1172.
- Ortalli, G. 1988. The invention of “bad wolf”. A report between history and ecosociology. *Convegno G.L.I., Parco Nazionale Foreste Casentinesi*.
- Ostrander, E. A., G. F. Sprague, and J. Rine. 1993. Identification and characterization of dinucleotide repeat (CA)_n markers for genetic mapping in dog. *Genomics* 16:207–213.
- Ouborg, N. J., C. Pertoldi, V. Loeschcke, R. K. Bijlsma, and P. W. Hedrick. 2010. Conservation genetics in transition to conservation genomics. *Trends in genetics: TIG* 26:177-87.
- Ovodov, N. D., S. J. Crockford, Y. V. Kuzmin, T. F. G. Higham, G. W. L. Hodgins, and J. van der Plicht. 2011. A 33,000-Year-Old Incipient Dog from the Altai Mountains of Siberia: Evidence of the Earliest Domestication Disrupted by the Last Glacial Maximum. *PLoS ONE* 6:e22821.

- O'Brien, S.J., M.E. Roelke, L. Marker, A. Newman, C. A. Winkler, D. Meltzer, L. Colly, J. F. Evermann, M. Bush, and D. E. Wildt. 1985. Genetic basis for species vulnerability in the cheetah. *Science* 227:1428–1434
- O'Connor, S. L., J. J. Lhost, E. A. Becker, A. M. Detmer, R. C. Johnson, C. E. Macnair, R. W. Wiseman, J. A. Karl, J. M. Greene, B. J. Burwitz, B. N. Bimber, S. M. Lank, J. J. Tuscher, E. T. Mee, N. J. Rose, R. C. Desrosiers, A. L. Hughes, T. C. Friedrich, M. Carrington, and D. H. O'Connor. 2010. MHC heterozygote advantage in simian immunodeficiency virus-infected Mauritian cynomolgus macaques. *Science translational medicine* 2:22ra18.
- Packard, J. M. 2003. Wolf behavior: reproductive, social and intelligent. In: L.D. Mech and L. Boitani (Eds.), *Wolves: Behavior, Ecology, and Conservation*. The University of Chicago press, pp. 35-55.
- Palombo, M., R. Sardella, and M. Novelli. 2008. Carnivora dispersal in Western Mediterranean during the last 2.6Ma. *Quaternary International* 179:176-189.
- Park, K., J. Kang, K. P. Subedi, J.-H. Ha, and C. Park. 2008. Canine polydactyl mutations with heterogeneous origin in the conserved intronic sequence of LMBR1. *Genetics* 179:2163-72.
- Parker, H. G., B. M. vonHoldt, P. Quignon, E. H. Margulies, S. Shao, et al. 2009. An expressed *fgf4* retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science* 325: 995–998.
- Peakall, R., and P. E. Smouse. 2006. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 6:288–295.
- Penn, D. J., and W. K. Potts. 1999. The Evolution of Mating Preferences and Major Histocompatibility Complex Genes. *Evolution* 153:145-164.
- Piertney, S. B., and M. K. Oliver. 2006. The evolutionary ecology of the major histocompatibility complex. *Heredity* 96:7-21.
- Potts, W. K., and E. K. Wakeland. 1990. Evolution of diversity at the major histocompatibility complex. *Trends in Ecology and Evolution* 5:181–186.
- Quignon, P., M. Rimbault, S. Robin, and F. Galibert. 2011. Genetics of canine olfaction and receptor diversity. *Mammalian genome* doi:10.1007/s00335-011-9371-1.
- Quilez, J., A. D. Short, V. Martinez, L. J. Kennedy, W. Ollier, A. Sanchez, L. Altet, and O. Francino. 2011. A selective sweep of >8 Mb on chromosome 26 in the Boxer genome. *BMC genomics* 12:339.
- Quinnell, R. J., L. J. Kennedy, A. Barnes, O. Courtenay, C. Dye, L. M. Garcez, M.-A. Shaw, S. D. Carter, W. Thomson, and W. E. R. Ollier. 2003. Susceptibility to visceral leishmaniasis in the domestic dog is associated with MHC class II polymorphism. *Immunogenetics* 55:23-8.

- R Development Core Team. 2009. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>.
- Radwan, J., A. Biedrzycka, and W. Babik. 2010. Does reduced MHC diversity decrease viability of vertebrate populations? *Biological Conservation* 143:537-544.
- Randi, E., V. Lucchini, M. F. Christensen, N. Mucci, S. M. Funk, G. Dolf, and V. Loeschcke. 2000. Mitochondrial DNA Variability in Italian and East European Wolves: Detecting the Consequences of Small Population Size and Hybridization. *Conservation Biology* 14:464-473.
- Randi, E., and V. Lucchini. 2002. Detecting rare introgression of domestic dog genes into wild wolf (*Canis lupus*) populations by Bayesian admixture analyses of microsatellite variation. *Conservation Genetics* 3:31-45.
- Randi, E.. 2008. Detecting hybridization between wild species and their domesticated relatives. *Molecular ecology* 17:285-93.
- Randi, E. 2010. Wolves in the Great Lakes region: a phylogeographic puzzle. *Molecular ecology* 19:4386-8.
- Randi, E. 2011. Genetics and conservation of wolves *Canis lupus* in Europe. *Mammal Review* 41:99-111.
- Rhotman, R. J., and L. D. Mech. 1979. Scent-marking in lone wolves and newly formed pairs, *Animal Behaviour* 27:750-760.
- Richardson, D. S., J. Komdeur, T. Burke, and T. von Schantz. 2005. MHC-based patterns of social and extra-pair mate choice in the Seychelles warbler. *Proceedings. Biological sciences / The Royal Society* 272:759-67.
- Robin, S., S. Tacher, M. Rimbault, A. Vaysse, S. Dréano, C. André, C. Hitte, and F. Galibert. 2009. Genetic diversity of canine olfactory receptors. *BMC genomics* 10:21.
- Robinson, J., K. Mistry, H. McWilliam, R. Lopez, and S. G. E. Marsh. 2010. IPD - the Immuno Polymorphism Database. *Nucleic Acids Research* 38:D863-839.
- Robinson, J., M. J. Waller, P. Parham, N. Groot, R. Bontrop, L. J. Kennedy, P. Stoehr, and S. G. E. Marsh. 2003. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic acids research* 31:311.
- Roelke, M. E., J. S. Martenson, and S. J. O'Brien. 1993. The consequences of demographic reduction and genetic depletion in the endangered Florida panther. *Current Biology* 3:340-350.
- Rouquier, S., and D. Giorgi. 2007. Olfactory receptor gene repertoires in mammals. *Mutation research* 616:95-102.
- Runstadler, J. A., J. M. Angles, and N. C. Pedersen. 2006. Dog leucocyte antigen class II diversity and relationships among indigenous dogs of the island nations of Indonesia (Bali), Australia and New Guinea. *Tissue antigens* 68:418-26.

- Ruth, T. K., D. W. Smith, M. A. Haroldson, P. C. Buotte, C. C. Schwartz, H. B. Quigley, S. Cherry, D. Tyers, K. Frey, and K. M. Murphy. 2003. Large-carnivore response to recreational big-game hunting along the Yellowstone National Park and Absaroka-Beartooth Wilderness boundary. *Wildlife Society Bulletin* 31:1150-1161.
- Ruvinsky, A., and J. Sampson. 2001. *The genetics of the dog* (A. Ruvinsky and J. Sampson, editors). CABI Publishing, New York.
- Saetre, P., J. Lindberg, J. A. Leonard, K. Olsson, U. Pettersson, H. Ellegren, T. F. Bergström, C. Vilà, and E. Jazin. 2004. From wild wolf to domestic dog: gene expression changes in the brain. *Brain research. Molecular brain research* 126:198-206.
- Salmon Hillbertz, N. H. C, M. Isaksson, E. K. Karlsson, E. Hellmen, G. R. Pielberg, et al. 2007. Duplication of FGF3, FGF4, FGF19 and ORAOV1 causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. *Nature Genetics* 39:1318–1320.
- Sand, H., C. Wikenros, P. Wabakken, and O. Liberg. 2006. Effects of hunting group size, snow depth and age on the success of wolves hunting moose. *Animal Behaviour* 72:781-789.
- Sanderson, H. S., and P. R. Clarke. 2006. Cell biology: Ran, mitosis and the cancer connection. *Current Biology* 16:R466-R468, doi:10.1016/j.cub.2006.05.032.
- Santos, P. S. C., T. Kellermann, B. Uchanska-Ziegler, and A. Ziegler. 2010. Genomic architecture of MHC-linked odorant receptor gene repertoires among 16 vertebrate species. *Immunogenetics* 62:569-84.
- Santos, P. S. C., J. A. Schinemann, J. Gabardo, and M. D. G. Bicalho. 2005. New evidence that the MHC influences odor perception in humans: a study with 58 Southern Brazilian students. *Hormones and behavior* 47:384-8.
- Savolainen, P., T. Leitner, A. N. Wilton, E. Matisoo-Smith, and J. Lundeberg. 2004. A detailed picture of the origin of the Australian dingo, obtained from the study of mitochondrial DNA. *Proceedings of the National Academy of Sciences of the United States of America* 101:12387-90.
- Savolainen, P., Y. Zhang, J. Luo, J. Lundeberg, and T. Leitner. 2002. Genetic evidence for an East Asian origin of domestic dogs. *Science (New York, N.Y.)* 298:1610-1613.
- Schleidt, W. M., and M. D. Shalter. 2003. Co-evolution of humans and canids – An alternative view of dog domestication: *Homo homini lupus?* *Evolution and Cognition* 9:57–72.
- Seddon, J. M. 2005. Canid-specific primers for molecular sexing using tissue or non-invasive samples. *Conservation Genetics* 6:147–149.
- Seddon, J. M., and H. Ellegren. 2004. A temporal analysis shows major histocompatibility complex loci in the Scandinavian wolf population are consistent with neutral evolution. *Proceedings. Biological sciences / The Royal Society* 271:2283-91.

- Seddon, J. M., and H. Ellegren. 2002. MHC class II genes in European wolves: a comparison with dogs. *Immunogenetics* 54:490-500.
- Sereni, E. 1961. *Storia Del Paesaggio Agrario Italiano*. Laterza, Bari.
- Setchell, J. M., M. J. E. Charpentier, K. M. Abbott, E. J. Wickings, and L. A. Knapp. 2010. Opposites attract: MHC-associated mate choice in a polygynous primate. *Journal of evolutionary biology* 23:136-48.
- Setchell, J. M., and E. Huchard. 2010. The hidden benefits of sex: evidence for MHC-associated mate choice in primate societies. *BioEssays*: news and reviews in molecular, cellular and developmental biology 32:940-8.
- Shimizu, S., N. Seki, T. Sugimoto, S. Horiguchi, H. Tanzawa, T. Hanazawa, and Y. Okamoto. 2007. Identification of molecular targets in head and neck squamous cell carcinomas based on genome-wide gene expression profiling. *Oncology Reports*, 18:1489-1497.
- Shin, J., M. Bossenz, Y. Chung, H. Ma, M. Byron, N. Taniguchi-Ishigaki, X. Zhu, B. Jiao, L. L. Hall, M. R. Green, S. N. Jones, I. Hermans-Borgmeyer, J. B. Lawrence, and I. Bach. 2010. Maternal Rnf12/RLIM is required for imprinted X-chromosome inactivation in mice. *Nature* 467:977-81.
- Shin, H. T., and M. W. Chang. 2001. Trichorhinophalangeal Syndrome, Type II (Langer-Giedion Syndrome). *Dermatology Online Journal* 7:8.
- Siddle, H.V., A. Kreiss, M. D. B. Eldridge, E. Noonan, C. J. Clarke, S. Pyecroft, G. M. Woods, and K. Belov. 2007. Transmission of a fatal clonal tumor by biting occurs due to depleted MHC diversity in a threatened carnivorous marsupial. *PNAS USA* 104:16221-16226.
- Siepel, A. G., Bejerano, J. S. Pedersen, A. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* 15:1034-1050.
- Sommer, R., and N. Benecke. 2005. Late-Pleistocene and early Holocene history of the canid fauna of Europe (Canidae). *Mammalian Biology - Zeitschrift fur Säugetierkunde* 70:227-241.
- Spehr, M., K. R. Kelliher, X.-H. Li, T. Boehm, T. Leinders-Zufall, and F. Zufall. 2006. Essential role of the main olfactory system in social recognition of major histocompatibility complex peptide ligands. *The Journal of neuroscience* 26:1961-70.
- Spurgin, L. G., and D. S. Richardson. 2010. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings. Biological sciences / The Royal Society* 277:979-88.
- Star, B., A. J. Nederbragt, S. Jentoft, U. Grimholt, M. Malmstrøm, T. F. Gregers, T. B. Rounge, J. Paulsen, M. H. Solbakken, A. Sharma, O. F. Wetten, A. Lanzén, R. Winer, J. Knight, J.-H. Vogel, B. Aken, Ø. Andersen, K. Lagesen, A. Tooming-Klunderud, R. B. Edvardsen, K. G. Tina, M. Espelund, C. Nepal, C. Previti, B. O. Karlsen, T. Moum, M. Skage, P. R. Berg, T. Gjøn, H. Kuhl, J. Thorsen, K. Malde, R. Reinhardt, L. Du, S. D.

- Johansen, S. Searle, S. Lien, F. Nilsen, I. Jonassen, S. W. Omholt, N. C. Stenseth, and K. S. Jakobsen. 2011. The genome sequence of Atlantic cod reveals a unique immune system. *Nature* 1-4.
- Stephens, M., and P. Donnelly. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics* 73:1162-1169.
- Stephens, M., N. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68:978-989.
- Stuglik, M. T., J. Radwan, and W. Babik. 2011. jMHC: software assistant for multilocus genotyping of gene families using next-generation amplicon sequencing. *Molecular ecology resources* 11:739-42.
- Sudhof, T. C. 2008. Neuroligins and neuexins link synaptic function to cognitive disease. *Nature* 455:903-911.
- Sundqvist, A.-K., S Björnerfeldt, J. A. Leonard, F. Hailer, A Hedhammar, H. Ellegren, and C. Vilà. 2006. Unequal contribution of sexes in the origin of dog breeds. *Genetics* 172:1121-8.
- Sundqvist, A. K., H. Ellegren, M. Olivier, and C. Vila. 2001. Y chromosome haplotyping in Scandinavian wolves (*Canis lupus*) based on microsatellite markers. *Molecular Ecology* 10:1959–1966.
- Sutter, N. B., C. D. Bustamante, K. Chase, M. M Gray, K. Zhao, L. Zhu, B. Padhukasahasram, E. Karlins, S. Davis, P. G Jones, P. Quignon, G. S. Johnson, H. G. Parker, N. Fretwell, D. S. Mosher, D. F. Lawler, E. Satyaraj, M. Nordborg, K G. Lark, R. K. Wayne, and E. A. Ostrander. 2007. A single IGF1 allele is a major determinant of small size in dogs. *Science* 316:112-5.
- Sutton, J. T., S. Nakagawa, B. C. Robertson, and I. G. Jamieson. 2011. Disentangling the roles of natural selection and genetic drift in shaping variation at MHC immunity genes. *Molecular ecology* 4408-4420.
- Suzuki, C., Y. Daigo, N. Ishikawa, T. Kato, S. Hayama, T. Ito, E. Tsuchiya, and Y. Nakamura. 2005. ANLN plays a critical role in human lung carcinogenesis through the activation of RHOA and by involvement in the phosphoinositide 3-Kinase/AKT pathway. *Cancer Research* 65:11314-11325.
- Tacher, S., P. Quignon, M. Rimbault, S. Dreano, C. Andre, and F. Galibert. 2005. Olfactory receptor sequence polymorphism within and between breeds of dogs. *The Journal of heredity* 96:812-6.
- Takanaga H., B. Mackenzie, J. B. Peng, and M. A. Hediger. 2005. Characterization of a branched-chain amino-acid transporter SBAT1 (SLC6A15) that is expressed in human brain. *Biochemical and Biophysical Research Communications* 337:892-900, doi:10.1016/j.bbrc.2005.09.128.

- Tamura, K., J. Dudley, M. Nei, and S. Kumar. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* 24:1596-1599.
- Tamura, K., M. Furihata, T. Tsunoda, et al. 2007. Molecular features of hormone-refractory prostate cancer cells by genome-wide gene expression profiles. *Cancer Research* 67:5117-5125.
- The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061-73.
- Theberge, J. B., and J. B. Falls. 1967. Howling as a means of communication in timber wolves, *American Zoologist*, 7:331-338.
- Thoss, M., P. Ilmonen, K. Musolf, and D J Penn. 2011. Major histocompatibility complex heterozygosity enhances reproductive success. *Molecular ecology* 20:1546-57.
- Topal, J., M. Gacsi, a Miklosi, Z. Viranyi, E. Kubinyi, and V. Csanyi. 2005. Attachment to humans: a comparative study on hand-reared wolves and differently socialized dog puppies. *Animal Behaviour* 70:1367-1375.
- Trut, L., I. Oskina, and A. Kharlamova. 2009. Animal evolution during domestication: the domesticated fox as a model. *BioEssays* 31:349-60.
- Turner, D. J., T. M. Keane, I. Sudbery, and D. J. Adams. 2009. Next-generation sequencing of vertebrate experimental organisms. *Mammalian genome* 20:327-38.
- Udell, M. A. R., N. R. Dorey, and C. D. L. Wynne. 2010. What did domestication do to dogs? A new account of dogs' sensitivity to human actions. *Biological reviews of the Cambridge Philosophical Society* 85:327-45.
- Udell, M. A. R., and C. D. L. Wynne. 2010. Ontogeny and phylogeny: both are essential to human-sensitive behaviour in the genus *Canis*. *Animal Behaviour* 79:e9-e14.
- Valière, N., L. Fumagalli, L. Gielly, C. Miquel, B. Lequette, M.-L. Poulle, J.-M. Weber, R. Arlettaz, and P. Taberlet. 2003. Long-distance wolf recolonization of France and Switzerland inferred from non-invasive genetic sampling over a period of 10 years. *Animal Conservation* 6:83-92.
- Vaysse, A., A. Ratnakumar, T. Derrien, E. Axelsson, G. Rosengren Pielberg, S. Sigurdsson, T. Fall, E. H. Seppälä, M. S. T. Hansen, C. T. Lawley, E. K. Karlsson, D. Bannasch, Carles Vilà, H. Lohi, F. Galibert, M. Fredholm, J. Häggström, Å. Hedhammar, C. André, K. Lindblad-Toh, C. Hitte, and M. T. Webster. 2011. Identification of Genomic Regions Associated with Phenotypic Variation between Dog Breeds using Selection Mapping. *PLoS Genetics* 7:e1002316.
- Verardi, A., V. Lucchini, and E. Randi. 2006. Detecting introgressive hybridization between free-ranging domestic dogs and wild wolves (*Canis lupus*) by admixture linkage disequilibrium analysis. *Molecular ecology* 15:2845-55.

- Vila, C., I. R. Amorim, J. A. Leonard, D. Posada, J. Castroviejo, F. Petrucci-Fonseca, K. A. Crandall, H. Ellegren, and R. K. Wayne. 1999. Mitochondrial DNA phylogeography and population history of the grey wolf *canis lupus*. *Molecular ecology* 8:2089-103.
- Vila, C., P. Savolainen, J. E. Maldonado, I. R. Amorim, J. E. Rice, R. L. Honeycutt, K. A. Crandall, J. Lundeberg, and R. K. Wayne. 1997. Multiple and Ancient Origins of the Domestic Dog. *Science* 276:1687-1689.
- Vilà C., and R. K. Wayne. 1999. Hybridization between wolves and dogs. *Conservation Biology* 13:195–198.
- Vilariño-Güell, C., C. Wider, O. Ross, B. Jasinska-Myga, J. Kachergus, S. Cobb, A. Soto-Ortolaza, B. Behrouz, M. Heckman, N. Diehl, C. Testa, Z. Wszolek, R. Uitti, J. Jankovic, E. Louis, L. Clark, A. Rajput, and M. Farrer. 2010. LINGO1 and LINGO2 variants are associated with essential tremor and Parkinson disease. *Neurogenetics* 11:408-411, doi: 10.1007/s10048-010-0241-x.
- Vonholdt, B. M., J. P. Pollinger, D. A. Earl, J. C. Knowles, A. R. Boyko, H. Parker, E. Geffen, M. Pilot, W. Jedrzejewski, B. Jedrzejewska, V. Sidorovich, C. Greco, E. Randi, M. Musiani, R. Kays, C. D. Bustamante, E. A. Ostrander, J. Novembre, and R. K. Wayne. 2011. A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. *Genome research* .
- Vonholdt, B. M., J. P. Pollinger, K. E. Lohmueller, E. Han, H. G. Parker, P. Quignon, J. D. Degenhardt, A. R. Boyko, D. A. Earl, A. Auton, A. Reynolds, K. Bryc, A. Brisbin, J. C. Knowles, D. S. Mosher, T. C. Spady, A. Elkahloun, E. Geffen, M. Pilot, W. Jedrzejewski, C. Greco, E. Randi, D. Bannasch, A. Wilton, J. Shearman, M. Musiani, M. Cargill, Paul G Jones, Z. Qian, W. Huang, Z.-L. Ding, Y.-P. Zhang, C. D Bustamante, E. A. Ostrander, J. Novembre, and R. K. Wayne. 2010. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* 464:898-902.
- Vonholdt, B. M., D. R. Stahler, D. W. Smith, D. A. Earl, J. P. Pollinger, and R. K. Wayne. 2008. The genealogy and genetic viability of reintroduced Yellowstone grey wolves. *Molecular ecology* 17:252-74.
- Wagner, J. L., R. C. Burnett, J. D. Works, and R. Storb. 1996. Molecular analysis of DLA-DRBB1 polymorphism. *Tissue antigens* 48:554-61.
- Wang, J., and A. W. Santure. 2009. Parentage and sibship inference from multi-locus genotype data under polygamy. *Genetics* 181:1579-1594.
- Wayne R. K., N. Lehman, M. W. Allard, and R. L. Honeycutt. 1992. Mitochondrial DNA variability of the gray wolf: genetic consequences of population decline and habitat fragmentation. *Conservation Biology* 6:559–569.
- Wayne R. K., N. Lehman, and T. K. Fuller. 1995. Conservation genetics of the gray wolf. In: Carbyn L. N., S. H. Fritts SH, D. R. Seip (Eds.), *Ecology and Conservation of Wolves in a Changing World*. Canadian Circumpolar Institute, Occasional Publication no.35. Edmonton, Alberta, pp. 399–407.

- Wayne, R. K., and E. A. Ostrander. 2007. Lessons learned from the dog genome. *Trends in genetics*: TIG 23:557-67.
- Wedekind, C., T. Seebeck, F. Bettens, and A. J. Paepke. 1995. MHC-dependent mate preferences in humans. *Proceedings. Biological sciences / The Royal Society* 260:245-9.
- Wegner, K. M. 2009. Massive parallel MHC genotyping: titanium that shines. *Molecular ecology* 18:1818-20.
- Wiedmann, R. T., T. P. L. Smith, and D. J. Nonneman. 2008. SNP discovery in swine by reduced representation and high throughput pyrosequencing. *BMC genetics* 9:81.
- Wilde, A. 2006. "HURP on" we're off to the kinetochore! *Journal of Cell Biology* 173:829-831, doi:10.1083/jcb.200605150.
- Wolfe, A. D., and J. J. Henry. 2006. Neuronal leucine-rich repeat 6 (XINLRR-6) is required for late lens and retina development in *Xenopus laevis*. *Developmental Dynamics* 235:1027-1041, doi:10.1002/dvdy.20691.
- Wong, A. K., A. L. Ruhe, B. L. Dumont, K. R. Robertson, G. Guerrero, S. M. Shull, J. S. Ziegler, L. V. Millon, K. W. Broman, B. A. Payseur, and M. W. Neff. 2010. A comprehensive linkage map of the dog genome. *Genetics* 184:595-605.
- Wu, C., C. Orozco, J. Boyer, M. Leglise, J. Goodale, S. Batalov, C. L. Hodge, J. Haase, J. Janes, J. W. Huss, and A. I. Su. 2009. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome biology* 10:R130.
- Xie, X., J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis. 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 434:338-45.
- Yoshikawa, Y., M. Morimatsu, and K. Ochiai. 2008. Novel variations and loss of heterozygosity of BRCA2 identified in a dog with mammary tumors. *American journal of veterinary research* 69:1323-1328.
- Yuhki, N., T. Beck, R. Stephens, B. Neelam, and S. J. O'Brien. 2007. Comparative genomic structure of human, dog, and cat MHC: HLA, DLA, and FLA. *The Journal of heredity* 98:390-9.
- Zeuner, F. E. 1963. *A History of Domesticated Animals*. Hutchinson & Co. Ltd., London.
- Zhang, Y., and A. Jeltsch. 2010. The application of next generation sequencing in DNA methylation analysis. *Genes* 1:85-101, doi:10.3390/genes1010085
- Zhang, H., Q. Wei, H. Zhang, and L. Chen. 2011. Comparison of the fraction of olfactory receptor pseudogenes in wolf (*Canis lupus*) with domestic dog (*Canis familiaris*). *Journal of Forestry Research* 22:275-280.
- Zhou, D., S. Li, J. Wen, X. Gong, L. Xu, and Y. Luo. 2008. Genome-wide computational analyses of microRNAs and their targets from *Canis familiaris*. *Computational biology and chemistry* 32:60-5.

- Ziegler, A., P. S. C. Santos, T. Kellermann, and B. Uchanska-Ziegler. 2010. Self/nonself perception, reproduction and the extended MHC. *Self/nonself* 1:176-191.
- Zimen, E., and L. Boitani. 1975. Number and distribution of wolves in Italy. *Zeitschrift für Säugetierkunde* 40:102–112.

6. Acknowledgements

Grazie a coloro i quali ritengono che una borsa di dottorato sia “un regalo al prof” per il quale il dottorando lavora, e che i progetti svolti presso enti esterni, come il mio, siano “soldi sprecati”.

Grazie a coloro i quali, nell’assegnazione delle borse Marco Polo, hanno ritenuto il mio curriculum e la qualità del mio progetto all’estero di secondaria importanza rispetto all’aver come tutor un professore ‘esterno’ al dipartimento, ovviamente senza avere il coraggio di scriverlo nel bando.

Grazie a tutti coloro i quali, bloccando l’accesso alle porzioni più rilevanti mercato del lavoro, hanno fatto sì che decine di migliaia di giovani eccellenti laureati si trovino ad elemosinare un lavoro interinale a 1000€ al mese.

Grazie a tutti loro, oggi l’Italia è quella che è.

Ma proprio per questo spetta a noi il compito di cambiarla.

Polemiche a parte, grazie di cuore a tutti quelli che mi hanno sostenuto in quest’avventura di dottorato.

In ordine sparso, grazie a tutta la mia famiglia per la fiducia e il sostegno costanti.

Grazie a Serena per l’appoggio instancabile ma soprattutto per i consigli sempre costruttivi, anche nei momenti più difficili – che purtroppo non sono mancati.

Grazie ad Ettore per la fiducia e il calore dal punto di vista umano, e per la supervisione eccellente dal punto di vista scientifico.

Grazie a Romolo ed Elena per l’esempio brillante di impegno quotidiano, e sul cui lavoro decennale si basano gran parte dei dati del presente studio; un grazie speciale a Romolo per l’appoggio, anche a distanza, nei momenti bui.

Grazie ad Andrea per aver condiviso con semplicità e simpatia la stanza dottorandi, oltre ad innumerevoli birre...

Grazie a Cocche perché ci sei sempre da vicino.

Grazie ad Andre per la compagnia nel bellissimo appartamento stile ‘Bologna anni ‘70’ in cui sono rimasto annidato nelle settimane frenetiche di stesura della tesi, e grazie a Stefano per l’amicizia e la caparbità.

Grazie a Marco e Maria-Grazia per l'esempio costante di gioia e dei valori veri della vita.

Grazie a tutti gli amici di Bologna per le serate svacco e quelli di S. Paolo per i festeggiamenti...alternativi.

Grazie ai membri del consiglio di dottorato, per i concorsi di selezione completamente trasparenti e per averci supportato nelle nostre esperienze all'estero.

Thanks to Bob for his enthusiasm, for the confidence he put on me and for the great opportunity of joining the wolf genome project.

Thanks to all the members of Los Lobos pack -Rena, Pedro and Xin- for the great time we spent together learning some genomics - and not dying trying!- but also during our deep lunchtime discussions about the world. Ahuu!

Thanks to John N for the great spirit and scientific excellence - it was a real pleasure working with your team! – and to the brother-in-town Adam.

Thanks to all the other Wayne and Novembre's Lab members: you are just too many to be mentioned, but I simply shared great moments with everyone of you. You know it!

Thanks to Fr. Peter for the great talks and advices about life and faith, and the other great people at UCC like Cecy and Fr. Paul.

And, as always, thanks to the wolves.

“Sbirciare l’infinito fa aumentare lo spazio, il respiro, la testa, di chi lo sta a osservare.

A forza di stupore la scienza progredi.

Provare meraviglia è un requisito scientifico, perché istiga a scoprire.

Se non c’è più la meraviglia nello scatto di chi si chiude in un laboratorio,

peggio per lui e peggio per la scienza.

Fu la sterminata immensità della notte a spalancare i pensieri dei nostri antenati.

Accorgersi che esiste l’infinito è già un inizio d’intesa tra la minima taglia della creatura

umana e l’universo”

Erri De Luca