Chapter

Canine Genetics and Genomics

Edo D'Agaro, Andrea Favaro and Davide Rosa

Abstract

In the past fifteen years, tremendous progress has been made in dog genomics. Several genetic aspects of cancer, heart disease, hip dysplasia, vision and hearing problems in dogs have been investigated and studied in detail. Genome-wide associative studies have made it possible to identify several genes associated with diseases, morphological and behavioral traits. The dog genome contains an extraordinary amount of genetic variability that distinguishes the different dog breeds. As a consequence of the selective programs, applied using stringent breed standards, each dog breed represents, today, a population isolated from the others. The availability of modern next generation sequencing (NGS) techniques and the identification of millions of single functional mutations (SNPs) has enabled us to obtain new and unknown detailed genomic data of the different breeds.

Keywords: canine genetics, canine genomics, bioinformatics, breeding, genetic diseases

1. Introduction

In recent years, genetic studies on dog genomics have multiplied worldwide. Currently, there are over 50 international laboratories which are involved in canine genome projects and several applications will be available in the near future from these studies. These new findings will improve our understanding of the selection process of the dogs and provide useful information for the study and control of genetic diseases.

2. Basic genetics

The single-control characters are influenced by genes located in a locus on one of the pairs of the chromosomes (78 in the dog) and have a binomial distribution. For example, the hair length in dogs is coded by two genes present at an autosomal locus. Short-haired animals have genotype LL (dominant homozygotes), while, long-haired animals have genotype ll (recessive homozygotes). From their mating originates short-haired animals with genotype Ll (heterozygotes), indistinguishable from short-haired parents. Even those characters that express different degrees of dominance, different from the Mendelian inheritance, are considered simple characters (e.g. incomplete or partial dominance). The simple characters are not influenced by the environment and, therefore, to each genotype corresponds a certain phenotype (P = G, where P = phenotype and G = genotype). The study of simple characters includes also multiple alleles (several alleles present in a population), pleitropy, association or linkage and incomplete penetrance. For characters with simple inheritance, it is easier to make selection than for multiple control characters. The multiple control characters are also called quantitative or polygenic characters. These characters are influenced by many genes distributed on several loci and they are influenced by environmental factors. The strong artificial selection exercised by man during the domestication process and during the creation of the different breeds has led to the setting of several characters. Color inheritance illustrates the case of separate loci that control the expression of the phenotype. The coat of dogs consists of two parts: top coat (protective function) and undercoat (heat-insulating function). Some breeds have no undercoat (e.g. Yorkshire). The color of the coat depends on the characteristics of the pigments contained in the medullary and cortical layers of the hair [1]. According to Willis [1], it is possible to explain all the colors by means of two chemical pigments: hemoglobin and melanin. More specifically, melanin is differentiated into eumelanin (black-brown) and pheomelanin (yellow-reddish). The synthesis of pigments in the hair of mammals depends on the interaction between the Agouti protein and the Melanocortin 1 receptor [2]. The coat colors in the dog are linked to the presence/absence of two types of melanin and their possible combinations. It is important to underline that melanin do not show a precise time of formation and they develop during the different phases of the fetal development and after birth [1]. The knowledge of the genetic inheritance of the morphological traits is very important in order to establish suitable selection objectives in the different breeds.

3. Relationship and inbreeding

Measurement of F coefficient (consanguinity) in a population can be considered as a measure of the increase in the proportion of homozygous individuals following an inbreeding mating (between relatives) [3]. The coefficient of consanguinity F can be calculated with the following methods: 1) pedigree 2) run of homozigosity (ROH); 3) genomic kinship matrix; 4) SNP genotyping [4, 5]. Inbreeding can occur in small closed populations due to mating between related animals. In a closed population, the decrease in the fraction of heterozygotes from one generation to the next may be referred to as ΔF. This value varies in relation to the size of the population: ΔF = 1 / 2Ne where Ne is the effective number or effective size of the population. In a population, Ne depends on the number of males (Nm), and on the number of females (Nf), in the following relationship:

$$1/Ne = 1/4\mathrm{Nf} + 1/4\,Nm; \Delta F = 1/8\mathrm{Nf} + 1/8\,Nm \tag{1}$$

The inbreeding coefficient, at a given t generation, can be calculated as a function of ΔF and t as:

$$\mathrm{F_t} = 1 - (1 - \Delta F)^{\mathrm{t}} \tag{2}$$

which shows the decrease (ΔF) of heterozygotes that occurs at each generation following inbreeding [6]. Lewis *et al.* [7] reported for 221 breeds of the UK Kennel Club a Ne that varies between 23.8 of the Manchester terrier breed to 918 of the Borzoi breed and an average value of F equal to 0.06. The deleterious effects of inbreeding are universally known. They can be summarized briefly in the increase in the frequency of all genetic defects and abnormalities (reproductive sphere, resistance to diseases, longevity, etc.). These findings are based on the results of experiments carried out on different breeds and for several generations. Leroy *et al.* [8] showed that the increase in inbreeding in the population has an effect on individual survival and litter size of different breeds. Deleterious effects begin to occur when the value of F is about 0.375. Lower values are not to be considered dangerous. It is worth noting that this is the level of inbreeding that is achieved in only two generations of full sibling mating. For this reason, it is recommended to avoid mating between close relatives. Consanguinity is influenced by the number of individuals used per generation [9]. As a general rule, individuals whose numbers are lower in the breeding population they exert a proportionately greater effect on consanguinity. This is true both in relation to the male/female ratio (depend more on the number of males) and the different numbers of breeders in the various generations. The actual number of breeding animals is the parameter used in small populations to determine the expected inbreeding coefficient. Since the less numerous sex is the most important, the actual number of the population can be calculated even if the number of the larger sex is not known (e.g. 2 males and the number of females is assumed to be infinite: 1/Ne = 1/ 4Nf = 1/4 (2) =1/8 * F = 1/16 = 0.0625). The family size is the number of offspring in each family who become parents in the next generation. In ideal conditions, the size of the population will remain constant in subsequent generations if each parent is replaced by another individual. In this case, the average number of offspring per parent is equal to 1 with an average family size of 2 (two parents). The Ne is also function of the variance of the family size. If males mate with more than one female, the number of offspring and thus the variance of the family size will differ between the two sexes. Several measures can be implemented to keep consanguinity within acceptable limits in the population: increase the number of breeders; mating of one male with a female (since the number within the sexes is the same, Ne will be maximized), reduce the variance size of the family (for a constant number of offspring for each family, the variance is equal to 0 and the Ne is double); avoid mating between siblings or cousins; avoid mating individuals in generations that overlap as inbreeding increases. If the management program includes the genetic improvement of one or more characters, selection must be carried out using selection indices that take into account of the level of relationship. The goal is to find the optimal number of offspring for each breeding animal and determine if a young animal (a candidate for selection) should be selected for breeding or not. This is done in an optimal way using the software EVA [10] that guarantees the achievement of the genetic progress and the maintenance an optimal genetic diversity in the population.

4. Breeding programs and strategies

The general actions to be taken in a program for the genetic improvement within a breed should include: 1) genomic identification and characterization of individuals, highlighting their potential in terms of their contribution to maintaining biodiversity, aptitude and use 2) monitoring of demographic parameters and assessment of the risk of reduced genetic variability 3) characterization and evaluation of the intra-breed genetic variability for proper management activities. Modern molecular techniques can be helpful for the improvement of management strategies, even for small breeds and for qualitative traits. The current hypothesis is to add molecular data to classical schemes (assisted selection) to improve their accuracy. The first step in planning an improvement program consists of: 1) a clear definition of the objectives 2) identification of the traits to be recorded 3) evaluation of the gene effect of the characters to be selected 4) estimate of the effect of the environment

Breed	DNA test	Physical test
Basenji	Fanconi	Eye assessment
		Hip score
	Progressive Retinal Atrophy	Thyroid
		Heart assessment
	Hemolytic anaemia	
	Pyruvate kinase deficiency	
	DNA inbreeding coefficient Factor	
	DNA identification Thyroid	
Border Collie	Neuronal Ceroid Lipofuscinosis	Elbow score
	Trapped Neutrophil Syndrome	Hip score
	Collie Eye Anomaly	Eye assessment
	Multi-Drug Resistance Gene 1	General vet check
	Imerslund-Grasbeck Syndrome	Chiropractor vet check
	Degenerative Myelopathy	Collie collaps
	Parentage (Orivet)	Hearing test
	Glaucoma	
German Shepherd	Degenerative Myelopathy	Hip score
	Ivermectin Sensitivity	Elbow score
	Long stock coat gene	
	Canine Renal Dysplasia	
	Dwarfism	
	Haemophilia	
Golden Retriever		
	Ichthyosis	Hip score
	Progressive Retinal Atrophy 1	Eye assessment
	Progressive Retinal Atrophy 2	Heart assessment
	Progressive Rod Cone Degeneration	Elbow score
		Dentition assesment

Table 1.
Genetic and physical testing used in genetic programs of common dog breeds.

(epigenetic effect) on the characters to be selected. In **Table 1** are reported the genetic and physical testing used in genetic programs of several dog breeds [11].

5. Genetic diseases and molecular diagnosis

In general, genetic diseases result from a mutation in a gene. In most cases, the mutations are traits that follow a simple Mendelian inheritance model (autosomal recessive, autosomal dominant or sex chromosome-linked character). Other hereditary diseases can be more complex and show reduced penetrance or multiple loci (multigenic disease). Genetic disorders can result from new mutations, but in most cases they result from old mutations passed on from one generation to the next. Mutated alleles can persist within a population for many reasons: 1. they can confer particular advantages in the state of heterozygotes; 2. the symptomatological signs can appear late 3. the mutation can be a recessive trait and therefore the defective allele can be spread in the population by healthy carriers. Without a mutation screening program, the carrier status can become evident only after the production of sick offspring.

The canine genome contains approximately 19,000 genes spread over 39 pairs of chromosomes (38 homologous chromosomes and 2 sex chromosomes). To date, nearly 400 hereditary diseases have been recognized in dogs. However, the precise ways in which these diseases are inherited are known for only about a third of them. In most cases, they are linked to autosomal recessive mutations. Bellumori *et al*. [12] report the prevalence of major genetic diseases in the United States for pure and mixed breeds. Pure breeds show more markedly some diseases including elbow dysplasia, cardiomyopathy, hypothyroidism and cataracts. The identification of the carriers can be implemented with the aid of two types of information: by pedigree or from a progeny test. In the first case, an animal showing the dominant phenotype (dominant phenotype) is known to be a carrier if one of the parents has the homozygous recessive genotype. In the second case, the farmer uses the information obtained from the offspring for the determination of the animal's genotype. Let us admit that a male is believed to be carrying a recessive allele. Special methods are required for the identification (and rejection) of carriers of the gene (suspected). This requires a reproduction test (test cross or progeny test) to determine whether the individual is dominant (suspected) or heterozygous. The genetic study of a hereditary diseases can follow additional strategies. Several genetic tests are now available for the identification of some hereditary disease [13]. The DNA-based diagnostic technique can be used to uniquely distinguish between sick and healthy subjects. These techniques allow the exclusion from reproduction of the carriers of frequent hereditary pathologies and they are a useful tool in validating the genealogical data reported in the pedigree.

6. Genomic analysis

6.1 Approach using candidate genes

The candidate gene approach consists in selecting a particular gene considered as the most likely site of a mutation. The main criteria for selecting a gene as a candidate are the following: 1) genes are selected because they are defective in similar animal species (usually humans or mice) 2) genes are selected based on their function. The analysis of the candidate gene consists in sequencing the entire gene and comparing two groups (healthy *vs* sick animals). However, the presence of a mutation in a gene is not in itself sufficient to identify the cause of the disorder. Unfortunately, for many genetic diseases the relative candidate gene has not been identified and very similar hereditary diseases can result from mutations on completely different genes. As an example, in the Bedlington terrier dog breed, the hereditary copper toxicosis is phenotypically identical to the Wilson's disease in humans. However, the gene involved in the human disease is not responsible for the disease in dogs. In conclusion, the approach with candidate genes has the advantage of allowing the identification of the specific mutation and therefore the creation of a targeted genetic tests.

6.2 Linkage analysis

The method of linkage analysis is based on completely different assumptions from the candidate gene approach. The main difference is that no assumptions are made about which gene is responsible for the disease, nor, more generally, the chromosomal tract involved. In this method, the whole genome is potentially subjected to analysis, without directing attention to any particular region. The search for the causal mutation takes place through the use of genetic markers whose chromosomal position is known. The more such markers are physically close to the mutation site, the more likely they will be co-inherited together with the mutation from one parental generation to the next. In a very simplified way, linkage analysis evaluates whether any of the variants of the markers appear in the population is associated with the presence of the disease. The ideal markers, and normally used to perform this type of study, are microsatellites, considered as practically ideal genetic markers because they are abundantly scattered throughout the genome and generally highly polymorphic. The number of microsatellites used to perform a linkage analysis is not fixed but generally the higher it is, the higher the probability that the study has success. This assumption derives from the fact that not directing attention towards specific genes and particular chromosomal portion, genome screening it must be as large as possible, i.e. it must contain the highest possible number of markers in order to understand the whole genome (so-called genome-wide screening). Generally, to perform a linkage study within a family tree informative are employed between 200 and 300 microsatellites using pedigrees with at least a hundred animals. For a given area of the genome, the probability of a recombination event occurring between a marker and a disease gene is directly proportional to their distance. The probability of occurrence of this event is expressed as a recombination fraction (θ). If θ is equal to 0.5, the marker and the disease gene are not linked and are therefore independently segregated. In other words, the probability that the marker and gene are inherited, associated or separated is identical. Conversely, if the marker and disease gene are linked together, the θ is less than 0.5. The lod score (Z) is the parameter which is used to estimate the linkage between 2 genetic loci. Z is the logarithm of the ratio between the probability that the 2 loci are linked ($\theta < 0.5$) and the probability that the 2 loci are randomly recombined ($\theta = 0.5$). Traditionally the linkage is accepted if the lod score is at least 3. Linkage analysis leads to the identification of a chromosomal region where the locus of the disease is probably located. The analysis must continue with the so-called refinement, that is, a further linkage analysis. Only later, the analysis proceeds through a gene candidate approach. All the genes of the region are identified and a sequence analysis is performed.

6.3 Genomic markers

6.3.1 Mitochondrial markers

Animal mtDNA is a cycular molecule ranging from 14,000 to 26,000 bp. The mtDNA codes for 13 proteins. Mitochondria contain most of the genes that code for cell energy production and electron transfer (NADH deydrogenase subunits, cytochrome oxidase subunits, ATPase 6 and 8, cytochrome b, rRNAr, RNA, 12S and 16S) [14, 15]. The choice of the sequence to be used for the genetic analysis depends on the phylogenetic hypothesis to be tested: D loop, sequences that evolve rapidly; cytocrome b, sequences that evolve moderately; Cytocrome oxidase I, sequences that evolve slowly. The mitochondrial control region (CR) sequence is the most popular marker. The mtDNA is uniparental (maternal line), characterized by a high evolution rate (5–10 times higher than nuclear genes) and the lack of introns and recombinations. The mtDNA is used to clarify the direction of hybridization and the incidence of introgression. In the case of hybridization, erroneous inferences can be obtained only using the evolutionary history of the females. In phylogeographic studies, information from various loci of the nuclear genome are also included [16–18]. The use of both parents allows a better analysis of the population structure.

6.3.2 Microsatellite markers

Nuclear microsatellites (one to six in tandem repeated nucleotides) are used in population genetics for the description of the population structure and kinship identification [19]. The reason for the wide use of microsatellites is due to the fact that are co-dominant, multi-allelic, highly reproducible and with a high resolution. The information per locus is about 10 times more informative than SNP markers. The most common repeats are di, tri and tetra-nucleotides. Microsatellite loci with a di-nucleotide motif are generally used, since they are easier to isolate and high density (on average every 30–50 kb) [20]. Microsatellites are also known as SSR (Simple Sequence Repeats) or STRs (Short Tandem Repeats). The maximum length is about 200 bp. Microsatellites are distributed throughout the genome with greater prevalence in non-coding regions. They are neutral in terms of selection. The typical problems encountered in the genotyping analysis are: homoplasy (condition of equality in the type and number of microsatellite repeats between two alleles) [21]; stutters (in the form of allelic pre-peaks); null alleles (NA) (possible mutations in the pairing site of the primers can prevent the pairing to the target sequence, causing the non-amplification of some alleles. The genetic analysis of microsatellites produce the following data: the distribution of allele frequencies for each microsatellite locus, the percentage of expected (H_E) and observed (H_O) heterozygosity, the estimates of the F_{st} values; Nei distances; conformity to the Hardy–Weinberg equilibrium (HWE) of the allele frequencies for each locus.

6.4 Next generation sequencing (NGS)

Starting in the 2000s, the analysis of SNPs led to the beginning of a new era in molecular genetics. The direct study of the genome using SNPs markers allows to integrate the genealogical information and to obtain high levels of accuracy in the estimation of the main genetic parameters of the population. The development of new sequencing techniques has made it possible to study the consequences of gene flow using a larger number of markers. At the beginning, the Sanger's technology was used to sequence the genomes of different animal species. This sequencing technique produces reads (>700 bp) with a very low error (<0.01%) and high cost (>600 US $ per Gb). This technique was subsequently improved through the use of the Celera assembler with a significant reduction in time and costs. New generation sequencing technologies (Next Generation Sequencing - NGS), also known as High Throughput Sequencing (HTS) technologies, have evolved rapidly offering an ever greater number of sequenced bases at a lower cost. In 2006, the first second-generation NGS technologies (Second-Generation Sequencing - SGS) appeared. Illumina (MiSeq, HiSeq and NovaSeq) is the most popular platform, due to its high performance and low cost. This technology is based on the fragmentation of DNA, amplification in multiple reactions in parallel, obtaining short reads, between 100 and 300 bp. Depending on the library, it is possible to sequence only one end of the fragment, single reads (single end) or both ends. The distance between the read pairs is called insert size (mate pair (2–5 kb); paired end (<1 kb)). Since 2013, the third-generation NGS techniques emerged, also known as the Single Molecule Sequencing (SMS) method. Single molecule sequencing produces long reads with higher costs (>2000 US$ per Gb). These techniques do not require the library amplification step and they are capable of directly sequencing a single DNA molecule, without applying any enzymatic or hybridization process. The main platforms of the third generation are Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). These platforms produce longer reads than the previous ones (5–50 kb) but have a much higher error rate. The Pacbio platform routinely generates reads with an N50 > 1 Mb and it has recently reduced the error rate with a new technique (circular consensus sequencing) and the production of high fidelity reads of 15 Kb. The most popular softwares used for the bio-informatic analysis are Canu; Marvel and Mecat Flye. Then, results obtained are cleaned with some software such as Racon; Nanopolish and Pilon. **Figure 1** shows an example of workflow using long reads.

After identifying the putative protein coding regions (CDSs), UCEs (Ultra Conserved Elements), it is possible to infer the correct reading pattern (Open Reading Frames, ORF) and translate the nucleotide sequences into amino acids [22]. In this way, we will obtain the set of predicted proteins encoded by the study genome. BLAST (nucleotide, protein, translated, genomes), HMMER or InterProScan databases can be used to functionally annotate these proteins. InterProScan provides the information on functional processes (GO terms) and metabolic pathways (KEGG). Once the functional and structural annotation has been obtained, the analysis of the functional elements of interest such as polymorphic positions or genes with differential expression can be performed. **Figure 2** shows an example of workflow for the genomic annotation analysis.

Orthmcm, Orthofinder; EggNog sofwares can be used for the homology analysis. Several studies, in recent years, have shown that the best way to understand complex systems (for example diseases) is to combine different omic data together. **Figure 3** shows a detailed analysis using omic data (genomic, transcrptomics, proteomics and metabolomics).

6.5 Reduced representation genome sequencing (RRGS)

Several new techniques have been developed in the last decade. The most popular is the restriction-site-associated DNA sequencing (RAD-Seq) [23] and the genotyping by sequencing method (GBS) [24]. The main advantage of RRGS methods is that it reduces the cost of analysis with an high coverage compared to the traditional sequencing methods. The *de novo* analysis does not require a priori knowledge of the reference genome sequence. Several applications of the RAD-Seq methods have been reported: population genetics studies (phylogenetic and

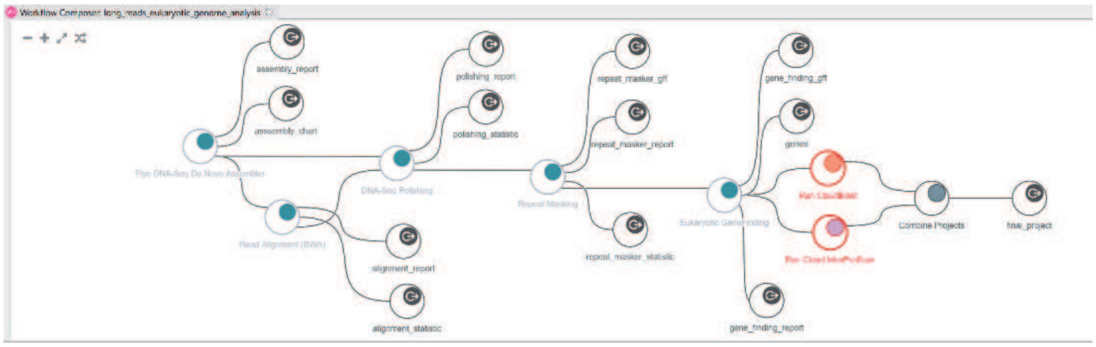

Figure 1.
Example of NGS bioinformatic analysis (long read sequencing).

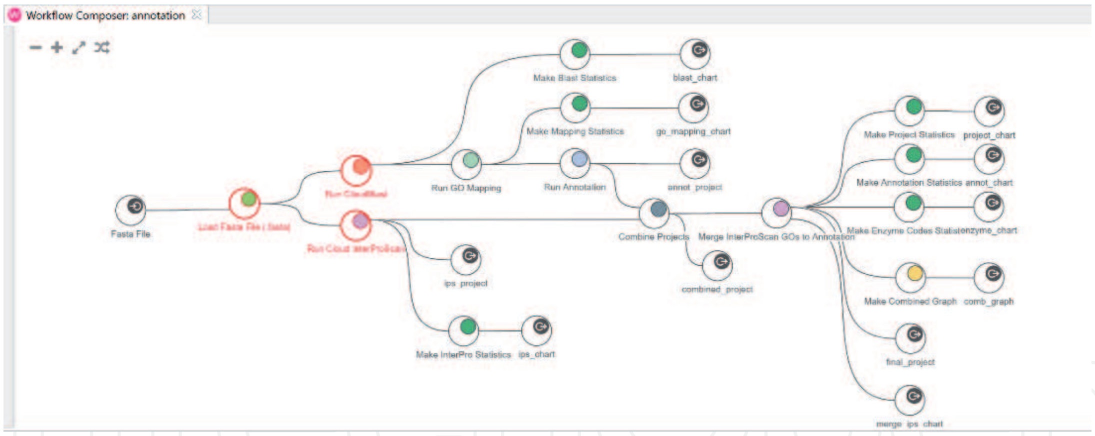

Figure 2.
Example of NGS annotation analysis.

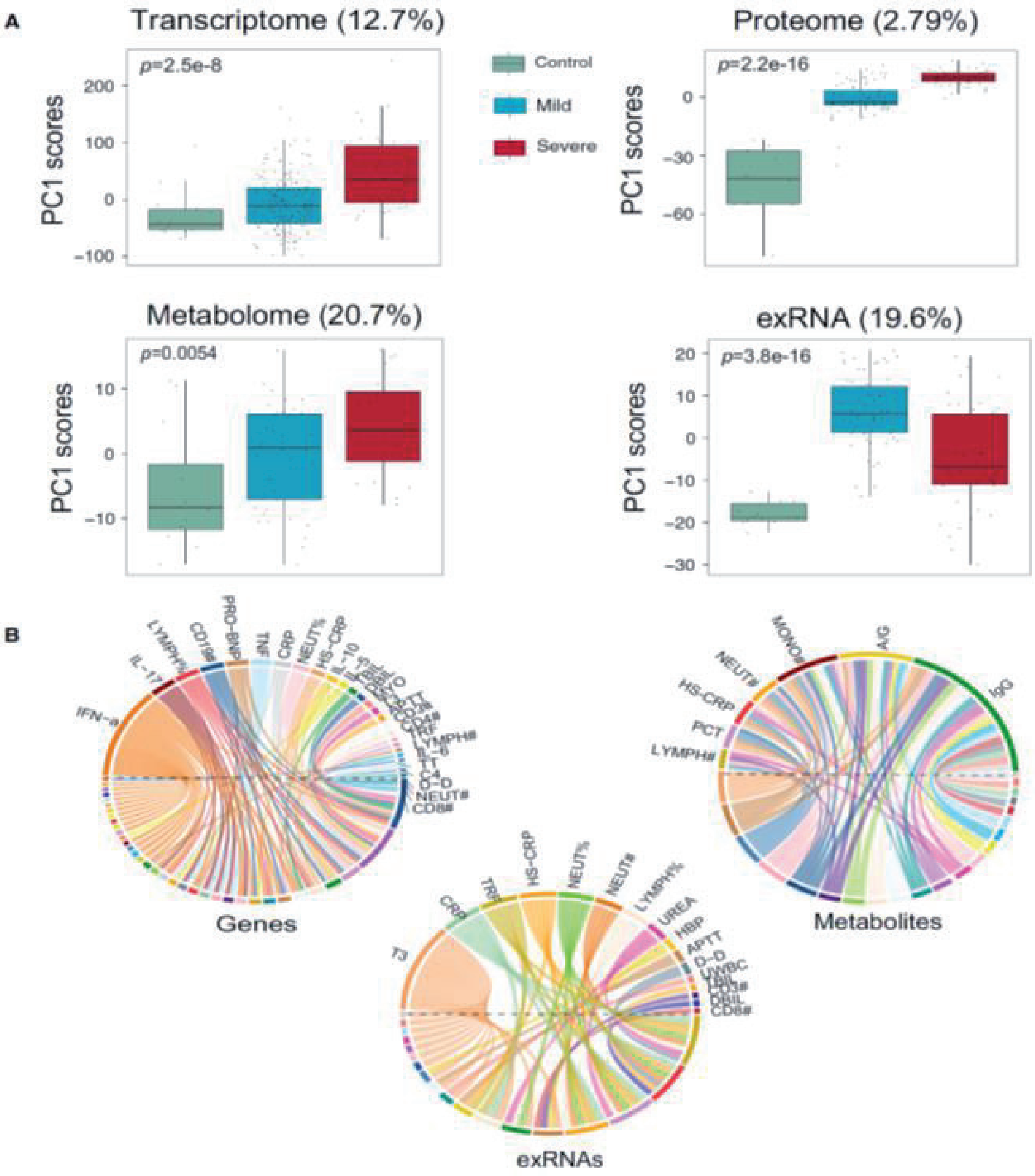

Figure 3.
Example of OMICs analysis (genomics,transcrptomics, proteomics and metabolomics).

phylogeographic), linkage mapping (fine scale) and genome scaffolding [25]. To avoid or reduce the bias, some variations of the original RAD Seq protocol have been proposed: ddRAD, ezRAD, 2b-RAD. Classic RAD reads are obtained between the restriction site and a random site while the ddRAD reads are obtained between two restriction sites. In particular, the ddRAD-SEq method increases the number of samples per sequencing line and develops a tagging approach by combining pairs of adapters. Another advantage is the selection of the fragment sizes. This reduces duplicate sampling of a region, thus requiring only half the reads to effectively achieve high levels of confidence for each SNP associated with a restriction enzyme cleavage site. All these properties make the ddRAD-Seq method robust, allowing to search for a smaller number of reads. The bio-informatic analysis of RAD-Seq data includes the following phases: quality control, trimming, reference genome or *de novo* mapping methods, SNP filtering/annotation. The results of RAD-Seq analysis are analyzed with different softwares such as Stacks, Ig-Tree, Uneak (Tassel), Pyrad; Ddocent; 2brad and Aftrrad. The most popular software is the Stacks program. RRL methods, in relation to the production of short reads, are not very useful for the construction of phylogenetic trees but are generally used for the analysis of SNPs.

6.6 Genome-wide genotyping arrays

In the recent years, the availability of massive genomic data obtained from the last generation sequencing techniques allowed the efficient identification of a large number of SNPs [26]. The GWAS is a method of investigation that allows to examine the entire genome by analyzing the single nucleotide polymorphism of genomic markers (SNPs) with the use of high density SNP arrays [27] (the last versions Illumina Canine HD SNP 170 K have hundreds of thousands of SNPs distributed throughout the genome). The study identified the genetic structure of the populations present in Italy and the selection signatures. Reduction of genotyping costs is achieved using inference methods such as the imputation. Imputation techniques allow to transfer information from DNA from high density bead chips to low density ones.

6.6.1 Genome-wide association studies

The genome-wide association studies (GWAS) have been proposed as an effective approach for the identification of many causative mutations and genetic factors that constitute the main traits. Unlike linkage studies, which consider the phenomenon of inheritance of chromosomal regions linked to the presence of a trait within a family, association studies consider instead the difference between the frequency of SNPs affecting the trait of interest. Association studies may be conducted through two approaches: direct and indirect. A direct association study is to catalog and test one by one all the possible causal mutations. However, the direct approach presents some practical problems. This strategy involves genome-wide identification of all genes (up to 19,000 genes) as well as of all SNPs. For these reasons, the use of the direct method is limited to a few cases and it has almost always replaced with the application of the indirect method. The indirect strategy avoids the need to catalog all mutations that could potentially give predisposition to a given trait and instead relies on the association between a giver phenotype and markers located near a strategic locus. These associations are obtained from studies of linkage disequilibrium (LD) between marker loci. The indirect strategy, then employs a dense map of polymorphic markers to explore the genome in a systematic way. The choice of markers differentiates further the indirect approach in two different strategies. In the first, markers are chosen very close to exon regions of known genes. The second employs markers located in large regions, virtually anywhere in the genome, thus considering the chromosomes in their entirety, including intronic regions. The choice of the marker falls on bi-allelic SNPs because of their high frequency in the animal genomes, for the low rate of mutation and for the ease with which it can be analyzed. Linkage means the presence of genes in closed loci on the same chromosome. LD is a combination of alleles at two or more loci that occurs more often than it does happen by chance. Two markers are in LD when they occur together in the same individual more frequently than would be expected by chance. The presence of a LD thus indicates co-segregation of two markers. In generally, the LD between two SNPs decreases with the physical distance and the extent of LD varies strongly among the regions of the genome. LD analysis is a valuable tool for fine mapping. Doherty [28] conducted an GWAS analysis using 9700 SNPs on 72,000 dogs (63 breeds). Eight SNPs were significantly correlated with the live weight and five SNPs with cancer mortality. Plassais [29] analyzed the genomes (WGS) of 722 dogs and used the Illumina canine HD SNP BeadArray to identify over 91 million SNPs. In this way the main SNPs coding for body weight and main morphological characters were identified. In **Table 2** is reported an example of SNP genotyping using a SNP chip array in dogs [30].

6.7 Scans for selective sweeps

The domestic dog is thought to be the most recent species of the canine family, within which three phylogenetic groups, or clades, are distinguished: the domestic dog belongs to the same clade as the gray wolf, coyote and jackals [31]. It is thought that the dog appeared about 40,000 years ago, and that the first steps in its domestication took place in East Asia [32]. Most of the domestic breeds we know today, however, are the result of human selection over the past two or three centuries. Many of the most popular modern breeds were created in Europe in the 19th century. Some of the breeds were already present in the ancient world as the greyhound and the dog of the pharaohs. Studies conducted at the genomic level have highlighted a stratification of genetic variability within dog breeds. The recent sequencing methods and the use of SNP arrays allow the screening of the whole genome for the presence of signatures of selection. Sequencing data are aligned to the reference genome to identify selective sweeps. The presence of genes with

Genomic analysis
Illumina CanineHD SNP chip (San Diego, CA)
Genotype SNP calls using Illumina's Genome Studio
Selection of samples with a >90% SNP call rate
SNPs with Gentrain scores >0.4
Minor allele frequency >1%
Bio-statistical analysis
FlashPCA - Principal components (PCs)
Admixture - two to ten adjusted cluster ancestry models
Beagle – calculation of IBD haplotype sharing analysis and phasing
VCFtools – calculation of the inbreeding coefficient
TreeMix – construction of a maximum likelihood tree - windows of 1000 SNPs using the flags -bootstrap and -k 1000 functions
R studio – construction of graphs and plots for all the analyses

Table 2.
SNP genotyping using a SNP chip array in dogs.

Dataset of 268 dogs representing 130 breeds
Phenotypes used in the study: canids catalog, kinship, aggressiveness, boldness, bulky, drop ears, furnishing, hairless, height, large ears, lengh of fur, life span, long legs, muscled, tail curl, weight, white chest, white head
GWAS
Samples with ≥10x coverage, selecting two males and two females
Gemma - linear-mixed model methods; elimination of variants with missing value >1
R Studio – Manhattan correlation and box-plots
Identification of positively selected genes
Vcftools60
Beagle - infer the haplotype phase
Xpclr - phased genotype input; non-overlapping windows (50 kb), 600 SNPs within each window; correlation level cutoff of 0.95.
XP-EHH - splitting the genome into non-overlapping segments of 50 kb

Table 3.
Example of GWAS and selective sweep analysis in dogs.

a large number of outliers indicates a positive or negative effect of selection. A genome scan approach can be used to distinguish genome-wide processes (expected to mainly reflect demographic histories) from processes at individual loci. Genome scans may suffer from inflated numbers of false positives under hierarchical spatial structure coupled with isolation by-distance dynamics. In the case of positive selection, there is an increase in the fitness of the population due to a new (or rare) mutation. In the case of hard sweeps, there is an increase in the frequency of some variants and in the linkage disequilibrium. Kim et al. [33] compared 127 dogs (sport-hunting vs. terrier) for sporting characteristics. Results of the study showed the main SNPs (cardio-circulatory, muscular and neuronal systems) and selection signature that are involved in the sport-hunting breeds. In **Table 3** is reported an example of GWAS and selective sweep analysis in dogs [29].

7. Genome applications in the canine sector

The canine genome project was launched in the early 1990s. After some preliminary results, in 2003, a fist sequence of the dog's genome was obtained from a female boxer which is now the reference sequence for the dog [34]. The availability of a high quality canine genome has revolutionized the way in which geneticists operate. The first version of the boxer's genome, carried out with a coverage of 7.5x, covered nearly 99 percent of the animal's genome. The genome sequence provided a first description of the organization, number of genes and the presence of repeated elements. To some surprise, they found a high presence of short interspersed nuclear elements (SINEs) throughout the genome, sometimes located in locations from which they could affect gene expression. For example, the insertion of a SINE into the gene encoding the hypocretin receptor (a neuropeptide hormone found in the hypothalamus) causes narcolepsy in the Doberman. Similarly, the insertion of a SINE element into the silv gene, which is known to be linked to the pigmentation process, is responsible for a particular mottled color called merle. The 2003 sequence comprises approximately 2.4 billion of bases and revealed the existence of approximately 19,000 genes. For about 75% of genes, the homology (resulting from shared ancestral material) between the dog, man and mouse is very high. The study also found that many genes have no gaps in their sequence, which is beneficial if you would like to study the correlation between a given gene and a disease. During its evolution, the dog's genome has accumulated more than two million of SNPs. These markers are proving crucial in understanding the role of genetic variability within one breed and in different breeds. SNPs, analyzed by means of DNA microarrays or bead arrays, can make an important contribution to GWAS (association studies) aimed at identifying the genes responsible for complex traits in dogs. A microarray with around 170,000 SNPs is currently available. By comparing data from dogs with a certain disease with healthy individuals, it is possible to quickly identify the genes responsible for the disease. Dog breeds differ not only in the overall body size but also in leg length, head shape and many other morphological characteristics. In the dog, the phenotypic variability of several traits is very high compared to the other living terrestrial mammals. The first molecular study on the genetic aspects of dog morphology was conducted at the University of Utah [35, 36]. Called Georgie Project (in memory of a dog), the study focused on the Portuguese water dog breed, ideal for this type of study because it comes from a small number of ancestors. In the project, DNA samples of more than a thousand dogs were collected. A completed genome scan using 500 microsatellite markers was carried out. For these animals, in addition to the genealogical and medical data, more than 90 anatomical measurements were obtained from a series of five radiographs taken on each animal during the first phase of the study. Based on the analysis of these data, four primary main components (CP) have been identified (**Figure 4**).

The analysis of the genome scans and principal components (CPs) revealed 44 putative QTLs (quantitative trait loci associated with a particular quantitative trait) on 22 chromosomes. QTLs are identified by means of a complicated statistical analysis and identify the genome regions that contribute to the expression of a certain trait. Of particular interest is the gene CFA15 on chromosome 15 which showed a strong association with the body size. Although, it is only one of seven loci thought to affect the body size, it was chosen as the starting point. To find the gene CFA15, several SNPs were identified and then the resulting set of genome-wide markers were genotyped. The distribution of these markers showed a single peak near the insulin-like growth factor-1 (IGF 1) gene, which codes for insulin-like growth factor which is known to code for the body size in humans and mice. IGF 1 was analyzed in detail, discovering that there are only two specific combinations of alleles (called haplotypes) and one of them is present in 96% of the population. The haplotype associated with the small size was called B, while the one associated

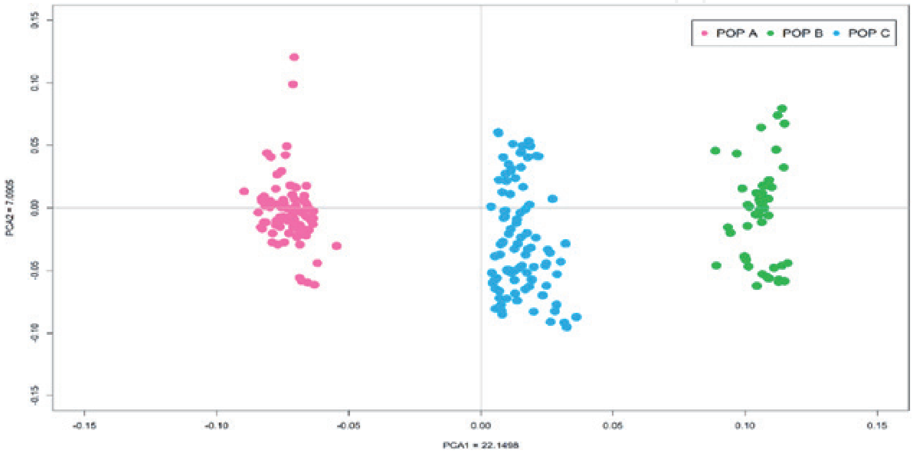

Figure 4.
Example of PCA (principal component analysis) of genotypic data (autosomal) of three dog populations.

with the largest size was called I. Homozygous dogs for the haplotype B showed a smaller average body size while, dogs homozygous for I were larger. Heterozygous dogs showed an intermediate size. The Georgie Project is important for the number of genes discovered. In addition to the genes related to the head shape, body size, leg length and many other traits, additional genes were discovered that control the sexual dimorphism [37, 38]. This dimorphism is observed in almost all mammals but its mechanisms it is not yet fully known. Indeed, it was found a gene on chromosome 15 which interacts with other genes to make males larger and females smaller. On average, females of the Portuguese water dog breed are 15% smaller than the males.

8. Future perspectives

In the past fifteen years, tremendous progress has been made in dog genomics [39–41]. Several genetic aspects of cancer, heart disease, hip dysplasia, vision and hearing problems in dogs have been investigated and studied in detail. Genome-wide associative studies have made possible to identify several genes associated with diseases, morphological and behavioral traits. The Dog10K project will produce 10,000 new dog genomes (20x) within five years [42]. The mapping of disease-associated genes will hopefully lead to the production of new genetic tests and improve the management of running breeding programs, which in turn will produce healthier and longer-living dogs. It will be easier to select for specific physical traits such as the size or coat color. Finally, perhaps we will be able to identity which genes are responsible for the typical behaviors of each breed.

Author details

Edo D'Agaro*, Andrea Favaro and Davide Rosa
Department of Agricultural, Food, Environment and Animal Sciences,
University of Udine, Udine

*Address all correspondence to: edo.dagaro@uniud.it

References

[1] Willis MB,. Genetics of the dog. Howell Book House. 1989; New York, USA.

[2] Berryere TG, Kerns JA, Barsh GS, Schmutz SM. Association of agouti allele with fawn or sable coat color in domestic dogs. Mamm Genom. 2005; 16: 262-272.

[3] Mellanby RJ, Ogden R, Clements DN, French AT, Gow AG, Powell R, Corcoran B, Schoeman JP, Summers KM. Population structure and genetic heterogeneity in popular dog breeds in the UK. Vet J. 2013; 196(1):92-97.

[4] Zhang Q, Calus MPL, Guldbrandtsen B, Lund MS, Sahana G. Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. BMC Genet. 2015; 16:88.

[5] Wang J. Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? Theor Popul Biol. 2016; 107: 4-13.

[6] Jansson M, Laikre L. Pedigree data indicate rapid inbreeding and loss of genetic diversity within populations of native, traditional dog breeds of conservation concern. Plos one. 2018; 13(9): e0202849.

[7] Lewis TW, AbhayaratneBM, BlottSC. Trends in genetic diversity for all Kennel Club registered pedigree dog breeds. Canine Genet Epidemiol. 2015; 2:13.

[8] Leroy G, Phocas F, Hedan B, Verrier E, Rognon X. Inbreeding impact on litter size and survival in selected canine breeds. Vet J. 2015; 203:74-78.

[9] Leroy G. Genetic diversity, inbreeding and breeding practices in dogs: Results from pedigree analyses. Vet J. 2011; 189:177-182.

[10] Berg P, Nielsen J, Sørensen MK. EVA: Realized and predicted optimal genetic contributions. WCGALP. 2006; 246.

[11] Czerwinski V, McArthur V, Smith B, Hynd P, Susan Hazel S. Selection of breeding stock among Australian purebred dog breeders, with particular emphasis on the dam. Animals. 2016; 6: 75.

[12] Bellumori TP, Famula TR, Bannasch DL, Belanger JM; Oberbauer AM, Bellumori TP, Famula TR Bannasch DL, Belanger JM, Oberbauer AM. J Am Vet Med Assoc. 2013; 242:1549-1555.

[13] Oberbauer AM, Belanger JM, Bellumori T, Bannasch DL, FamulaTR. Ten inherited disorders in purebred dogs by functional breed groupings. Canine Genet Epidemiol. 2015; 2:9.

[14] Avise JC. Perspective: conservation genetics enters the genomics era. Cons Genet. 2010; 11:665-669.

[15] Duleba A, Skonieczna K, Bogdanowicz W, Malyarchuk B, Grzybowski T. Complete mitochondrial genome database and standardized classification system for *Canis lupus familiaris*. Forensic Sci Int-Gen. 2015; 19:123-129.

[16] Ersmark E, Klütsch CFC, Chan YL, Sinding M-H S, Fain SR, Illarionova NA. From the past to the present: Wolf phylogeography and demographic history based on the mitochondrial control region. Front Ecol Evol. 2016; 4:134.

[17] Hendrickson M, Remm J, Pilot M, Godinho R, Stronen AV, Baltrūnaité L. Wolf population genetics in Europe: a systematic review, meta-analysis and suggestions for conservation and management. Biol Rev. 2017; 92:1601-1629.

[18] Toth G, Gaspari Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 2000*;* 10:967-981.

[19] Zane L, Bargelloni L, Patarnello T. Strategies for microsatellite isolation: a review. Mol Ecol. 2002; 11:1-16.

[20] Estoup A, Rousset F, Michalakis Y, Cornuet J-M, Adriamanga M, Guyomard R. Comparative analysis of microsatellite and allozyme markers: a case study investigating microgeographic differentiation in brown trout (*Salmo trutta*). Mol Ecol. 1998; 7:339-353.

[21] Grabherr M, Haas B, Yassour M,LevinJ, Thompson D, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q.Trinity:reconstructing full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol. 2011; 29(7):644-652.

[22] Martin SH, Davey JW, Jiggins CD, Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Mol. Biol. Evol. 2015; 32:244-257.

[23] Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, Casler MD. Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. Plos Genet. 2013; 9(1):e1003215.

[24] Baird NA, Etter PD, Atwood TS et al. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. Plos one. 2008; 3: e3376.

[25] Daetwyler HD, Hickey JM, Henshall JM, Dominik S, Gredler B. Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. Anim Prod Sci. 2010; 50:1004-1010.

[26] Ramos AM, Crooijmans RPMA, Affara NA, Amaral A J, Archibald AL.   ration sequencing technology. Plos one. 2009; 4: e6524.

[27] Meuwissen THE. Genomic selection: marker assisted selection on a genome wide scale. J Anim Breed Genet. 2007; 124: 321-322.

[28] Doherty A, Lopes, Ford CT, Monaco G, Guest P, de Magalhães JP. A scan for genes associated with cancer mortality and longevity in pedigree dog breeds. Mamm Genome. 2020; 31:215-227.

[29] Plassais J, Kim j, Davis BW, Karyadi DM, Hogan AN, HarrisAC, Decker B, Parker HG, Ostrander EA. Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. Nature Comm. 2019; 10:1489.

[30] Ali MB, Evans JM, Parker HG, Kim J, Pearce-Kelling S, Whitaker DT, et al. (2020) Genetic analysis of the modern Australian labradoodle dog breed reveals an excess of the poodle genome. PLoS Genet. 2020; 16(9): e1008956.

[31] Vonholdt BM, Pollinger JP, Lohmueller KE, et al. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature. 2010; 464:898-902.

[32] Wayne RK, Ostrander EA. Lessons learned from the dog genome. Trends in Genet. 2007; 23:11.

[33] Kim J, Williams FJ, Dreger DL, Plassais J, Davis BW, Parker HG, Ostrander EA. Genetic selection of athletic success in sport-hunting dogs. Pnas. 2018; 115:30.

[34] Parker HG, Dreger DL, Rimbault M, Davis BW, Mullen AB, Carpintero-Ramirez G, Ostrander EA. Genomic analyses reveal the influence of geographic origin migration, and hybridization on modern dog breed development. Cell Rep. 2017; 19: 697-708.

[35] Chase K, Carrier DR, Adler FR, Jarvik T, Ostrander EA, Lorentzen TD, Lark KG. Genetic basis for systems of skeletal quantitative traits: Principal component analysis of the canid skeleton. Pnas. 2002; 99(15):9930-9935.

[36] Chase K, Carrier DR, Adler FR, Ostrander EA, Lark KG. Interaction between the X chromosome and an autosome regulates size sexual dimorphism in Portuguese Water Dogs. Genome Research. 2005; 15:1820-1824.

[37] Lindblad-Toh K., Wade CM, Lander ES. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 2005; 438:803-819.

[38] Rimbault M, Beale HC, Schoenebeck JJ, Hoopes BC, Allen JJ , Kilroy-Glynn P, Wayne RK, Sutter NB, Ostrander EA. Derived variants at six genes explain nearly half of size reduction in dog breeds. Genome Res. 2013; 23:1985-1995.

[39] Parker HG, Meurs KM, Ostrander EA. Finding cardiovascular disease genes in the dog. J Vet Cardiol. 2006; 8:115-127.

[40] Parker HG and Ostrander FA. Canine genomics and genetics: Running with the pack. Plos Genet. 2005;1(5):e58, 2005.

[41] Dreger DL, Davis BW, Cocco R, Sechi S, Di Cerbo A, Parker HG, et al. Commonalities in development of pure breeds and population isolates revealed in the genome of the Sardinian Fonni's Dog. Genetics. 2016; 204:737-55.

[42] Ostrander EA, Wayne RK, Freedman A H, Davis BW. Demographic history, selection and functional diversity of the canine genome. Nat Rev Genet. 2017; 18:705-720.