The canine era

van Steenbeek, F. G.

2016-10


http://hdl.handle.net/10138/167363
https://doi.org/10.1111/age.12460

Downloaded from Helda, University of Helsinki institutional repository.
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.
Please cite the original version.
The canine era: the rise of a biomedical model

F. G. van Steenbeek*, M. K. Hytönen†, P. A. J. Leegwater* and H. Lohi†

*Department of Clinical Sciences of Companion Animals, Faculty of Veterinary Medicine, Utrecht University, Yalelaan 104, 3508 TD, Utrecht, the Netherlands. †Research Programs Unit, Molecular Neurology, Department of Veterinary Biosciences 00014, Folkhälso Research Center, University of Helsinki, Helsinki, Finland.

Summary
Since the annotation of its genome a decade ago, the dog has proven to be an excellent model for the study of inherited diseases. A large variety of spontaneous simple and complex phenotypes occur in dogs, providing physiologically relevant models to corresponding human conditions. In addition, gene discovery is facilitated in clinically less heterogeneous purebred dogs with closed population structures because smaller study cohorts and fewer markers are often sufficient to expose causal variants. Here, we review the development of genomic resources from microsatellites to whole-genome sequencing and give examples of successful findings that have followed the technological progress. The increasing amount of whole-genome sequence data warrants better functional annotation of the canine genome to more effectively utilise this unique model to understand genetic contributions in morphological, behavioural and other complex traits.

Keywords animal model, dog, genetic tools, next generation sequencing

Unique population history and genetic architecture facilitate gene mapping
Domestication of the dog occurred thousands of years ago. The exact period in history and the region of origin have been extensively discussed and are still under debate (Savolainen et al. 2002; Pang et al. 2009; Boyko 2011; vonHoldt et al. 2011; Axelsson et al. 2013; Thalmann et al. 2013; Wang et al. 2013, 2016; Freedman et al. 2014; Shannon et al. 2015). The dog was used as a sentry, a food source and for hunting, but its role as companion animal swiftly resulted in the removal of the dog from its pack to become a pet. The French Philosopher Voltaire was the first to launch the term ‘man’s best friend’:

‘It seems that nature has given the dog to man for his defense and for his pleasure. Of all the animals it is the most faithful: it is the best friend man can have’ (Voltaire, Dictionnaire Philosophique, 1764).

Selection and breed formation have resulted in a broad variety of dog phenotypes. Due to man’s influence, the appearance of the dog varies greatly; for instance, there is up to a 40-fold difference in size between the Great Dane and the Chihuahua (Wayne & Ostrander 2007). The vast differences among breeds are the result of just a small number of gene variants that define, for instance, coat colour (Karlsson et al. 2007; Dreger et al. 2013), coat texture (Salmon Hillbertz et al. 2007; Cadieu et al. 2009) and size (Sutter et al. 2007), indicating that the genetic architecture is simplified and involves large gene effects. For example, the background of coat colour of Labrador Retrievers, Golden Retrievers and Flat-coated Retrievers is defined by two genomic loci (Fig. 1) (Lavrijsen et al. 2014a, b). The first locus on chromosome 5 is in close vicinity to the melanocortin 1 receptor (alpha melanocyte-stimulating hormone receptor) gene (MC1R), and the second locus on chromosome 11 includes the tyrosinase-related protein 1 gene (TYRP1). Both genes are known to be involved in mammalian coat coloration (Éverts et al. 2000; Newton et al. 2000; Schmutz et al. 2002).

Several examples demonstrate how selection for phenotypic characteristics results in severe health issues for canines. A brachycephalic phenotype selected for in breeds such as the Boxer, Boston Terrier, Pekingese and Bulldog causes breathing abnormalities because of the soft tissue blockage of the airways during respiration (Packer et al. 2015). Cavalier King Charles Spaniels are commonly affected by syringomyelia by which fluid-filled cavities develop within the spinal cord near the brain (Parker et al. 2011). This disease results from the shortening of the caudal skull, termed Chiari-like malformation (Cerda-Gonzalez et al. 2009).
The characteristic wrinkled skin of Shar-Pei is associated with familial Shar-Pei fever (FSF), characterised by periodic fever and inflammation similar to human familial Mediterranean fever (French FMF Consortium 1997). High concentrations of hyaluronan in the skin cause the wrinkling phenotype (Zanna et al. 2008). FSF was found to be associated with the expression level of the gene hyaluronan synthase 2 (HAS2), and the copy number of a 16.1-kb repeat fragment close to the gene has been correlated with both HAS2 expression and disease (Olsson et al. 2011). The repeat is expected to contain a regulatory element that regulates the gene expression of HAS2. An increased copy number of the duplication gives rise to an increased copy number of the potential enhancer elements, increasing the expression of HAS2 and consequently the elevating levels of hyaluronan. The large amount of hyaluronan is postulated to be responsible for the fever and inflammation, and, as a result, selection for appearance directly causes a selection for disease risk.

The extent of white colour in Dalmatians correlates with congenital sensorineural deafness (Stritzel et al. 2009). Involvement of microphthalmia-associated transcription factor (MITF) defines the extent of extreme white found in the coat (Karlsson et al. 2007; Baranowska Korberg et al. 2014) due to its function in melanogenesis. MITF is also known to be secondarily associated with hearing loss in relation to the regulation of pigmentation (Waardenburg/Tietz syndrome) (TIETZ 1963) and deafness in mice (Motohashi et al. 1994; Yoshida et al. 1996). A causative mutation for deafness in Dalmatians has not yet been found.

The dog is a relatively large animal as compared to mice, and many canine disorders resemble the corresponding human conditions better than do induced rodent models. Numerous examples exist across disease groups, including epilepsy (Lohi et al. 2005), narcolepsy (Lin et al. 1999), developmental conditions (Bannasch et al. 2010; Hytonen et al. 2012; Wolf et al. 2015), skin diseases (Grall et al. 2012; Tengvall et al. 2013), eye disorders (Sargan et al. 2007; Ahonen et al. 2013) and cancer (Arendt et al. 2015). Unique population history and breed structure have resulted in a genetic architecture that facilitates the exploration of the genetic basis of simple and complex disorders (Ostrander & Kruglyak 1997). Extensive linkage disequilibrium combined with low haplotype diversity within breeds requires a lower number of markers to test the genome and also fewer samples for statistical power in genome-wide association studies (GWAS) as compared to human GWAS (Sutter & Ostrander 2004). An informative example is the canine systemic lupus erythematosus study that identified a limited set of risk loci but with each of them having a high contribution to the disease (Wilbe et al. 2010). Additionally, interbreed mapping with shared conditions provides a powerful approach to narrow down the associated loci for the identification of the causal variant, as demonstrated by the identification of a FOXI3 frameshift mutation in canine ectodermal dysplasia (Drogemuller et al. 2008). A 1.7-Mb disease locus in Chinese Crested Dogs was narrowed down to a 102-kb region by interbreed analysis in Peruvian and Mexican Hairless breeds. To date, there are over 200 known pathogenic variations across disorders in dogs, and this number is rapidly increasing (OMIA database).

**Evolving genetic tools**

The annotation of the canine genome has allowed for a rapid development of powerful genomic tools for gene discovery (Fig. 2). Sparse microsatellite experiments (Clark et al. 2004) have been replaced by high-density SNP arrays and resequencing experiments. One of the first successes of the microsatellite approach was the localisation of the gene causing copper toxicosis in Bedlington Terriers with a set of 213 microsatellite markers (Yuzbashyan-Gurkan et al. 1997). Eventually, this led to the
The discovery of the causative mutation in COMMD1 (van De Sluis et al. 2002). Other microsatellite experiments were used to localise the first canine epilepsy gene (Lohi et al. 2005) and the dwarfism mutation discovery in German Shepherds (Voorbij et al. 2011). The first higher coverage canine DNA sequence became available in 2004. Sequencing of the DNA of the Boxer Tasha by the Broad Institute resulted in a 7.5 \times coverage genome, which was a significant step forward in canine genetic research (Lindblad-Toh et al. 2005). More than 30 million DNA sequence reads were generated in just 6 months at a cost of 30 million dollars and resulted in the dog being one of the first mammals to have its genome sequenced. Owing to the availability of the reference genome, the IGF1 gene could be rapidly identified as a determinant of dog size (Sutter et al. 2007).

Prior to the boxer reference sequence, a low coverage poodle genome was already available (Kirkness et al. 2003). The annotations of the canine genomes and the subsequent identification of large numbers of SNPs have enabled the development of series of SNP arrays with an increased resolution in gene mapping. The first arrays contained only approximately 1500 SNPs, but provided faster experiments and comparable power to that of the existing panel of 256 microsatellite markers, as evidenced by the identification of white coat spotting in Boxers on chromosome 20 (Leegwater et al. 2007). The development of higher-density SNP arrays allowed genome-wide association analyses beyond family relationships and across breed populations. Again, breed-defining traits such as ridgeback, white spotting, chondrodysplasia and coat type were among the first to be mapped with these arrays (Karlsson et al. 2007; Cadieu et al. 2009; Parker et al. 2009). The further development of higher-resolution SNP assays (Ke et al. 2011) resulted in the successful application of the Affymetrix 50k platinum panel (Awano et al. 2009; Roque et al. 2011, 2012; Seppala et al. 2012) and the Illumina 22k array (Kyostila et al. 2012), both of which in turn were replaced by the Illumina CanineHD array (Vaysse et al. 2011) with 173k markers (eurolupa.eu).

These developments have significantly facilitated the gene discovery across traits and breeds, with high relevance for human medicine (Lequarre et al. 2011). A recent example is the identification of the CCDC39 defect first in canine primary ciliary dyskinesia (PCD), leading to the discovery of many loss-of-function mutations in the human orthologue and PCD (Merveille et al. 2011). Another example, this time in Golden Retrievers, is the identification of the PNPLA1 mutation in autosomal recessive congenital ichthyosis. This canine discovery led to the discovery of two causative mutations in the human orthologue in human ichthyosis (Grull et al. 2012).

Is there still a need for higher-resolution SNP arrays to facilitate disease gene mapping in dogs? The long linkage disequilibrium structure and the limited number of haplotypes within breeds allow for the efficient testing of dog genomes with much fewer markers in comparison with the human genome (Lindblad-Toh et al. 2005). The current

© 2016 Stichting International Foundation for Animal Genetics, 47, 519–527
et al. revealed the mutation for progressive renal atrophy in the Whole exome sequencing and other conditions (Vaysse et al. 2014a,b) and other regions associated with mast cell tumours could be identified. By genotyping cases from Golden Retrievers can canine mast cell tumours, this approach has been proven successful. By genotyping cases from Golden Retrievers originating from different populations, distinct chromosomal regions associated with mast cell tumours could be identified (Arendt et al. 2015). Sequence capture of the different associated regions and the subsequent sequencing identified a SNP in GUSB as a cause of the disease. Even in complex disorders, such as canine mast cell tumours, this approach has been proven successful. By genotyping cases from Golden Retrievers originating from different populations, distinct chromosomal regions associated with mast cell tumours could be identified (Arendt et al. 2015). Sequence capture of the different associated regions and the subsequent sequencing identified a SNP in GNAI2. In addition, several genes involved in hyaluronan synthesis have been implicated in the development of these tumours. Comparable sequencing approaches have been successfully applied in canine ophthalmologic disorders (Downs & Mellersh 2014), orthopaedic research (Hytonen et al. 2012; Lavrijsen et al. 2014a,b) and other conditions (Vaysse et al. 2011; Forman et al. 2012, 2013; Kim et al. 2012; Yokoyama et al. 2012; Tengvall et al. 2013; Agler et al. 2014; Massey et al. 2014).

Whole exome sequencing

The first published whole-exome experiment in dogs revealed the mutation for progressive renal atrophy in the Papillon and Phalène breeds (Ahonen et al. 2013). The study utilised a combined approach of genetic linkage analysis and exome sequencing to map the locus and to identify the causative variant in CNGB1. Another exome sequencing example suggested that variants in NEB could be causal for Basset Hounds suffering from primary angle closure glaucoma (Ahram et al. 2015). Exome sequencing in a trio of one affected individual and its healthy parents resulted in the identification of four non-synonymous variants within the identified chromosome 19 region of interest, one of which is postulated to be causative.

Exome designs have been based on the CanFam2.0 reference. The older reference versions have limited genome coverage and lack many coding, small RNA and regulatory regions that are included in the recent CanFam3.1-based designs (Broeckx et al. 2015). The new exome capture reagents provide powerful and cost-efficient resequencing tools for gene discovery in many Mendelian disorders that are likely to harbour coding mutations.

Whole-genome sequencing

The cost of whole-genome sequencing (WGS) continues to decrease and offers a comprehensive analysis of the entire DNA sequence. At the same time, WGS avoids technical enrichment biases that are present in capture reagents and methods in targeted analyses, including exome studies. The increasing amount of available WGS data provides powerful filtering strategies and facilitates mutation discovery in coding and non-coding regions. Several examples already exist. A breed-wide study investigating cranial development combined GWAS and WGS approaches followed by a functional validation in zebrafish (Schoenebeck et al. 2012). Genetic mapping in a cohort of 576 dogs from 72 breeds in four different categories of skull shape identified four different associated loci. Genome-wide sequencing was performed in 11 dog breeds and focused on the most promising, previously unexplored regions. The 190-kb region included two candidates, of which BMP3 revealed a missense mutation. The BMP3 morpholino-injected zebrafish showed the loss of jaw structures and frontal bossing as well as cartilage abnormalities.

Other examples of studies with locus-mapping strategies prior to WGS include the identification of ADAMTS20 as having a risk variant for cleft lip and palate (Wolf et al. 2015), the discovery of four risk loci involved in both B-cell lymphoma and hemangiosarcoma (Tonomura et al. 2015) and the categorisation of genomic alterations in canine transmissible venereal tumours by comparing these to 186 canine whole-genome sequences (Decker et al. 2015). In the case of the discovery of the frameshift mutations involved in Imerslund–Grasbeck syndrome in Beagles and Border Collies (Drogemuller et al. 2014), the genes CUBN and AMN were known from human cases. Together, these genes
represent 80 exons. Performing genome-wide sequencing of single cases was cheaper and faster than was classical DNA sequencing of all exons one by one. Therefore, genome-wide sequencing is also considered to be a powerful and cost- and time-efficient tool if strong candidate genes are known (Droegemueller et al. 2014; Guo et al. 2014; Willet et al. 2015).

One can easily foresee that eventually DNA sequence analysis will make SNP arrays redundant. Genotyping by sequencing may potentially provide a single method in the future for gene mapping and sequencing purposes (Cooke et al. 2016). The first example of the combination of mapping and mutation discovery by WGS was used to provide insight into signatures of selection that are associated with high altitude. The strongest Z-transformed $F_{ST}$ values when comparing breeds living at high, middle and low altitudes were found within a region comprising EPAS1, carrying a most likely causal mutation in its PAS domain (Liu et al. 2014). In order to obtain an insight into extensive genomic alterations in canine spontaneous mammary cancer, a series of mammary carcinomas (seven simple and four complex) have been sequenced at whole-genome, whole-exome and whole-transcriptome resolutions. Based on this elegant combination of three high-throughput screenings, it was postulated that simple carcinomas originate from genomic aberrations and complex carcinomas are caused by epigenetic changes. Combining multiple NGS strategies strongly enhances the comprehension of the underlying biology of disorders.

**Future perspectives**

**1000 Canomes (exomes and genomes)**

Ongoing disease-related whole-exome sequencing and WGS efforts are rapidly progressing in the canine genetics community. These efforts will yield sequences from hundreds of dogs across breeds and wolf populations to reveal millions of common and rare variants. It is important to further encourage researchers to share their data among groups to aid and empower individual studies. A corresponding successful collaborative example includes the 1000 bull genome project that facilitates monoergic and complex disease mapping in bovine (Daetwyler et al. 2014). The ultimate goal in the canine community should be the establishment of a public database that includes all genome sequences and variants together with the sufficient phenotypic information from the sequenced dogs to allow the efficient use of data. Fortunately, such an international effort involving 10,000 dog genomes has been recently initiated (http://www.dog10kgenomes.org/), and individual research groups have already started to share genome data on a smaller scale to support each other’s mapping efforts. The publicly available human 1000 genomes data (1000 Genomes Project Consortium et al. 2010) and more recent 2500 genomes data (Sudmant et al. 2015) as well as the Genome of the Netherlands Project (1000 Genomes Project Consortium et al. 2010; Genome of the Netherlands Consortium 2014) have been instrumental in characterising geographic, structural and functional spectrums of genetic variation that contributes to human diseases. ‘The 1000 Canomes’ would serve the same purpose in dogs, significantly expediting gene discovery and comparative studies across species.

**Functional annotation of the canine genome**

The sequence data itself may have limited use without a comprehensive functional annotation of the dog genome. Several improvements have been accomplished in the CanFam2.0 reference genome containing several sequencing errors and gaps (Hoepner et al. 2014). Combined BAC and RNA sequencing efforts from 10 different canine tissues resulted in the elimination of most of the gaps and annotation of novel protein-coding genes and additional isoforms per gene in the current CanFam3.1 reference. However, although the current reference is useful for many studies, it still suffers from gaps and incomplete and inaccurate annotation of many genes, particularly in the 5’ end and regulatory regions of the genes. The use of advanced long read sequencing technologies, such as the PacBio long sequencing read approach, would help to fill the remaining gaps, as recently demonstrated for a canine epilepsy locus (Koskinen et al. 2015). The current short read WGS technology is not optimal for discovering structural variants in genomes and could also benefit from long read WGS approaches. Although comparative genomics between canine and human or mouse genomes may help to specify candidate functional variation in the regulatory regions, more comprehensive annotation of the entire genome would aid genetic and epigenetic studies in dogs, particularly in complex diseases such as cancer and behavioural traits. Ongoing efforts are expected to provide the next generation of improvements for the reference, further facilitating the use of dogs as comparative models for many traits.

Over the past decade, canine genetic research has focused largely on coding region variations (Merveille et al. 2011; Grall et al. 2012). Whether this is a useful strategy depends on the type of the disorder being investigated. The success rate of focusing on only coding variants is higher in monogenic Mendelian disorders than in complex disorders. In complex disorders, causative variants have been reported frequently in introns, influencing exon skipping (Lee et al. 2012). Furthermore, intergenic variants have been found to regulate tissue-specific expression (Grundberg et al. 2012). The first examples of non-coding variants and their role in complex disorders in dogs have been published, confirming the power of the canine model (Karlsson et al. 2013; Tang et al. 2014). The fact that roughly 5% of the
human genome is evolutionarily conserved when compared with rodents implies that these regions are likely functionally important. Based on the finding that only one-third of these genomic regions code for proteins, it has been hypothesised that intergenic regions do not consist merely of so-called junk DNA, but do indeed harbour functional elements.

Based on this theory, The Encyclopedia of DNA Elements (ENCODE) Consortium had its kickoff in 2004 (ENCODE Project Consortium 2004) to discover regulatory elements. Numerous datasets covering many different cell types were used for the search of functional DNA elements, resulting in a series of fundamental findings (Pennisi 2012; Skipper et al. 2012; Stamatoyannopoulos 2012). The consortium managed to systematically map regions of transcription, transcription factor associations, chromatin structures and histone modifications on the human genome. These findings opened up a discussion as to which parts of the genome are truly functional. According to the consortium, approximately 80% of the genome was found to be functional. Rands et al. (2014) questioned these data and postulated that only 8.2% is subject to negative selection and therefore functional. The ENCODE data describe protein–DNA interactions and define transcribed bases as functional. Proteins in nucleosomes that pack DNA, transcription binding factors and other gene expression regulating elements are included in these data, and this is why the estimate of the ENCODE Consortium for the fraction of functional DNA reached 80%.

Another major effort to identify functional areas of the genome was the Functional Annotation of the Mammalian Genome 5 (FANTOM5) Project, which provided comprehensive expression profiles, transcription start sites, enhancer maps and functional annotation of mammalian cell-type-specific transcriptomes with wide applications in biomedical research (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al. 2014). Similar ENCODE and FANTOM approaches to functional annotation of the canine genome would significantly facilitate the identification of causal variants across the genome to fully utilise the dog genome would significantly facilitate the identification of causal variants across the genome to fully utilise the dog genome. Extreme morphological and behavioural variation as well as the high number of complex diseases in dogs suggests a strong involvement of regulatory variation contributing to traits. Solving complex disorders could benefit greatly from such a CANCODE approach.

References


Baranowska Korberg I., Sundstrom E., Meadows J.R. et al. (2014) Improved canine exome designs, featuring ncRNAs and increased coverage of protein coding genes. Scientific Reports 5, 12810.


Downs L.M. & Mellersh C.S. (2014) An intronic SINE insertion in FAM161A that causes exon-skipping is associated with


Sutter N.B., Bustamante C.D., Chase K et al. (2007) A single IFI1 allele is a major determinant of small size in dogs. Science 316, 112–5.


