Bioinformatics Mining for Disease Causing Mutations

Using the Dog Genome as a Model for Human Disease

Katarina Truvé
Faculty of Veterinary Medicine and Animal Sciences
Department of Animal Breeding and Genetics
Uppsala

Doctoral Thesis
Swedish University of Agricultural Sciences
Uppsala 2012

Acta Universitatis agriculturae Sueciae
2012:64

Cover: “Man’s best friend in sickness and in health”
(photo: Staffan Truvé)

ISSN 1652-6880
ISBN 978-91-576-7711-2

Print: SLU Service/Repro, Uppsala 2012

Bioinformatics Mining for Disease Causing Mutations using the Dog Genome as a Model for Human Disease

Abstract

Humans and dogs share many common diseases, and it has been shown that the identification of mutations that cause disease in dogs can help unravel the genetic basis for a similar disease in humans. Mapping of traits and disease in dogs is not a new idea, but the sequencing of the whole dog genome, the creation of a dense SNP maps followed by the development of SNP arrays for high throughput genotyping has led to new facilitated mapping procedures. Each dog breed can be seen as a genetic isolate and certain breeds are often predisposed to specific diseases. Because of the genomic structure of the dog genome and the availability of new resources for disease mapping, the dog has been proposed to be especially advantageous for the mapping of complex disease that is difficult to map in human outbred populations.

In this thesis, the aim has been to identify disease-causing mutations for three complex diseases in dogs with the presence of similar conditions in humans. Emphasis has been on bioinformatics analyses of genome-wide SNP and large re-sequencing data.

In the dog breed Nova Scotia duck tolling retriever it is common with an immune-mediated disease complex that resembles human systemic lupus erythematosus (SLE). In paper I we used a two-stage genome-wide association mapping method and successfully located several susceptibility loci in dogs for this disease complex. In paper II we identified a mutation that had been under selection in the Shar-Pei breed, causing both a breed-defining wrinkled skin phenotype and an autoinflammatory fever disease. Because the locus had been under selection we used an alternative mapping approach, called homozygosity mapping to identify the locus, followed by re-sequencing using next generation sequencing technologies. In paper III we report the development of a web-based tool that facilitates analyses and extraction of essential information from the large amount of data produced by next generation sequencing projects. In paper IV we used across-breed genome-wide association mapping to identify risk factors for glioma, a type of malignant brain tumor fatal to both human and dogs. For the three diseases excellent candidate genes have been identified, and continued research might has the potential to lead to better treatment options and thus benefit both dogs and humans.

Keywords: dog, genome-wide association mapping, homozygosity mapping, next generation sequencing, glioma, autoinflammatory disease, systemic lupus erythematosus (SLE)

Author's address: Katarina Truvé, SLU, Department of Animal breeding and Genetics
P.O. Box 7023, 750 07 Uppsala Sweden
E-mail: Katarina.Truve@slu.se

Dedication

To everyone with an interest in improved health and better treatment options for humans and their best friends

”A healthy person has many wishes, but the sick person has only one” – Indian proverb

Contents

List of Publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I Wilbe M, Jokinen P*, **Truvé K***, Seppala EH, Karlsson EK, Biagi T, Hughes A, Bannasch D, Andersson G, Hansson-Hamlin H, Lohi H, Lindblad-Toh K. (2010) Genome-wide association mapping identifies multiple loci for a canine SLE-related disease complex. *Nature Genetics*. 42(3):250-254.

II Olsson M, Meadows JR*, **Truvé K***, Rosengren Pielberg G*, Puppo F*, Mauceli E, Quilez J, Tonomura N, Zanna G, Docampo MJ, Bassols A, Avery AC, Karlsson EK, Thomas A, Kastner DL, Bongcam-Rudloff E, Webster MT, Sanchez A, Hedhammar A, Remmers EF, Andersson L, Ferrer L, Tintle L, Lindblad-Toh K. (2011) A novel unstable duplication upstream of HAS2 predisposes to a breed-defining skin phenotype and a periodic fever syndrome in Chinese Shar-Pei dogs. *PLoS Genetics*. 7(3):e1001332.

III **Truvé K**, Eriksson O, Norling M, Wilbe M, Mauceli E, Lindblad-Toh K, Bongcam-Rudloff E. (2011) SEQscoring: a tool to facilitate the interpretation of data generated with next generation sequencing technologies. *EMBnet journal,* 17.1 38-45.

IV **Truvé K,** Dickinson P, York D, Rosengren Pielberg G, Perloski M, Murén E, Fuxelius HH, Andersson G, Hedhammar Å, Bongcam-Rudloff E, Lindblad-Toh K, Bannasch D. (2012) Identification of a glioma susceptibility locus in the wake of selective dog breeding for brachycephaly. Manuscript.

Papers I-III are reproduced with the permission of the publishers.

*These authors contributed equally

Abbreviations

Amstaff	American Staffordshire terrier
ANA	antinuclear autoantibodies
Bp	basepair
CCDC39	coiled-coil domain containing 39
CFA	c*anis familiaris* chromosome
CMH	Cochran-Mantel-Haenszel
DAPP1	dual adaptor of phosphotyrosine and 3-phosphoinositides
DNA	deoxyribonucleic acid
FMF	Familial Mediterranean Fever
FSF	Familial Shar-Pei Fever
GC	genomic control
GWAM	genome-wide association mapping
GWAS	genome-wide association study
HA	hyaluronic acid
HAS1	HA synthase 1
HAS2	HA synthase 2
HAS2as	HAS2 antisense gene
HAS3	HA synthase 3
HOMER2	homer homolog 2
IBD	identical-by-descent
IMRD	immune-mediated rheumatic disease
kb	kilobases
LD	linkage disequilibrium
LINE	long interspersed nucleotide element
log2	binary logarithm
MAF	minor allele frequency
Mb	megabases
MURR 1	copper metabolism gene
NF-AT	nuclear factor of activated T cells
NGS	next generation sequencing
NSDTR	Nova Scotia duck tolling retriever
PCD	primary ciliary dyskinesia
PPP3CA	protein phosphatase 3, catalytic subunit, alpha isoform
PTPN3	protein tyrosine phosphatase, non-receptor type 3
RNA	ribonucleic acid
SAA	serum amyloid A protein
SINE	short interspersed nucleotide element

SLE	systemic lupus erythematosus
SMOC2	SPARC related modular calcium binding 2
SNP	single nucleotide polymorphism
SRMA	steroid-responsive meningitis-arteritis
UCSC	University of California, Santa Cruz

1 Introduction

1.1 Using the dog as a model for human disease

The domestic dog (*Canis lupus familiaris)* is an excellent model species for the study of disease genetics. Humans share many common diseases with their canine friends, *e.g.* cancer, autoimmune diseases, epilepsy and heart disease. Disease manifestations are often similar in dogs and humans and most of our genes are orthologous (Karlsson & Lindblad-Toh, 2008). There are more than 400 dog breeds (Wilcox & Walkowicz, 1995), all with differing behavioral and morphological characteristics. Most dog breeds were created during the past two centuries by strong artificial selection leading to relatively inbred populations with sometimes unintended consequences concerning the health (Lindblad-Toh *et al.*, 2005). There are often high incidences of specific diseases in certain breeds, explained by the random amplification of risk alleles during population bottlenecks or accidental enrichment because of hitchhiking of mutations near selected traits or pleiotropic effects of selected variants (Karlsson & Lindblad-Toh, 2008; Patterson *et al.*, 1988).

Complex diseases are more difficult to map than monogenic classical Mendelian recessive or dominant traits. In general for a complex disease the phenotype is caused by the interaction of several genes, the environment and stochastic factors. (Lander & Schork, 1994). The recent breed-creation and low genomic diversity within dog breeds implicates that increased risk is attributable to only a few disease alleles of strong effect, making disease mapping potentially easier in dogs compared to human, especially for complex disease (Ostrander & Kruglyak, 2000).

The diseases that were chosen for study in this thesis concordantly have similarities with human disorders, all of them showing a complex pattern. Firstly, in the breed Nova Scotia duck tolling retriever it is common with immune-mediated diseases including a disease-complex that resembles human

systemic lupus erythematosus (SLE). Secondly, many dogs from the Shar-Pei breed suffer from a hereditary periodic fever syndrome that has similarities with several human auto-inflammatory syndromes. Thirdly, Brachycephalic (short-nosed) dog breeds such as Boxer, Bulldog and Boston terrier have an increased incidence of glioma, a type of brain tumors that are devastating to both dogs and humans.

Identification of causative loci in dogs has the potential to increase knowledge about genes and pathways relevant to human disease, which might has the potential to lead to better diagnostics and improved treatment options for both species.

1.2 Shaping the genome structure of the domestic dog

Historical events have shaped the genome of the pure bred dog in a way that makes it excellent for genetic disease mapping. The genome structure of an individual breed bears evidence of two widely spaced major population bottlenecks. The first bottleneck occurred at domestication and the second at breed-creation with subsequent inbreeding and the enrichment of inherited breed-specific diseases (Parker *et al.*, 2009; Drogemuller *et al.*, 2008; Lindblad-Toh *et al.*, 2005; Ostrander & Wayne, 2005).

There is evidence that the dog is derived from gray wolves only and no other wild canids (Lindblad-Toh *et al.*, 2005; Savolainen *et al.*, 2002; Vila *et al.*, 1997). It is believed that domestication began more than 15,000 years ago, since there is ample archeological evidence of domesticated dogs from that time. Selection for desirable traits *e.g*. ability to hunt, guard, and herd, and for morphological traits like size and shape likely have prehistoric roots (Larson *et al.*, 2012). There are several examples of geographically distributed dog populations sharing identical mutations for phenotypes causing for instance hairlessness in Chinese and Mexican breeds (Drogemuller *et al.*, 2008) and the ridge on the back of sub-Saharan African and Thai breeds (Salmon Hillbertz *et al.*, 2007). At least 19 breeds share the same mutation causing short legs (Parker *et al.*, 2009). These mutations are not likely to have arisen multiple times but imply that there has been a significant mixture of genetic material before the gene-pools were closed during the creation of the currently existing breeds (Larson *et al.*, 2012).

A few founder dogs have typically been used in the creation of each breed. Usually a breed standard has been agreed upon that further reduces diversity, encouraging breeding of individuals that are fairly similar in type. This breeding procedure has resulted in the loss of genetic diversity within a breed and a greater variation across breeds (Ostrander & Wayne, 2005).

The two major bottlenecks, domestication and breed-creation, have resulted in long haplotype blocks, *i.e*. stretches with no recombination, within modern dog breeds (500 kb to 1Mb) and thus extensive linkage disequilibrium (LD), while short LD and short haplotype blocks (≈10 kb) are revealed as remnants from the ancestral dog population by comparison across breeds (Lindblad-Toh

et al., 2005). It should be noted though that there are considerable differences in extent of LD between breeds due to differences in breed popularity and local population bottlenecks. Compared to humans, LD in dogs is 20-50 times more extensive (Lindblad-Toh *et al.*, 2005; Ostrander & Wayne, 2005; Sutter & Ostrander, 2004).

Figure 1 illustrates how diversity is reduced during breed-creation. Long haplotype blocks are the results of the small number of meiosis that have taken place since breed creation. Note that since some shorter haplotype blocks residing within the longer blocks are shared by all haplotypes within a breed, they are in fact not visible, as long as they are not compared to breeds carrying other variants.

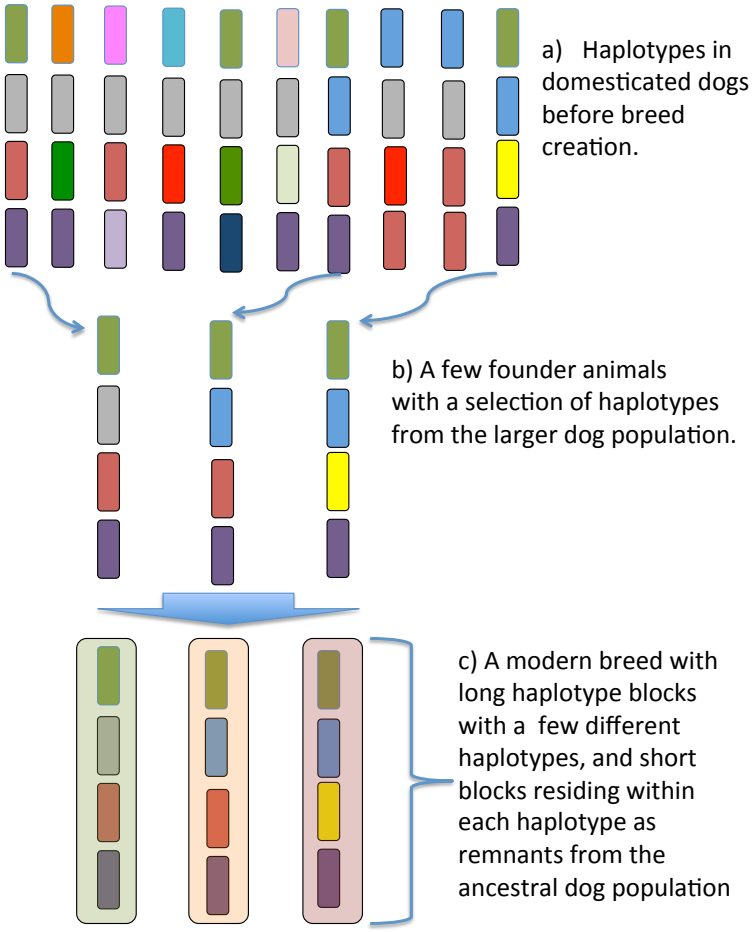


Figure 1. Reduction of diversity during breed-creation.
a) Small colored blocks represent ancestral haplotype blocks combined in diverse ways in the dog population before breed-creation. b) At breed-creation, the diversity of haplotypes is reduced, and only a portion of the variation in the common dog gene pool is brought to the created isolated population. c) Colored boxes surrounding the ancient shorter blocks illustrate longer haplotype blocks in the modern breed where no recombination has yet taken place.

1.3 Resources that facilitate trait mapping in dogs

Trait mapping in dogs has been facilitated by recent development of several new resources. A high-quality draft sequence of a female Boxer dog was published in 2005. The genome was covered ≈ 7.5 times by redundant sequence data, and the assembly included ≈99% of the euchromatic genome (Lindblad-Toh *et al.*, 2005). The dog genome contains ≈ 2.4 x 10^9 base pairs and is divided into 38 autosomal chromosomes and the X and Y sex chromosomes. To be able to extract the full potential of the dog genome a dense SNP map is needed that can be used to explore the genetic variation within and among dog breeds. Hence the same authors produced a catalogue of > 2.5 million SNPs in three complementary ways; (1) by identifying heterozygous SNPs in the sequenced Boxer, (2) comparing the genome of the Boxer with the partial sequence of a standard Poodle (Kirkness *et al.*, 2003) and (3) the generation of random reads from nine additional breeds, four wolves and one coyote (Lindblad-Toh *et al.*, 2005).

Several SNP genotyping arrays have been developed to facilitate high-throughput mapping. These are the 26578 (27K) and the 49663 (50K) Affymetrix SNP arrays, and a 22,362 SNP Illumina array (Karlsson & Lindblad-Toh, 2008). The latest contribution is the 173,622 canine HD Illumina SNP array with a mean spacing of 13 kb (Vaysse *et al.*, 2011).

In addition to knowledge about the dog genome, another very important aspect is that the dog has been intensely studied in medical practice. Furthermore, detailed family history and pathology data are often available (Hedhammar *et al.*, 2011; Patterson, 2000).

1.4 Genetic mapping of traits and diseases

Genetic mapping has the goal to find allelic variants that are causative of certain traits or that increase the risk of certain diseases. The underlying assumption is thus that the mutation occurred in a common ancestor to the individuals that share the phenotype in question, and that the mutation has been segregating through the generations. Alleles that are inherited from a common ancestor are said to be identical-by-descent (IBD). Figure 2 illustrates how a disease-causing mutation for a recessive disease is transmitted in a hypothetical pedigree, and is more likely to occur in two copies because of inbreeding.

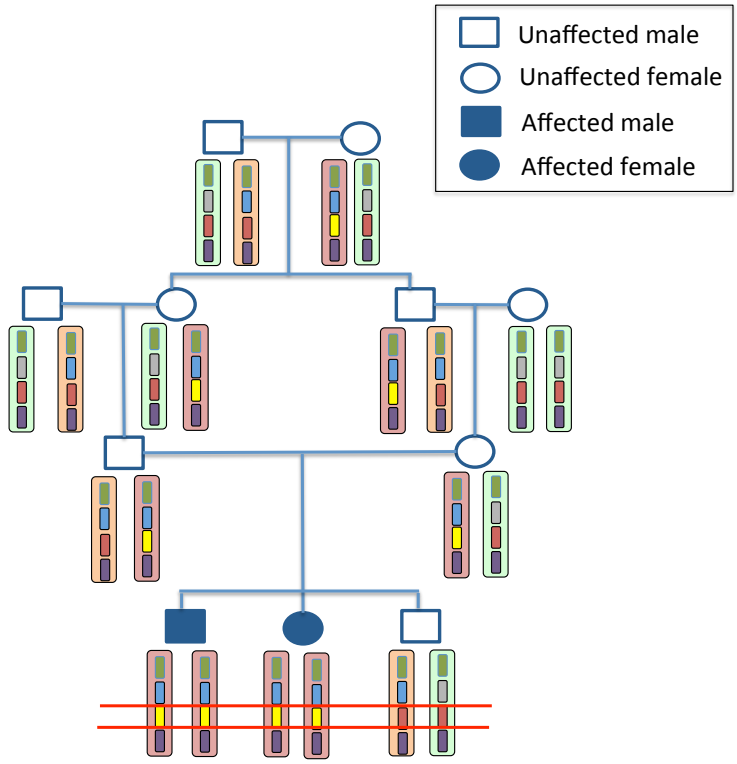


Figure 2. Pedigree with an inherited recessive mutation.
The figure illustrates how a mutation is located within the "short yellow haplotype", and how the mutation causes a change in phenotype when present in two copies.

To find a shared mutation that is IBD is a fundamental goal for disease mapping and applies also to dominant and complex modes of inheritance. Mapping can however be less straightforward than the illustration suggests, for reasons like incomplete penetrance, environmental and stochastic factors, and involvement of several genes such as different modifiers in the genetic background that complicates the picture.

1.4.1 Linkage mapping

A type of genetic mapping approach that was not used in this thesis but is worth mentioning is linkage mapping. Linkage mapping has frequently been used in the past and has clearly been the method of choice for simple

Mendelian traits. The method is based on proposing a model for inheritance which requires a pedigree and a test for correlated transmission of trait/disease and allele within a pedigree thus using individuals with known relatedness (Lander & Schork, 1994). Linkage studies do not need a control group and do not suffer from problems with heterogeneity and population stratification, as can be the case for genome-wide association studies (GWAS). A drawback is that identified regions are often large (>5 Mb), and test statistics used are complicated (Ott *et al.*, 2011).

Trait mapping in dogs began in the 1990s using large multigenerational pedigrees and many trait loci have been identified. For reviews see (Karlsson & Lindblad-Toh, 2008; Sutter & Ostrander, 2004).

In some cases the knowledge gained in dogs has opened doors for a wider understanding of similar disease conditions in humans. Copper toxicosis in Bedlington terrier is one such example. It has been shown that a mutation in the *MURR1* gene impairs the biliary excretion of copper in Bedlington terriers (van De Sluis *et al.*, 2002). Wilson disease is a similar condition in humans resulting in toxic copper accumulation. An investigation of the *MURR1* gene in humans with Wilson disease suggested that variants in this gene were associated with an early onset of disease (Stuehler *et al.*, 2004). Another example is narcolepsy, for which an autosomal recessive mutation in the hypocretin 2 receptor gene was identified in a canine model. The results revealed a family of genes that were shown to encode major sleep-modulating neurotransmitters, enabling new treatments for narcoleptic human patients (Lin *et al.*, 1999).

1.4.2 Genome-wide association mapping

Over the past few years, linkage mapping has lost its predominance in favor of genome-wide association mapping (GWAM). It is easier to collect case-control samples than family collections and the method makes it possible to map not only Mendelian traits, but also common risk factors for complex diseases. GWAM is based on tests of several markers for being in linkage disequilibrium with a trait or disease. True association and linkage disequilibrium (LD) arises either when a marker allele actually is the cause of disease, or if the marker allele is located close enough to the causative mutation. The mutation is considered to be segregating from a common ancestor and show correlation with a marker if linkage has not yet been eroded by recombination (Lander & Schork, 1994). In practice, allele frequencies for affected and unaffected individuals from a population are compared using several markers and if one allele is significantly more frequent in cases than controls, then that allele is consequently associated with the trait.

LD is dependent on population history and it may be advantageous to use isolated populations for mapping where LD is extensive, which require fewer markers (Lander & Schork, 1994). Isolated populations are also more homogenous, which can be of advantage especially for complex traits. Genetic

heterogeneity implies that a risk haplotype co-segregates with a disease in some families but not others (Lander & Schork, 1994). In outbred human populations, GWAS have difficulties in detecting rare variants due to heterogeneity (Ott *et al.*, 2011). Another potential pitfall of association studies is the presence of stratification in the population, *i.e*. subgroups with differing allele frequencies that might cause false positive association. Cases and controls should therefore, if possible, be matched to be as similar as possible except for the trait under investigation.

Dog breeds have all the advantages of isolated populations for GWAM, as each breed can be seen as a genetic isolate. In the case where traits or disease are shared between several breeds the genetic structure of the dog genome makes it possible to efficiently perform the mapping using a *two-stage strategy*. It is likely that most disease-causing mutations arose prior to breed-creation since there has been little time for novel mutations to accumulate since breed-creation (Karlsson & Lindblad-Toh, 2008). The *first stage* is performed within a single breed where it is sufficient to use a sparse set of markers (≈15,000) to identify an associated region of ≈1Mb. Simulations have shown that a recessive trait can be mapped with as few as 20 cases and 20 control (Lindblad-Toh *et al.*, 2005) while a risk factor that multiplies risk with a factor 5 for a complex disease has been simulated to be detected in 97% of data sets, using 100 cases and 100 controls and a set of 15,000 markers (Lindblad-Toh *et al.*, 2005). In the *second fine-mapping stage* a denser set of SNPs and more breeds are included with the potential to narrow the region(s) to a few hundred kilobases (Karlsson & Lindblad-Toh, 2008). Proof of principle, has been shown by the mapping of two monogenic traits: white spotting in Boxer and other breeds as well as the hair ridge in Rhodesian ridgebacks using the 27,000 SNP array and only 10 cases and 10 controls (Karlsson *et al.*, 2007; Salmon Hillbertz *et al.*, 2007).

In certain situations extensive inbreeding can have negative consequences in the form of large homozygous regions that are essentially invisible to association mapping, since this method relies on allelic segregation of surrounding markers (Karlsson & Lindblad-Toh, 2008). When a trait is fixed within a breed and shared by several breeds an alternative approach is to perform across-breed GWAS. In these studies some individuals from several different breeds with or without the trait in question are compared. Allele frequencies differ extensively between breeds, making the risk of false positives higher than for studies performed within a single breed. With many different breeds included in the study, the chance is higher that only trait-related alleles will differ consistently (Karlsson & Lindblad-Toh, 2008). Across-breed mapping is most likely to be successful for mapping traits that have been under selection. In general, LD decays much faster between breeds, but for traits that have been under selection, the causative variant is likely located in a fixed long haplotype. Thus, LD is likely increased in the region under selection, given that affected breeds share a common founder ancestor for the trait in question (Vaysse *et al.*, 2011).

1.4.3 Homozygosity mapping

Homozygosity mapping is a method that has been applied to find an allele causing a rare disease with a recessive mode of inheritance. The region surrounding the disease allele is thus homozygous in affected individuals and the method relies on detecting long stretches of homozygosity (Ott *et al.*, 2011). Since regions need to be long to be identified, this method is best suited for recent mutations that occur in families or in isolated populations. In the case of a recent mutation, homozygosity mapping is an option also in dogs. A success story is exemplified by the use of only five cases and 15 controls of the dog breed Old English Sheepdog that identified a mutation in the *CCDC39* gene as being responsible for a chronic airway inflammation with similarities to the human disease primary ciliary dyskinesia (PCD). Loss of function in the orthologous human gene was thereafter found in a substantial fraction of PCD patients (Merveille *et al.*, 2011; Ott *et al.*, 2011). Strong selection for a desirable trait can lead to homozygous (fixed) regions in all dogs within a breed. Such regions could also be possible to find with homozygosity mapping by screening the genome for large regions of homozygosity. The thick heavily wrinkled skin of Shar-Pei dogs is a selected trait that is unique to that breed. Sometimes a specific fixed trait can be shared among several breeds exemplified by chondrodysplasia (short legs) or brachycephaly (short nose). In the case of a shared mutation, it is possible to search for shared regions of homozygosity across such breeds. In those cases it could also be possible to perform across-breed association mapping as explained above.

1.5 Next generation sequencing

In the 1970s it became possible to clone and sequence deoxyribonucleic acid (DNA), and thereby tie genetic linkage to the underlying DNA sequence (Altshuler *et al.*, 2008). For about three decades the primary choice for DNA sequencing was methods requiring electrophoretic separations of DNA fragments. It was the development of high-resolution methods that could separate DNA fragments differing in size by just one base that made sequencing possible in the first place. For reviews see (Shendure & Ji, 2008; Shendure *et al.*, 2004). Automation, parallelization, and refinements of Sanger sequencing was the road to increased cost-effectiveness (Shendure *et al.*, 2004). The goal to sequence the entire human genome was set already in 1985 and motivated a cost reduction from US$ 10 per finished base to 10 bases per US$ 1 (Shendure *et al.*, 2004). In 2001, two draft sequences of the human genome were published (Lander *et al.*, 2001; Venter *et al.*, 2001). In the wake of the human genome project several academic and commercial efforts were

initiated with the aim to develop new ultra low cost sequencing technologies (Shendure *et al.*, 2004).

Over the past seven years these new technologies have been evolving rapidly, and massively parallel DNA sequencing platforms are now widely available. These technologies produce relatively short reads compared to Sanger sequencing. The utility of short reads became stronger and more valuable with the availability of whole genome assemblies for human and other species. Those assemblies provided reference genome sequences from various different species against which short reads could be mapped, thereby providing information about genetic variation (Shendure & Ji, 2008). The term "second-generation" sequencing was proposed by (Shendure & Ji, 2008) used for implementations of parallelized cyclic-array sequencing (*e.g.* as in the commercial products: 454 Genome Sequencer, Illumina Genome Analyzer and the SOLiD platform). Electrophoretic separation is no longer needed, but sequencing is performed by synthesis in iterative cycles of enzymatic manipulation and image base data collection. Millions of PCR colonies of DNA fragments are immobilized to an array and are simultaneously sequenced in parallel. Now "third generation" technologies are emerging that promise even higher throughput, longer read lengths, smaller amounts of starting materials, higher consensus accuracy and even lower costs. For a review see (Schadt *et al.*, 2010).

Next generation sequencing (NGS) can be applied for a variety of reasons and it has the potential to dramatically accelerate biological and biomedical research (Schadt *et al.*, 2010; Shendure & Ji, 2008).

Genome-wide mapping studies in dogs where the mutation is shared by several breeds could identify candidate regions less than 100 kb, but candidate regions might be up to 2 Mb long for mutations that are specific to a single breed as discussed above (Karlsson & Lindblad-Toh, 2008). Re-sequencing of the entire region in several cases and controls would have been too costly prior to development of NGS technologies, but now provide the potential to identify the causative mutation. In this thesis the Illumina NGS technology was used for targeted discovery of sequence variation in genomic regions associated to specific traits or disease (Paper II and IV). In paper III we assessed the challenge of extracting the most essential information from the very large amount of data that is produced by next generation sequencing.

2 Aims of the thesis

The overall aim of this thesis has been to mine the dog genome for genetic risk factors underlying canine genetic diseases, using bioinformatics methods.

The specific aims were to:

I. Perform a genome-wide association study in the dog breed Nova Scotia duck tolling retriever to identify susceptibility loci for an immune-mediated disease complex with similarities to human systemic lupus erythematosus (SLE).

II. Map the locus for the characteristic wrinkled skin phenotype of the Shar-Pei breed, and to map the breed-specific Familial Shar-Pei Fever (FSF), a disease with similarities to several human autoinflammatory syndromes.

III. Develop an easy to use web-based tool for analyses of data from NGS projects, in order to facilitate the identification of causative mutations in targeted case-control re-sequencing projects.

IV. Identify the loci predisposing to an increased risk of gliomas (primary brain tumors) in brachycephalic (short-nosed) dog breeds, and place it in context of pathways for gliomagenesis relevant for human disease.

3 Genome-wide association mapping identifies several loci for an SLE-related disease in Nova Scotia duck tolling retrievers (Paper I)

3.1 Background

The incidence of autoimmune disease is overrepresented in the breed Nova Scotia duck tolling retriever (NSDTR). In particular, two types of immune-mediated phenotypes are diagnosed more frequently in NSDTR compared with other breeds. These are immune-mediated rheumatic disease (IMRD) and steroid-responsive meningitis-arteritis (SRMA) (Hansson-Hamlin & Lilliehook, 2009; Anfinsen *et al.*, 2008; Redman, 2002). A primary question in this project was to determine whether these represent two separate disorders, or if they share common genetic risk factors. The IMRD disease complex shares many similar clinical features with human systemic lupus erythematosus (SLE). IMRD-affected dogs frequently display antinuclear autoantibodies (ANA) (70%) and arthritis (100%). Other symptoms sometimes seen in both dogs and humans include fever and affection of skin, liver and kidneys (Hansson-Hamlin & Lilliehook, 2009; Koskenmies *et al.*, 2008; Tan *et al.*, 1982). Dogs diagnosed with SRMA usually have a typical acute course of disease with severe neck pain, stiffness and fever. In most cases, treatment with corticosteroids gives a good response (Anfinsen *et al.*, 2008).

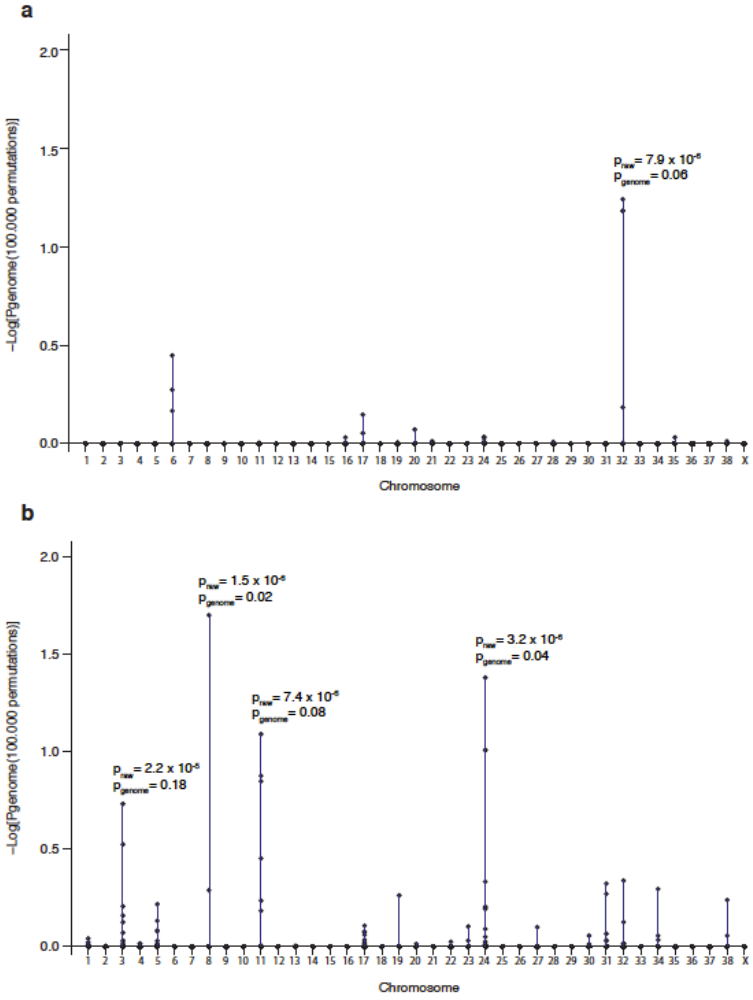
Most autoimmune diseases in humans are the result of a complex inheritance involving several genes, stochastic and environmental factors (Mariani, 2004). Similarly, pedigree analysis has indicated that the SLE-related diseases seen in NSDTRs involve polygenic inheritance. The NSDTR breed

has gone through a severe bottleneck, when a very small population survived two devastating outbreaks of canine distemper virus in 1908 and 1912 (Strang & MacMillan, 1996).

In this paper we performed a Genome-Wide Association Study (GWAS) in NSDTR to locate candidate loci for this immune-mediated disease complex. Simulations have shown that for a complex trait, risk alleles that increase risk with a factor of 5 can be detected (in 97% of cases) using only ≈ 15,000 SNPs and ≈100 cases and ≈100 controls (Lindblad-Toh *et al.*, 2005). Given the recent population bottleneck and the high prevalence of disease in this breed, we expected to find a few strong risk factors.

3.2 Methods and Results

To perform genome-wide association mapping we used the Illumina canine SNP array with 22,000 markers. We used 81 cases (37 diagnosed with IMRD whereof 22 were ANA-positive and 44 diagnosed with SRMA) and 57 controls to identify five candidate loci located on CFA 3, 8, 11, 24 and 32. Analyses were performed for all cases together and for ANA-positive IMRD and SRMA cases separately (Figure 3).


a
−Log[Pgenome(100,000 permutations)]
2.0
1.5
1.0
0.5
0.0
p_{new} = 7.9 x 10^{-6}
p_{genome} = 0.06
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 X
Chromosome
b
−Log[Pgenome(100,000 permutations)]
2.0
1.5
1.0
0.5
0.0
p_{new} = 1.5 x 10^{-6}
p_{genome} = 0.02
p_{new} = 3.2 x 10^{-6}
p_{genome} = 0.04
p_{new} = 7.4 x 10^{-6}
p_{genome} = 0.08
p_{new} = 2.2 x 10^{-6}
p_{genome} = 0.18
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 X
Chromosome

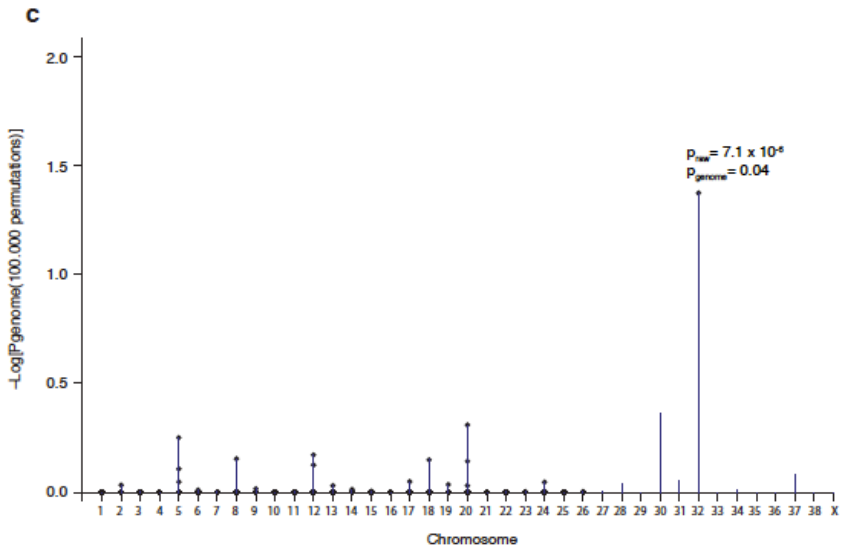


Figure 3. (a) All cases analyzed together showed one strongly associated peak on CFA 32. (b) Associated regions were located on four chromosomes for ANA-positive dogs (CFA 3, 8, 11 and 24). c) One region on CFA 32 was significantly associated with SRMA. (Figure modified from paper I)

The software tool PLINK (Purcell *et al.*, 2007) was used to perform all statistical tests including association calculations and to adjust for population stratification and multiple testing. The dogs used in the analyses were from Finland and Sweden and multidimensional scaling plots showed the presence of some population stratification. Identity-by-state (IBS) clustering was therefore used to separate the samples into two groups. The Cochran-Mantel-Haenszel (CMH) association statistics was then used to test for SNP-disease association conditional on the clustering. This method was used to avoid false positives due to subpopulations with differing allele frequencies, as well as to increase the power to detect true positives. The initial p-values were in the range 10^{-5}-10^{-6}. In a second step, we used permutation testing (*i.e.* label-swapping of phenotypes) conditional on the clustering into two groups, thus both controlling for stratification and correcting for multiple testing. Corrected p-values (P_{genome}) based on 100,000 permutations, reached genome-wide significance (cut-off 0.05) for three loci, located on chromosome 8 (P_{genome} ≈0.02) and chromosome 24 (P_{genome} ≈ 0.04) for ANA-positive dogs and on chromosome 32 (P_{genome} ≈0.04) for dogs diagnosed with SRMA (Figure 3).

All five loci were included in the proceeding fine-mapping. According to recommended state of the art, the whole genome mapping within one breed should be followed by fine-mapping in multiple breeds in order to identify

shorter across-breed shared haplotypes (Karlsson & Lindblad-Toh, 2008). Unfortunately, since these SLE-like diseases are rare in other breeds, we did not have sufficient number of samples to be able to narrow down the regions using this method. Nevertheless, fine-mapping also has the purpose to validate associated regions by adding additional cases and controls. SNPs were selected with a density of ≈ 1 SNP/10 kb in associated regions. Additional NSDTRs were included adding up to a total of 324 dogs (82 with IMRD, of which 32 showed ANA-positivity, 78 with SRMA and 173 controls; nine dogs were classified as having both IMRD and SRMA). Three loci (CFA 3, 11, 24) associated with ANA-positivity were strongly validated, with p-values in the range 10^{-11}-10^{-13}. The other two loci (CFA 8 and 32) were validated, but with weaker p-values of 10^{-5} and 10^{-8} in the fine-mapping stage. In the fine-mapping stage, a region of 1.6 Mb on CFA 32 was found to be the one most associated with SRMA, but also associated with ANA-positivity and consequently with all cases together.

3.3 Discussion and future prospects

In conclusion, five loci were identified as associated with the SLE-related disease complex in NSDTRs using a two-stage mapping strategy. The associated regions contain several genes making it difficult to determine which are causative even though several of the genes are excellent candidates based on their biological function. Strikingly, four of the candidate genes (*PPP3CA, HOMER2, DAPP1,* and *PTPN3*) from three of the associated regions are all involved in regulation of the nuclear factor of activated T cells (NF-AT) pathway. It is well established that the NF-AT transcription factors are involved in T-cell activation as well as the generation of peripheral tolerance against self-antigens (Serfling *et al.*, 2006). Activation of calcineurin (encoded by *PPP3CA)*, a well-known target for immunosuppressive drugs, results in the activation of the NF-AT pathway (Clipstone & Crabtree, 1992). Both calcineurin and NF-AT, have been reported to be differentially expressed in human patients with SLE compared to healthy individuals (Kyttaris *et al.*, 2007; Guerini, 1997). For more details about the candidate genes in associated regions, see paper I.

As expected, it seems like a few loci, each with a strong effect to influence development of this immune-mediated disease complex are present in this dog breed, and that a combination of these genetic risk factors might be sufficient to predispose to disease. Since strong risk factors that are rare escape GWAS in humans, we suggest that studies in dog can be a valuable complement in identifying genes and pathways that are involved in complex disease development and thus benefit both species. Hybrid capture followed by next generation sequencing was planned as the next step, to investigate associated regions in more detail and to be able to identify causative variants.

4 Homozygosity mapping identifies a locus under selection for the characteristic skin phenotype in Shar-Pei dogs (Paper II)

4.1 Background

The dog breed Shar-Pei has a breed-defining wrinkled skin phenotype that has been strongly selected for. The skin is thickened and folded and the muzzle is heavily padded. An ancestral type of Shar-Pei with less accentuated skin condition exists and is referred to as the "traditional" type Shar-Pei, while the type that has been selected for the skin phenotype is called the "meatmouth" type. It is known that the major component of the thickened skin in Shar-Pei dogs is hyaluronic acid (HA) and that meatmouth Shar-Peis also have two- to five-fold higher serum-levels of HA (Zanna *et al.*, 2008). A similar condition has been described in humans (Ramsden *et al.*, 2000) and termed hyaluronanosis.

Apart from the skin condition, an autoinflammatory disease named Familial Shar-Pei Fever (FSF) is very common among meatmouth Shar-Peis. In 1992, as many as 23% of US individuals within the breed were estimated to be affected (Rivas *et al.*, 1992). The disease clinically resembles some hereditary periodic fever syndromes seen in humans, such as Familial Mediterranean Fever (FMF). The condition in both human and dog is characterized by recurrent fever of unknown origin and local inflammation, usually affecting major joints. As a complication of recurrent or chronic inflammation, human patients as well as Shar-Pei dogs are at risk of developing reactive systemic AA amyloidosis and subsequent kidney or liver failure (Stojanov & Kastner, 2005; Rivas *et al.*, 1992). (AA amyloidosis is a form of amyloidosis associated with serum amyloid A protein (SAA) (Lachmann *et al.*, 2007).)

In an effort to identify the genetic risk factor/s for this disease we first performed a GWAS using 18 cases and 18 controls, but without obtaining any significant results. (Data not shown in paper.) Dog breeds offer the power of genetically isolated populations for association mapping, but in cases when homozygosity is extensive, large genomic regions will escape detection in association mapping. Hence we assumed that the selection for the skin phenotype, might have concealed the region responsible for the fever, and that finding the locus for the skin phenotype might also lead us to the mutation causing the fever syndrome. It was also possible that the study was underpowered, and in parallel more samples were collected.

4.2 Methods and results

We screened the genome of the Shar-Pei for signs of selective sweeps, by searching for long stretches of homozygosity or a reduction in heterozygosity. We investigated the Shar-Pei genome separately by searching for shared homozygosity in 50 Shar-Pei dogs, and also compared it to 230 control dogs from 24 other breeds. We used a sliding window approach to investigate the ratio of heterozygosity compared to other breeds in every window of 10 consecutive SNPs, using 50.000 SNPs evenly distributed throughout the genome (Affymetrix 50 K canine SNP array). The strongest signal of reduced heterozygosity was localized to a region of ≈3,7 Mb on CFA 13 (CanFam 2.0 chr13:23,4879,92-27,227,623). The ratio of heterozygosity in Shar-Pei dogs was 10-fold more reduced compared to the average ratio found in dogs from the control breeds (Fig 4 A) and nearly complete homozygosity was observed in Shar-Peis near the HA synthase 2 (*HAS2)* gene (Fig 4 C). HA is synthesized by three different HA synthases, (HAS1, HAS2, HAS3) with HAS2 being the rate limiting enzyme (Weigel *et al.*, 1997). Given the strong signal of selection and knowing that the skin condition is due to excess deposits of HA, this emerged as an obvious candidate region for the mutation causing the wrinkled skin.

In parallel, more dogs had been collected to perform a GWAS for the fever syndrome, and importantly the classification of phenotypes were more strictly defined than previously. In total 22 cases and 17 controls were compared and now reached genome-wide significance (best SNP $P_{raw} = 2.3 \times 10^{-6}$, $P_{genome} \approx 0.01$ based on 100.000 permutations)(Fig 4 B). A set of 17,227 SNPs common to both the 27K and 50K array, were used in the analysis since all individuals were not run on the same type of array. The software package PLINK (Purcell *et al.*, 2007) was used for association analysis. As can be seen in Fig 4C the most associated SNPs (blue line) were located on CFA 13 close to the region under selection. Since association cannot be detected where there is no variation, and thus goes down where homozygosity (red line) goes up, it is hard to determine the most likely location for a causative mutation.

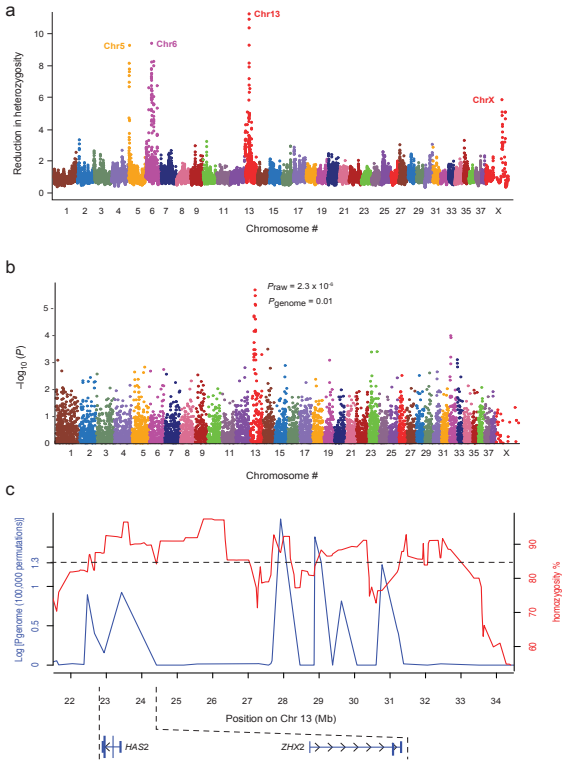


Figure 4. Signs of a selective sweep and association to FSF
a) The largest reduction in heterozygosity in Shar-Peis (n=50) compared to 24 other breeds (n=230) was observed on chromosome 13. The relative heterozygosity was calculated sliding through 50,000 genome-wide SNPs with a window size of 10 consecutive SNPs. (b) A case-control study for Familial Shar-Pei Fever (FSF) showed the strongest association on CFA 13 (best SNP P_{raw} = 2.13 x 10^{-6}, P_{genome} ≈0.01 based on 100.000 permutations). (c) Association with FSF co-localized with signals of selection on CFA 13. SNPs associated with FSF (blue line) were interspersed with signals of selection (red line). (Figure modified from paper II)

To search for the mutation causing the wrinkled skin phenotype (hyaluronanosis) we targeted the candidate region using a custom-designed sequence capture array from Roche NimbleGen. The target capture was followed by re-sequencing of a 1.5 Mb large region (CanFam 2.0 chr13:22,937,592 24,414,650). This region was considered the most likely for having been under selection, and included the candidate *HAS2* gene. The

region also included a non-coding RNA, *HAS2 antisense*, that has been proposed to act as a negative regulator of *HAS2* (Chao & Spicer, 2005). Seven dogs were re-sequenced, two meatmouth Shar-Peis with high HA serum levels, two traditional type Shar-Peis and three dogs from other control breeds. The obtained sequence reads were mapped to the Boxer reference genome sequence, and ≈1500 SNPs and ≈670 indels were identified in each dog. Nine variants (eight SNPs and one indel) were both unique to the two sequenced meatmouth Shar-Peis and were located within conserved elements. Additional genotyping of these variants in several Shar-Peis and dogs from other breeds showed that they were not specific to Shar-Peis and subsequently excluded as being causative for the phenotype. Comparisons of the coverage in the target region revealed two overlapping duplications in Shar-Pei dogs. Copy number analyses were performed in several dogs for the two duplications, and confirmed that the duplications were unique to Shar-Peis. The larger duplication with a size of 16,1 kb was unique to meatmouth Shar-Peis (CanFam 2.0 Chr13: 23,746,089–23,762,189) while a smaller duplication of 14,3 kb seemed to have its origin in the traditional Shar-Pei (CanFam 2.0 Chr13:23,743,906–23,758,214).

Further it was discovered that there was a significant correlation between meatmouth copy number and the Familial Shar-Pei Fever (FSF) ($p=<0.0001$, Mann Whitney test). The fever syndrome affected most dogs with more than six copies, whereas dogs with less than four copies were not.

In a limited study, using dermal fibroblasts from six meatmouth Shar-Peis, correlation between duplication copy number and RNA expression of *HAS2* and the *HAS2* antisense gene (*HAS2as*) was examined. Both genes showed a strong correlation between increased expression and increasing copy number.

4.3 Discussion and future prospects

In this study we used homozygosity mapping to identify a region under selection in Shar-Peis. We used this approach because we searched for a homozygous region shared by all individuals in the breed, created by strong artificial selection. Homozygosity mapping can also be used to identify homozygous regions present only within cases for a recessive trait. The mapping procedure is principally the same, but the homozygous region arose for a different reason. Following homozygosity mapping, next generation sequencing of the identified target region was performed. Genome-wide association mapping for FSF suggested that the region containing a susceptibility locus for the fever syndrome co-localize with the region under selection for the wrinkled skin phenotype. We identified a duplication of 16,1 kb unique to meatmouth Shar-Peis located 350 kb upstream of *HAS2*. Based on our findings we postulated that this duplication is the causative mutation for both hyaluronanosis and Shar-Pei fever because of the observed correlation

between the number of copies and susceptibility to FSF. We suggested that one or more regulatory elements in the duplication influence *HAS2* mRNA expression, leading to the higher levels of HA observed in dogs from this breed. We proposed that the duplication found in traditional type Shar-Pei dogs was the first duplication event making this region unstable, facilitating the second meatmouth duplication to occur by unequal crossing over.

HA has a high turnover, and is degraded into polymers of decreasing size. These hyaluronan fragments have wide-ranging and sometimes opposing biological functions. Large polymers are space filling and immunosuppressive, while smaller fragments seem to act as "danger signals" and are inflammatory and immune-stimulatory (Stern *et al.*, 2006).

The functional consequences of excessive HA in Shar-Peis need to be investigated further, but given the function of fragmented HA it is expected that the strong selection for the hyaluronanosis phenotype is contributing to induction of recurrent episodes of fever and inflammation. Approximately 60% of human patients with similar fever diseases are currently unexplained and we therefore suggest that the possible involvement of HA regulators could be relevant to investigate in more depth also in human patients.

Finally, this study illustrates how a copy number variation can affect the phenotype and how strong artificial selection for a desired trait may have a pleiotropic effect, with negative impact on the health of our companion animals.

5 SEQscoring: a tool to facilitate the analysis of data from next generation sequencing projects (Paper III)

5.1 Background

The goal of a GWAS is to locate a region that is associated with a trait or disease. Subsequent fine-mapping with a denser set of genetic markers can then reduce the size of the associated region. Finally, re-sequencing of the associated region is required to identify candidate mutation/s that needs to be functionally validated as being the causative mutation. Before the development of next generation sequencing (NGS) technologies, mutation detection by re-sequencing was a tedious and expensive task. For large regions, re-sequencing efforts have for that reason typically been limited to protein coding genes. An identified mutation in a coding exon will be the easiest to interpret, and several disease causing mutations inherited in a Mendelian fashion has been identified in coding exons (Altshuler *et al.*, 2008). Yet, many regions outside genes have important regulatory effects, for example, with respect to the location, timing, and amount of gene expression. Regulatory mutations are likely to be common particularly in complex diseases (Epstein, 2009). NGS allows rapid, affordable, and comprehensive re-sequencing, and thus provide the opportunity to investigate larger genomic regions for identification of disease-causing mutations. Causation of identified candidate mutations needs then to be confirmed by functional assays.

In the two projects previously described in this thesis (Papers I and II) the subsequent step after genomic mapping was to perform re-sequencing. The genomic susceptibility region in Shar-Peis for harboring the causative mutation

selected for the skin phenotype was ≈1.5 Mb in size. Conveniently we were in phase with the recent introduction of NGS technologies on the market and it was thus feasible to re-sequence the entire region in a few cases and controls. After re-sequencing we were faced with the large amount of data produced, and stood in front of the next challenge of how to extract the most essential information and how to be able to identify the most likely causative mutation. During the course of data analyses, the idea was born to make use of our lessons learnt, and develop a web-based tool that would be easy to use.

It has been shown that elements that are conserved across species, and thus are under purifying selection, are more likely to have biological function (Birney *et al.*, 2007; Drake *et al.*, 2006; Woolfe *et al.*, 2005; Margulies *et al.*, 2003). Consequently, the web-based software tool SEQscoring that we developed, utilizes the power of comparative genomics and was developed to score variants according to the degree of conservation at their location. SEQscoring also assesses the pattern of variants between cases and controls in order to identify a set of the most likely causative mutations for a trait.

5.2 Methods and Results

Several programs have been developed to map millions of reads to a reference sequence and to call variants, *e.g*. BWA (Li & Durbin, 2009), SAMtools (Li *et al.*, 2009) and MAQ (Li *et al.*, 2008). SEQscoring supports several file formats as input data. In figure 5 an overview of SEQscoring functionalities is shown.

In the "Scoring module", variants are scored according to the degree of conservation at the genomic position. SEQscoring keeps a local database of species alignments from some different sources where the degree of conservation has been assessed (*e.g*. 29 mammal constraint scores, 16 amniota vertebrates and human/mouse/rat/dog/ comparison (Lindblad-Toh *et al.*, 2011; Paten *et al.*, 2008; Siepel *et al.*, 2005). The conservation score source is selected by the user and a file with all detected variants is uploaded to the website. All variants are accordingly checked and recorded with a score from the database, and the information is returned to the user.

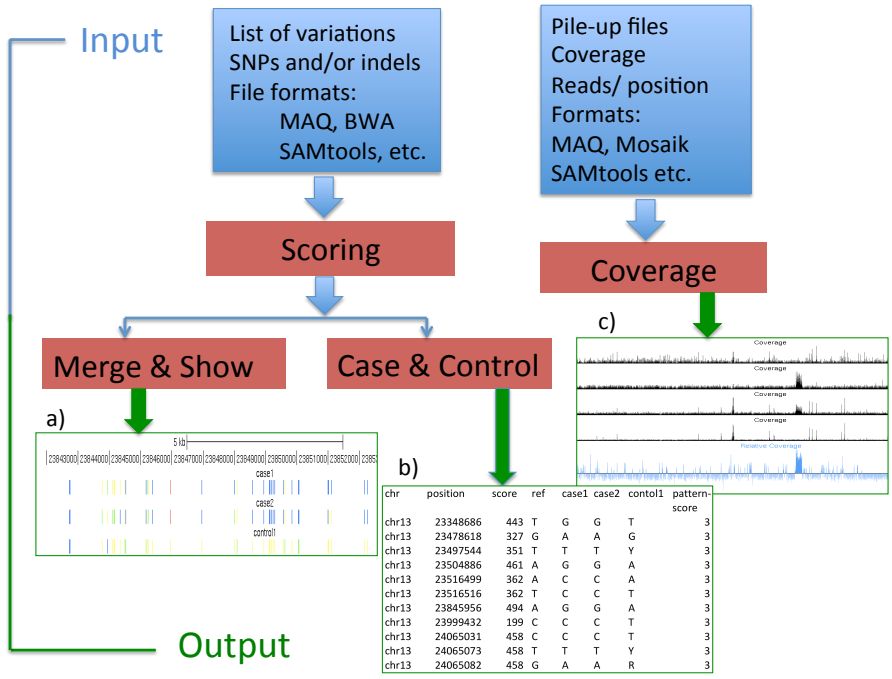


Figure 5. Overview of SEQscoring functionalities.
Input to the program is submitted by the user in the form of lists of variants and/or information about coverage/position. Variants are scored according to evolutionary conservation at the genomic position for the variant. a) SNPs are color-coded (as explained in the text) and individual samples can be merged and displayed in the UCSC browser for visual inspection of shared haplotypes and presence of variants at conserved positions (red). b) The Case & Control module performs pairwise comparisons of all samples and helps the user to rank the variants by concordance with an expected pattern for a causative variant. c) Coverage calculations are performed to find differences between cases and controls, with the goal to localize structural variants, like deletions or duplications. (Figure modified from paper III.)

The "Merge & Show" module merges the results for all samples and facilitates interpretation by visualization. Variants are displayed in the UCSC genome browser for easy comparison of samples, and investigation of haplotype structure. The SNPs are color-coded in the following way: homozygous SNPs within or near (± 5bp) conserved elements are colored red; heterozygous SNPs within or near (± 5bp) are colored pink; non-conserved homozygous SNPs

equal to the reference are colored yellow; homozygous SNPs deviating from the reference are colored blue; heterozygous non-conserved SNPs are colored green.

The "Case & Control" module helps the user to rank variants by differences between cases and controls. The user is offered three options: 1) to rank variants according to an expected pattern between cases and controls by pair-wise comparison of all samples; 2) to transform data for performing traditional association studies, 3) to compare genomic regions, by using a window of a specific size and, sliding through all consecutive variants.

Using the first option, variants are selected as the most likely for being causative, based on whether they are located within a conserved element and whether segregating as expected for a phenotype-genotype correlation. The second option is useful if the number of samples is large enough for doing an association study. The third option, to compare genomic regions, is most useful for identifying selective sweeps or homozygous regions harboring a mutation for a recessive trait.

The "Coverage module" aims to find structural variations, such as deletions or duplications (copy number variations). Coverage for different samples can vary, and therefore data is normalized, to obtain comparable figures. The ratio of average coverage between cases and controls in consecutive windows of user-specified size are calculated, and the results are visualized as graphs that can be displayed in the UCSC genome browser.

In the paper we exemplify the use of SEQscoring by describing the selection of a set of the most likely candidate mutations in the Shar-Pei project. We had re-sequenced a ≈1.5 Mb target region using Illumina Genome Analyzer, and mapped the reads to the CanFam 2.0 (Lindblad-Toh *et al.*, 2005) reference genome sequence using the software tool MAQ (Li *et al.*, 2008). We analysed two "meatmouth" Shar-Peis with an excess of serum HA and compared them to three normal controls from other breeds. ≈1500 SNPs/ per individual were detected. We scored the variants by conservation according to the UCSC PhastCons alignment of four species (Siepel *et al.*, 2005) and used the "Merge and Show" module to merge the information for the individual samples. A total of 3430 SNPs were detected that differed compared to the reference, and out of these 84 were located within conserved elements. The "Case & Control" module allowed ranking of the SNPs and we found that only eight of the conserved SNPs had a pattern where the two Shar-Peis were alike and differed from the controls. Using the "Coverage module" coverage was checked for every 10^{th} position and the average coverage for cases was compared to the average coverage for controls for every consecutive window with a size of 1 kb. As can been seen in figure 6 there was one distinct peak of

excessive coverage in the two Shar-Peis. The blue graph shows the log2 values of the ratios between cases and controls. In paper II we showed that Shar-Peis have a 16.1 kb duplication at the site of this peak.

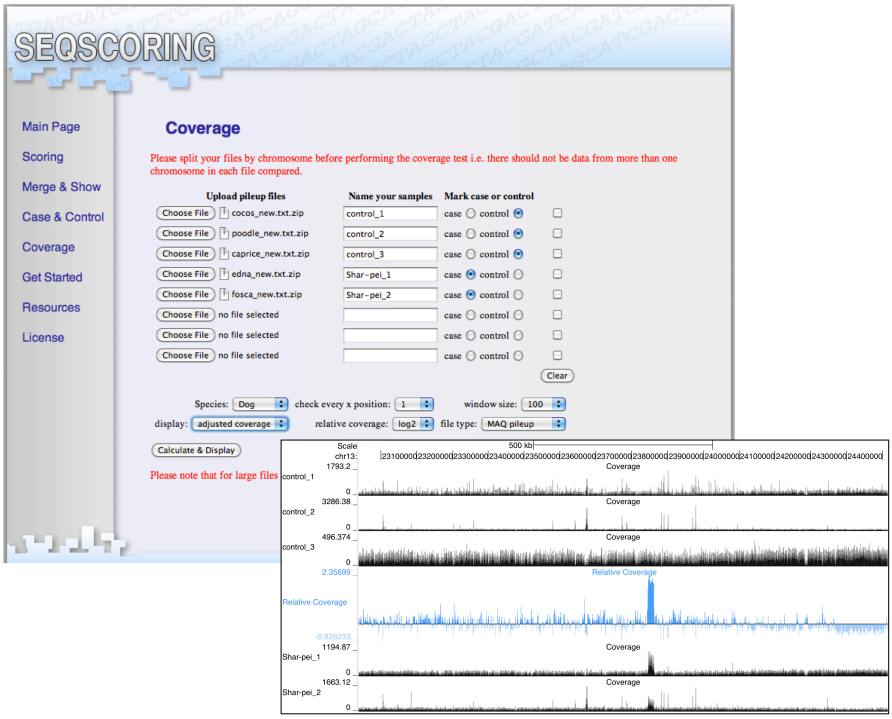


Figure 6. Copy number detection by coverage comparison.
An example of how the form is filled out in SEQscoring for doing coverage comparisons. The results are displayed below the form, with the blue graph showing the coverage ratio (log2) between cases and controls. The site of the peak was shown to harbor a copy number variation.

5.3 Discussion and future prospects

The publicly available SEQscoring tool was developed to facilitate the interpretation of data from NGS-projects. At the time we expected the user to re-sequence a limited set of samples (≈6-12) and our goal was to evaluate the

vast amount of variants present and extract a set of the most likely causative mutations for a trait or disease. We proposed a method where extracted variants should be validated by genotyping in a larger cohort of cases and controls. The program has been frequently used in our laboratory and by our collaborators. The methodology has been proven successful in several cases to narrow the possibilities and pinpoint highly likely susceptibility mutations.

We focus on filtering out conserved variants, but we recommend our users to select also non-conserved variants for evaluation in a larger cohort, since functional elements sometimes show a low degree of sequence conservation. Development is moving towards even lower costs, giving the opportunity to pool and re-sequence many more samples. SEQscoring also offers the opportunity to transform data to perform association studies (PLINK (Purcell *et al.*, 2007) format), which may be more suitable in projects including large sample sets.

Some limitations should be mentioned for targeted re-sequencing projects. It will not be possible to detect larger insertions relative to the reference genome sequence, since such sequences will not be included in the probes used to capture the target region. Since the read length is quite short (≈30-100 bp) it limits the size of repetitive regions that can be read through, which makes differences in size of microsatellites, presence of long interspersed nucleotide elements (LINEs) and short interspersed nucleotide elements (SINEs) etc. hard to detect. It appears that we are moving towards longer reads for many of the new technologies. Longer paired end reads and the use of *de novo* assembly into longer contigs prior to mapping might help overcome some of these limitations.

There are several functionalities that could be added to SEQscoring in the future. For instance, the user can currently only determine if a variant is located within a conserved variant. A lot more information could be useful for the user, as if the variant is located in a protein coding exon, if the mutation will cause a codon change, and if there is a regulatory element at the site with known function. SEQscoring could also be given functions to evaluate data from whole genome RNA sequencing. Here it would be useful to extract information about new transcripts, and new splice variants, as well as differences in expression of different transcripts in cases and controls. Protein interactions is another compelling feature to add, to get an integrated picture of how extracted candidates fit in a network, and thereby get a broader understanding of pathways and genes involved for a certain condition.

6 Across-breed genome-wide association mapping identifies a glioma susceptibility locus

6.1 Background

Gliomas are primary brain tumours derived from glial cells, and the most common form of malignant brain tumors in humans. Gliomas are classified according to different grades of malignancy and patients with the most malignant form, grade IV, has an approximate survival time of one year (Louis, 2006). Compared to humans, dogs have a similar or higher incidence of gliomas (Dobson *et al.*, 2002; Hayes *et al.*, 1975). It has been observed that brachycephalic (short-nosed) dog breeds have a considerable elevated risk of developing gliomas (Hayes *et al.*, 1975).

Brachycephaly is a trait that has been under strong selection. A short broad skull and a severely shortened muzzle characterize the phenotype. The breeds with the most elevated risk, Boxer, Boston terrier and Bulldog share a common ancestor, an "ancestral Bulldog" (Hayes *et al.*, 1975). The "ancestral" Bulldog was bred for the sport "bull baiting" before it was forbidden in England 1835. According to historical records, the original Bulldog was crossed with Pugs at that time, leading to a more extreme brachycephalic phenotype that is seen in modern breeds (Voss, 1933).

Brachycephalic breeds like Pugs and Pekingese do not seem to have an elevated risk (Hayes *et al.*, 1975), making us hypothesize that glioma risk factors are descending in the ancestral Bulldog line, and thus arose before the cross with the Pug. Because of the strongly elevated risk in certain breeds with common ancestry, we expect them to carry shared genetic risk factors for

glioma. In humans, some rare inherited mutations have been identified. It has been agreed that there is a need to expand research in genetics and molecular epidemiology of brain tumors. Even though rare inherited mutations account for few cases they are important for identifying pathways for gliomagenesis (Bondy *et al.*, 2008). We propose that mutations identified in dogs will likewise benefit both dogs and humans by the identification of genes involved in glioma pathways.

In a previous study we performed across-breed genome-wide association mapping and identified a locus at CFA 1 with strong evidence of having been under selection for brachycephaly (Bannasch *et al.*, 2010). In this study we performed an across-breed genome-wide study for glioma followed by targeted re-sequencing to search for mutations causing brachycephaly and/or glioma.

6.2 Methods and Results

We performed a GWAS for glioma using 39 cases and 142 controls from 25 different pure bred and four mixed dog breeds. The latest developed Illumina array with 173,622 SNPs was used. The strongest association was found on CFA 26 (Fig 7). We noted that the population structure caused inflated p-values, and genomic control (GC) was therefore used to correct the chi-square test statistics. The software tool PLINK (Purcell *et al.*, 2007) was used for the analysis.

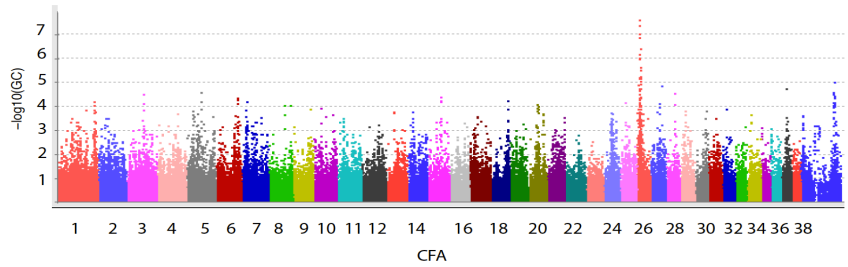


Figure 7. Across-breed GWAS for glioma.
The strongest association for glioma appears on chromosome 26

Five breeds among the cases are related to the "ancestral Bulldog": Boxer, English Bulldog, French Bulldog, Boston terrier, and Staffordshire terrier. We investigated the possibility of a selective sweep common to those breeds and found a completely homozygous region with respect to SNP markers spanning ≈ 4Mb, including the most glioma-associated region. There were no comparable signs of selection in this region in Pugs that are brachycephalic but not related to the ancestral Bulldog.

The next step was to go from SNP markers, to investigate the full sequence in order to identify candidate mutations causing brachycephaly and/or glioma. We performed targeted re-sequencing of in total 7 Mb from the previously identified brachycephaly locus on CFA 1 and the newly identified glioma locus on CFA 26. In total nine dogs were re-sequenced (two Boxers, one Pug, one English Bulldog, one French Bulldog, one Boston terrier, one Dachshund, one Welsh Corgi, and one Basset hound). Four of the dogs had been diagnosed with glioma.

In total 20,998 SNPs were identified. We used SEQscoring (Truvé *et al.*, 2011) (Paper III) as previously described and found that out of these, 1,086 SNPs were located within conserved elements. The SNPs were ranked according to the expected pattern of segregation between cases and controls for causative variants. In addition, two structural variations were identified (one 160 bp duplication, and one 2200 bp insertion). The structural variants and the top ranked SNPs from both conserved and non-conserved SNPs were selected for validation by genotyping.

In total 100 candidate SNPs were successfully genotyped in 168 dogs. Since we were interested in both glioma and brachycephaly the data was divided in subgroups for the analyses. The two structural variants were less associated than the best SNPs. One SNP located on CFA 1 was very strongly associated with brachycephaly. We located strongly associated candidate SNPs for glioma

on CFA 26. The best two SNPs were located in introns of excellent candidate genes based on biological function, and within newly discovered intronic transcripts with unknown function. A third SNP got our attention since it caused a non-synonymous codon change.

6.3 Discussion and future prospects

We have mapped a locus on CFA 26 associated with glioma and shown that this region has probably been under selection and is thus shared by all the high-risk breeds. When emphasis was on finding a brachycephaly locus including brachycephalic breeds with no evidence of an increased glioma risk, the strongest association was found on CFA 1. The history of breeds and our results, suggest that the locus on CFA 26 is descending in breeds related to an ancestral Bulldog, while the CFA 1 locus is inherited from an ancestor of the Pug.

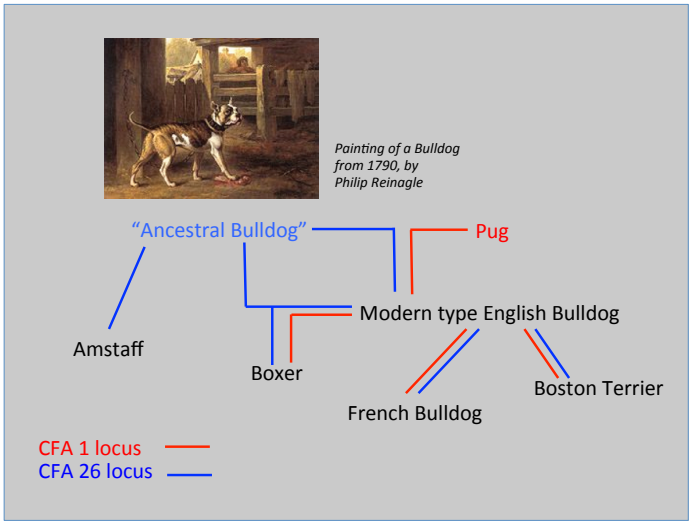


Figure 8. Suggested segregation of loci under selection.
The sweep signal on CFA26 seems to have originated in the "ancestral Bulldog" while the brachycephaly association on CFA1 seem to have its origin in an ancestor to the Pug. (Painting of "ancestral Bulldog" from 1790.)

The SNP most associated to brachycephaly as validated here was located in an intron of the gene *SMOC2*. This gene was already suggested as a candidate in our previous paper (Bannasch *et al.*, 2010), but this present research either brought us much closer to or actually hit a causative variant, since the best validated SNP was much more associated than the best SNP from the GWAS. The position of the SNP is not conserved, and its function is thus not obvious. The *SMOC2* gene product is a matricellular protein that is expressed in the craniofacial region of developing mouse embryos and therefore *SMOC2* is a strong candidate gene (Bloch-Zupan *et al.*, 2011). It is possible that the site of the SNP harbors a transcription-factor binding site that regulates *SMOC2* transcription but further research is required to functionally evaluate the mutation.

For glioma we identified three genes as very likely candidates for being involved in disease development. We think that follow up research in humans should not be limited to the SNPs found in dogs, but all differences that could be found between normal and tumor cells for those genes could be of interest in the unraveling of cause for disease development. GWAS studies in humans have shown that a single locus can harbor several different risk variants. Even if individual SNPs might have a small effect, drug targeting of the encoded protein could have a much greater effect (Altshuler *et al.*, 2008). The most important result from mapping studies in both humans and dogs is the gained knowledge about disease mechanism and disease pathways. In the manuscript (Paper IV) we report a literature research assessing how our candidates interact with genes with an established role in pathways for gliomagenesis. To our knowledge, none of our candidate genes has been proposed to be involved in the development of glioma before, but recent research in other forms of cancer (as described in the manuscript) make us propose that we have identified putative drug-targets for glioma.

For our glioma candidate genes there are several different known transcripts. In addition, recent RNA sequencing in dogs (unpublished data) suggests the existence of small intronic transcripts and more splice variants than what has previously been detected. Intriguingly for one of the candidate genes other research has shown that some splice variants are more commonly expressed in cancer cells including glioma (Davare *et al.*, 2011; Frigo *et al.*, 2011; Hsu *et al.*, 2001). Given this we propose that it is important to investigate the expression pattern of all transcripts of these genes in more detail. At present, we are investigating the presence of some of the transcripts in glioma and normal cells for both dogs and humans. A more comprehensive analysis comparing whole genome RNA sequence data for glioma and normal cells, in both dogs and humans would probably be the best way to continue the

research. For one of the candidates other research has shown that inhibiting its expression blocks migration of cancer cells (See paper IV). Of great interest is to test if this would hold true also for glioma cells. Preferable such inhibition studies could distinguish between different splice variants and intronic transcripts. Finally because of the similarity between disease development in dogs and humans the dog also offers an excellent model to test new treatment options.

7 Conclusions

The results described in this thesis confirm that the dog is an excellent model for disease mapping. In addition it can be concluded that the dog with advantage can be used also for mapping of complex disease. Since it is much more problematic to map complex disease in human populations the dog model will likely continue be a helpful complement in revealing pathways for the development of human disease.

Three different complex diseases were mapped in dogs and we also showed examples of the use of three different mapping procedures. Conclusions were drawn and the results and future prospects were discussed for each project separately in this thesis. However, the main conclusions for each project are as follows:

- Five loci have been identified that are strongly associated with an SLE-related disease in Nova Scotia duck tolling retriever. Strong candidate genes involved in regulation of T-cell activation was found in these regions. In this work GWAS within one breed was performed followed by fine-mapping to validate the regions of association. Mapping within breed is preferable performed when possible for best taking advantage of the genomic structure of the dog.

- A duplication of 16,1 kb unique to meatmouth Shar-Peis was identified 350 kb upstream of the *HAS2* gene. We concluded that this duplication is likely the causative mutation for both the heavily wrinkled and thickened skin phenotype in Shar-Pei dogs, and for the fever syndrome common in this breed. In this study we showed that homozygosity mapping can be used to identify a mutation that has been strongly selected for and is fixed within a breed.

- A web-based software tool SEQscoring was developed. The use of targeted NGS following the mapping of an associated locus offers both opportunities and challenges. SEQscoring has been shown to be useful for extracting and identifying a set of the most likely causative mutations from the vast amount of variants identified in NGS projects. The extracted set of mutations should then be validated in a larger cohort of samples, followed by functional studies.

- A locus associated with glioma was mapped on CFA 26. Mutations were identified in genes that closely interact with genes known to be involved in human glioma pathways. In this study we performed across-breed GWAS. Across-breed mapping should be performed with care, but will work best for traits or diseases in breeds sharing a recent common ancestor.

8 Populärvetenskaplig sammanfattning

8.1 Hunden kan användas för att hitta sjukdomsframkallande mutationer

Hundar och människor drabbas av samma sjukdomar som t.ex. cancer, epilepsi, autoimmuna sjukdomar och hjärt- och kärlsjukdomar. Våra gener är också mycket lika hundens gener, vilket gör att om vi kan hitta orsaken till en ärftlig sjukdom i hundens arvsmassa, så kan denna kunskap hjälpa oss att förstå också hur sjukdomen uppkommer och utvecklas hos oss människor. Det finns idag mer än 400 hundraser och de flesta har skapats genom avel under de senaste 200 åren. De flesta hundar inom en ras är som bekant ganska lika varandra och ofta är det bara ett fåtal individer som från början varit med och bildat varje ras. Renrasiga hundar har stamtavlor som visar deras släktträd och ingen avel över rasgränser får förekomma. På grund av denna strikta avel är variationen i arvsmassan mellan olika hundar inom en hundras mycket lägre än variationen mellan olika människors arvsmassa. Vissa sjukdomar är också ofta mycket vanligare inom vissa hundraser. Att det har blivit så beror antagligen både på slumpen och på att vissa sjukdomsframkallande mutationer av misstag har ”liftat” med en egenskap som selekterats för inom en ras. Detta kan ske då en mutation som orsakar sjukdom ligger nära något som man vill selektera för i arvsmassan. En och samma mutation kan också i vissa fall ha flera effekter både önskade och oönskade.

För att hitta var i arvsmassan mutationerna ligger görs genetisk kartläggning med hjälp av s.k. genetiska markörer d.v.s. positioner i arvsmassan som ofta varierar mellan individer. Flera hundar som har diagnosticerats med en sjukdom jämförs gentemot friska kontroller. Om ett område identifieras där de sjuka hundarna har mer lika markörer och skiljer sig från kontrollerna, så ligger troligtvis den sjukdomsframkallande mutationen i närheten d.v.s. detta område i arvsmassan är associerat med sjukdomen. För att hitta själva mutationen

måste varje position i arvsmassan inom detta område analyseras. Många sjukdomar kan bero på flera orsaker. Det kan vara flera olika gener som samverkar, eller defekter i olika gener som ger samma symptom. Miljön och slumpmässiga faktorer kan också vara avgörande för huruvida man drabbas av sjukdom. Sjukdomar som påverkas av flera faktorer kallas för komplexa sjukdomar. Det kan vara lättare att hitta orsaken till komplexa sjukdomar hos hund än hos människan på grund av att hundar från en och samma ras har mindre variation i sin arvsmassa samt att det inom en ras troligen bara finns ett fåtal ofta starka riskfaktorer om sjukdomen är överrepresenterad inom rasen. Arvsmassan som också kallas för genomet består av långa dubbel-strängar av deoxiribonukleinsyra (DNA). DNA kan utvinnas från ett vanligt blodprov.

I den här avhandlingen har tre olika sjukdomar hos hund med förekomst av liknande komplexa sjukdomar hos människa studerats. Syftet har varit att hitta regionen i genomet som är associerad med sjukdom och därefter identifiera mutationen som orsakar sjukdom. Hos hundrasen Nova Scotia duck tolling retriever är det vanligare än hos andra hundraser med en autoimmun sjukdom som liknar sjukdomen SLE (systemisk lupus erythematosus) hos människa. Hundrasen Shar-Pei får ofta en sjukdom som ger feber och svullna leder. Denna sjukdom kallas Shar-Pei feber och liknar så kallade autoinflammatoriska sjukdomar som också drabbar människor. Både människor och hundar kan drabbas av hjärntumörer. Den vanligaste maligna formen hos människa kallas gliom. Vissa raser med kort nos som boxer, Boston terrier och bulldog har en mycket högre risk än andra raser att drabbas av gliom.

8.2 En SLE-liknande sjukdom kartlagd hos Nova Scotia duck tolling retriever

I avhandlingens första delarbete beskrivs hur vi sökt efter genetiska riskfaktorer för ett SLE-liknande sjukdomskomplex hos Nova Scotia duck tolling retriever. Vi använde DNA från 81 sjuka hundar och 57 kontroller och analyserade dessa med 22000 SNP markörer i hundens DNA. Vi identifierade 5 områden i genomet som är associerade med det SLE-liknande sjukdomskomplexet. För att verifiera att vi hittat rätt område utfördes en så kallad fin-mappning där vi analyserade de identifierade områdena i mer detalj, genom att analysera fler markörer inom regionerna och genom att inkludera

fler hundar i analysen. Inom dessa områden finns flera gener. Några av de mest associerade generna har en känd biologisk funktion som tyder på att det är mycket troligt att dessa gener är inblandade i sjukdomens uppkomst. Vid autoimmuna sjukdomar attackerar kroppens egna immunförsvar olika vävnader i kroppen. En typ av celler i immunförsvaret som brukar vara inblandade är så kallade T-celler. Flera av de gener som vi fann vara associerade är kända för att vara inblandade i samma process av händelser som leder till aktivering av T-celler. Nästa steg är att identifiera mutationerna, och att undersöka om motsvarande gener är involverade i sjukdomen SLE hos människor.

8.3 Selektiv avel för rynkig hud hos hundrasen Shar-Pei kan medföra sjukdom

I delarbete II beskrivs hur vi hittade mutationen som orsakar en febersjukdom hos Shar-Pei hundar. Vi hittade till att börja med ingen association och antog därför att det möjligen var för lite variation i området för att denna metod skulle fungera. Shar-Pei hundar har en karakteristisk skrynklig hud-fenotyp som det har kraftigt selekterats för inom aveln. När det görs en sådan kraftig selektering medför det att, i området omkring den orsakande mutationen blir hundarna inom rasen mer lika än vad de är i andra delar av genomet. Vi prövade därför en annan kartläggningsmetod, som går ut på att hitta just sådana områden där individer är mer lika varandra. Vi hittade ett sådant område precis bredvid en gen som är involverad i syntes av hyaluronsyra. Det är känt att det finns mycket hyaluronsyra i huden hos Shar-Pei hundar och att detta troligen orsakar den tjocka veckade huden. Vi var därför övertygade om att vi hittat det område som selekterats för och gjorde ett antagande att den mutation som orsakar febern fanns dold i detta område. Vi sekvenserade därför 1.5 miljoner baspar av DNA för att försöka hitta mutationen. Vi hittade ett område på 16.1 tusen baspar som hade blivit duplicerat och alltså fanns i flera kopior. Vi kunde därefter visa att denna mutation säkerligen orsakar både den skrynkliga huden och medför risk för att utveckla febersjukdomen. Denna studie pekade också på att hyaluronsyra eller andra gener som samverkar med detta ämne i kroppen kan vara intressant att studera i mer detalj hos människa där det för de flesta fall av autoinflammatoriska sjukdomar inte finns någon känd orsak.

8.4 SEQscoring ett verktyg för att underlätta analys av stora mängder DNA sekvens

I delarbete III beskriver vi ett datorverktyg som vi utvecklat för att underlätta analys av stora mängder DNA sekvens. Detta verktyg finns på en websida och är tillgängligt för alla som vill använda det. Som beskrivet ovan är första steget att hitta en region i genomet där en mutation för en sjukdom eller egenskap kan finnas. Nästa steg är att sekvensera detta område. På varje position i DNA sekvensen finns en av fyra så kallade nukleotider eller baser, adenin, cytosin, tymin och guanin. Dessa brukar skrivas med bokstäverna A, C, T och G. Den sekvens som ska analyseras kan vara flera miljoner baser lång. I en så lång sekvens kan många av positionerna variera mellan individer, och svårigheten är att veta vilka varianter som är mest troliga att orsaka sjukdom. Vissa DNA sekvenser är konserverade och alltså väldigt lika mellan arter. Det är mest troligt att de konserverade områdena har en funktion, och mutationer i sådana områden av arvsmassan är alltså mest troliga att orsaka sjukdom eller andra förändringar. Programmet som vi utvecklat kallas för SEQscoring och hjälper användaren genom att markera vilka varianter som är konserverade. Om flera individer sekvenserats, görs också jämförelser mellan fall och kontroller för att filtrera fram ett litet antal av de mest troliga varianterna. Dessa varianter kan därefter undersökas vidare hos fler individer, för att slutligen kunna avgöra vilken eller vilka varianter som är orsakande eller medför ökad risk för sjukdom. När en sådan här stor sekvensering utförs erhålls i första steget miljarder av korta sekvenser på mellan 30-100 baser för varje individ. Eftersom hela hundgenomet har sekvenserats så kan dessa korta sekvenser passas in mot denna referens, och enstaka positioner som varierar kan identifieras. SEQscoring kan också användas för att kontrollera hur mycket täckning det finns för olika regioner, och därigenom kan ett område som har blivit duplicerat eller deleterats i genomet upptäckas.

8.5 Kartläggning av ökad genetisk risk att drabbas av hjärntumör

I artikel IV beskrivs hur vi hittat ett område i genomet som är associerat med risk att få en typ av hjärntumör som kallas gliom. I denna studie användes 173622 markörer jämt spridda över hela genomet. Vi jämförde 39 fall av gliom gentemot 142 friska kontroller. Både gliomfallen och de friska kontrollerna kom från flera olika hundraser. När man utför en associationsstudie på detta sätt mellan raser är variationen mycket högre än inom en ras. Det är större risk att hitta falska positiva och det behövs fler markörer för att hitta rätt område. Vi hittade ett område som var mycket starkt associerat. Det är känt att vissa

raser med kort nos har högre risk att utveckla gliom, och dessa raser har ett gemensamt ursprung från en typ av ”original bulldog”. De raser som har högst risk att utveckla gliom är engelsk bulldog, Boston terrier och boxer. Vi kunde konstatera att i det associerade området för gliom var dessa hög-risk- raser extremt lika. Efter associations-studien gjorde vi en omfattande sekvensering och använde SEQscoring för att filtrera ut de mest troliga mutationerna. De mutationer vi hittat ligger i gener som tidigare inte indikerats att vara inblandade i utvecklandet av gliom. En litteraturstudie visar att de samverkar med gener som har en känd roll i gliomutveckling. Än mer positivt är att när genuttrycket av en av dessa kandidat-gener hindras i andra typer av cancer har mycket goda resultat uppnåtts. Detta ger hopp om att även tillväxt och spridning av gliomtumörer kan hämmas genom att inhibera denna gen. Nu behövs funktionella studier av dessa kandidat-gener, och förhoppningsvis kan det på sikt leda fram till nya mediciner, där även hunden kan vara till hjälp för att testa fram nya behandlingsformer.

References

Altshuler, D., Daly, M.J. & Lander, E.S. (2008). Genetic mapping in human disease. *Science* 322(5903), 881-8.

Anfinsen, K.P., Berendt, M., Liste, F.J., Haagensen, T.R., Indrebo, A., Lingaas, F., Stigen, O. & Alban, L. (2008). A retrospective epidemiological study of clinical signs and familial predisposition associated with aseptic meningitis in the Norwegian population of Nova Scotia duck tolling retrievers born 1994-2003. *Canadian journal of veterinary research = Revue canadienne de recherche veterinaire* 72(4), 350-5.

Bannasch, D., Young, A., Myers, J., Truve, K., Dickinson, P., Gregg, J., Davis, R., Bongcam-Rudloff, E., Webster, M.T., Lindblad-Toh, K. & Pedersen, N. (2010). Localization of canine brachycephaly using an across breed mapping approach. *PloS one* 5(3), e9632.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., Kuehn, M.S., Taylor, C.M., Neph, S., Koch, C.M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J.A., Andrews, R.M., Flicek, P., Boyle, P.J., Cao, H., Carter, N.P., Clelland, G.K., Davis, S., Day, N., Dhami, P., Dillon, S.C., Dorschner, M.O., Fiegler, H., Giresi, P.G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K.D., Johnson, B.E., Johnson, E.M., Frum, T.T., Rosenzweig, E.R., Karnani, N., Lee, K., Lefebvre, G.C., Navas, P.A., Neri, F., Parker, S.C., Sabo, P.J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F.S., Dekker, J., Lieb, J.D., Tullius, T.D., Crawford, G.E., Sunyaev, S., Noble, W.S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I.L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H.A., Sekinger, E.A., Lagarde, J., Abril, J.F., Shahab, A., Flamm, C., Fried, C., Hackermuller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J.S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M.C., Thomas, D.J., Weirauch, M.T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K.G., Sung, W.K., Ooi, H.S.,

Chiu, K.P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M.L., Valencia, A., Choo, S.W., Choo, C.Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T.G., Brown, J.B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C.N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J.S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R.M., Rogers, J., Stadler, P.F., Lowe, T.M., Wei, C.L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S.E., Fu, Y., Green, E.D., Karaoz, U., Siepel, A., Taylor, J., Liefer, L.A., Wetterstrand, K.A., Good, P.J., Feingold, E.A., Guyer, M.S., Cooper, G.M., Asimenos, G., Dewey, C.N., Hou, M., Nikolaev, S., Montoya-Burgos, J.I., Loytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N.R., Holmes, I., Mullikin, J.C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W.J., Stone, E.A., Batzoglou, S., Goldman, N., Hardison, R.C., Haussler, D., Miller, W., Sidow, A., Trinklein, N.D., Zhang, Z.D., Barrera, L., Stuart, R., King, D.C., Ameur, A., Enroth, S., Bieda, M.C., Kim, J., Bhinge, A.A., Jiang, N., Liu, J., Yao, F., Vega, V.B., Lee, C.W., Ng, P., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M.J., Inman, D., Singer, M.A., Richmond, T.A., Munn, K.J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J.C., Couttet, P., Bruce, A.W., Dovey, O.M., Ellis, P.D., Langford, C.F., Nix, D.A., Euskirchen, G., Hartman, S., Urban, A.E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T.H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C.K., Rosenfeld, M.G., Aldred, S.F., Cooper, S.J., Halees, A., Lin, J.M., Shulha, H.P., Xu, M., Haidar, J.N., Yu, Y., Iyer, V.R., Green, R.D., Wadelius, C., Farnham, P.J., Ren, B., Harte, R.A., Hinrichs, A.S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A.S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R.M., Karolchik, D., Armengol, L., Bird, C.P., de Bakker, P.I., Kern, A.D., Lopez-Bigas, N., Martin, J.D., Stranger, B.E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrimsdottir, I.B., Huppert, J., Zody, M.C., Abecasis, G.R., Estivill, X., Bouffard, G.G., Guan, X., Hansen, N.F., Idol, J.R., Maduro, V.V., Maskeri, B., McDowell, J.C., Park, M., Thomas, P.J., Young, A.C., Blakesley, R.W., Muzny, D.M., Sodergren, E., Wheeler, D.A., Worley, K.C., Jiang, H., Weinstock, G.M., Gibbs, R.A., Graves, T., Fulton, R., Mardis, E.R., Wilson, R.K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D.B., Chang, J.L., Lindblad-Toh, K., Lander, E.S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B. & de Jong, P.J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146), 799-816.

Bloch-Zupan, A., Jamet, X., Etard, C., Laugel, V., Muller, J., Geoffroy, V., Strauss, J.P., Pelletier, V., Marion, V., Poch, O., Strahle, U., Stoetzel, C. & Dollfus, H. (2011). Homozygosity mapping and candidate prioritization identify mutations, missed by whole-exome sequencing, in SMOC2, causing major dental developmental defects. *American journal of human genetics* 89(6), 773-81.

Bondy, M.L., Scheurer, M.E., Malmer, B., Barnholtz-Sloan, J.S., Davis, F.G., Il'yasova, D., Kruchko, C., McCarthy, B.J., Rajaraman, P., Schwartzbaum, J.A., Sadetzki, S., Schlehofer, B., Tihan, T., Wiemels, J.L., Wrensch, M. & Buffler, P.A. (2008). Brain tumor epidemiology: consensus from the Brain Tumor Epidemiology Consortium. *Cancer* 113(7 Suppl), 1953-68.

Chao, H. & Spicer, A.P. (2005). Natural antisense mRNAs to hyaluronan synthase 2 inhibit hyaluronan biosynthesis and cell proliferation. *The Journal of biological chemistry* 280(30), 27513-22.

Clipstone, N.A. & Crabtree, G.R. (1992). Identification of calcineurin as a key signalling enzyme in T-lymphocyte activation. *Nature* 357(6380), 695-7.

Davare, M.A., Saneyoshi, T. & Soderling, T.R. (2011). Calmodulin-kinases regulate basal and estrogen stimulated medulloblastoma migration via Rac1. *Journal of neuro-oncology* 104(1), 65-82.

Dobson, J.M., Samuel, S., Milstein, H., Rogers, K. & Wood, J.L. (2002). Canine neoplasia in the UK: estimates of incidence rates from a population of insured dogs. *The Journal of small animal practice* 43(6), 240-6.

Drake, J.A., Bird, C., Nemesh, J., Thomas, D.J., Newton-Cheh, C., Reymond, A., Excoffier, L., Attar, H., Antonarakis, S.E., Dermitzakis, E.T. & Hirschhorn, J.N. (2006). Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nature genetics* 38(2), 223-7.

Drogemuller, C., Karlsson, E.K., Hytonen, M.K., Perloski, M., Dolf, G., Sainio, K., Lohi, H., Lindblad-Toh, K. & Leeb, T. (2008). A mutation in hairless dogs implicates FOXI3 in ectodermal development. *Science* 321(5895), 1462.

Epstein, D.J. (2009). Cis-regulatory mutations in human disease. *Briefings in functional genomics & proteomics* 8(4), 310-6.

Frigo, D.E., Howe, M.K., Wittmann, B.M., Brunner, A.M., Cushman, I., Wang, Q., Brown, M., Means, A.R. & McDonnell, D.P. (2011). CaM kinase kinase beta-mediated activation of the growth regulatory kinase AMPK is required for androgen-dependent migration of prostate cancer cells. *Cancer research* 71(2), 528-37.

Guerini, D. (1997). Calcineurin: not just a simple protein phosphatase. *Biochemical and biophysical research communications* 235(2), 271-5.

Hansson-Hamlin, H. & Lilliehook, I. (2009). A possible systemic rheumatic disorder in the Nova Scotia duck tolling retriever. *Acta veterinaria Scandinavica* 51, 16.

Hayes, H.M., Priester, W.A., Jr. & Pendergrass, T.W. (1975). Occurrence of nervous-tissue tumors in cattle, horses, cats and dogs. *International journal of cancer. Journal international du cancer* 15(1), 39-47.

Hedhammar, A.A., Malm, S. & Bonnett, B. (2011). International and collaborative strategies to enhance genetic health in purebred dogs. *Veterinary journal* 189(2), 189-96.

Hsu, L.S., Chen, G.D., Lee, L.S., Chi, C.W., Cheng, J.F. & Chen, J.Y. (2001). Human Ca2+/calmodulin-dependent protein kinase kinase beta gene encodes multiple isoforms that display distinct kinase activity. *The Journal of biological chemistry* 276(33), 31113-23.

Karlsson, E.K., Baranowska, I., Wade, C.M., Salmon Hillbertz, N.H., Zody, M.C., Anderson, N., Biagi, T.M., Patterson, N., Pielberg, G.R., Kulbokas, E.J., 3rd, Comstock, K.E., Keller, E.T., Mesirov, J.P., von Euler, H., Kampe, O., Hedhammar, A., Lander, E.S., Andersson, G., Andersson, L. & Lindblad-Toh, K. (2007). Efficient mapping of mendelian traits in dogs through genome-wide association. *Nature genetics* 39(11), 1321-8.

Karlsson, E.K. & Lindblad-Toh, K. (2008). Leader of the pack: gene mapping in dogs and other model organisms. *Nature reviews. Genetics* 9(9), 713-25.

Kirkness, E.F., Bafna, V., Halpern, A.L., Levy, S., Remington, K., Rusch, D.B., Delcher, A.L., Pop, M., Wang, W., Fraser, C.M. & Venter, J.C. (2003). The dog genome: survey sequencing and comparative analysis. *Science* 301(5641), 1898-903.

Koskenmies, S., Jarvinen, T.M., Onkamo, P., Panelius, J., Tuovinen, U., Hasan, T., Ranki, A. & Saarialho-Kere, U. (2008). Clinical and laboratory characteristics of Finnish lupus erythematosus patients with cutaneous manifestations. *Lupus* 17(4), 337-47.

Kyttaris, V.C., Wang, Y., Juang, Y.T., Weinstein, A. & Tsokos, G.C. (2007). Increased levels of NF-ATc2 differentially regulate CD154 and IL-2 genes in T cells from patients with systemic lupus erythematosus. *Journal of immunology* 178(3), 1960-6.

Lachmann, H.J., Goodman, H.J., Gilbertson, J.A., Gallimore, J.R., Sabin, C.A., Gillmore, J.D. & Hawkins, P.N. (2007). Natural history and outcome in systemic AA amyloidosis. *The New England journal of medicine* 356(23), 2361-71.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T.,

Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S. & Chen, Y.J. (2001). Initial sequencing and analysis of the human genome. *Nature* 409(6822), 860-921.

Lander, E.S. & Schork, N.J. (1994). Genetic dissection of complex traits. *Science* 265(5181), 2037-48.

Larson, G., Karlsson, E.K., Perri, A., Webster, M.T., Ho, S.Y., Peters, J., Stahl, P.W., Piper, P.J., Lingaas, F., Fredholm, M., Comstock, K.E., Modiano, J.F., Schelling, C., Agoulnik, A.I., Leegwater, P.A., Dobney, K., Vigne, J.D., Vila, C., Andersson, L. & Lindblad-Toh, K. (2012). Rethinking dog domestication by integrating genetics, archeology, and biogeography. *Proceedings of the National Academy of Sciences of the United States of America* 109(23), 8878-83.

Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14), 1754-60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16), 2078-9.

Li, H., Ruan, J. & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* 18(11), 1851-8.

Lin, L., Faraco, J., Li, R., Kadotani, H., Rogers, W., Lin, X., Qiu, X., de Jong, P.J., Nishino, S. & Mignot, E. (1999). The sleep disorder canine narcolepsy is caused by a mutation in the hypocretin (orexin) receptor 2 gene. *Cell* 98(3), 365-76.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L.D., Lowe, C.B., Holloway, A.K., Clamp, M., Gnerre, S., Alfoldi, J., Beal, K., Chang, J., Clawson, H., Cuff, J., Di Palma, F., Fitzgerald, S., Flicek, P., Guttman, M., Hubisz, M.J., Jaffe, D.B., Jungreis, I., Kent, W.J., Kostka, D., Lara, M., Martins, A.L., Massingham, T., Moltke, I., Raney, B.J., Rasmussen, M.D., Robinson, J., Stark, A., Vilella, A.J., Wen, J., Xie, X., Zody, M.C., Baldwin, J., Bloom, T., Chin, C.W., Heiman, D., Nicol, R., Nusbaum, C., Young, S., Wilkinson, J., Worley, K.C., Kovar, C.L., Muzny, D.M., Gibbs, R.A., Cree, A., Dihn, H.H., Fowler, G., Jhangiani, S., Joshi, V., Lee, S., Lewis, L.R., Nazareth, L.V., Okwuonu, G., Santibanez, J., Warren, W.C., Mardis, E.R., Weinstock, G.M., Wilson, R.K., Delehaunty, K., Dooling, D., Fronik, C., Fulton, L., Fulton, B., Graves, T., Minx, P., Sodergren, E., Birney, E., Margulies, E.H., Herrero, J., Green, E.D., Haussler, D., Siepel, A., Goldman, N., Pollard, K.S., Pedersen, J.S., Lander, E.S. & Kellis, M. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478(7370), 476-82.

Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas, E.J., 3rd, Zody, M.C., Mauceli, E., Xie, X., Breen, M., Wayne, R.K., Ostrander, E.A., Ponting, C.P., Galibert, F., Smith, D.R., DeJong, P.J., Kirkness, E., Alvarez, P., Biagi, T., Brockman, W., Butler, J., Chin, C.W., Cook, A., Cuff, J., Daly, M.J., DeCaprio, D., Gnerre, S., Grabherr, M., Kellis, M., Kleber, M., Bardeleben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K.P., Parker, H.G., Pollinger, J.P., Searle, S.M., Sutter, N.B., Thomas, R., Webber, C., Baldwin, J., Abebe, A., Abouelleil, A., Aftuck, L., Ait-Zahra, M., Aldredge, T., Allen, N., An, P., Anderson, S., Antoine, C., Arachchi, H., Aslam, A., Ayotte, L., Bachantsang, P., Barry, A., Bayul, T., Benamara, M., Berlin, A., Bessette, D., Blitshteyn, B., Bloom, T., Blye, J., Boguslavskiy, L., Bonnet, C., Boukhgalter, B., Brown, A., Cahill, P., Calixte, N., Camarata, J., Cheshatsang, Y., Chu, J., Citroen, M., Collymore, A., Cooke, P., Dawoe, T., Daza, R., Decktor, K., DeGray, S., Dhargay, N., Dooley, K., Dorje, P., Dorjee, K., Dorris, L., Duffey, N., Dupes, A., Egbiremolen, O., Elong, R., Falk, J., Farina, A., Faro, S., Ferguson, D., Ferreira, P., Fisher, S., FitzGerald, M., Foley, K., Foley, C., Franke, A., Friedrich, D., Gage, D., Garber, M., Gearin, G., Giannoukos, G., Goode, T., Goyette, A., Graham, J., Grandbois, E., Gyaltsen, K., Hafez, N., Hagopian, D., Hagos, B., Hall, J., Healy, C., Hegarty, R., Honan, T., Horn, A., Houde, N., Hughes, L., Hunnicutt, L., Husby, M., Jester, B., Jones, C., Kamat, A., Kanga, B., Kells, C., Khazanovich, D., Kieu, A.C., Kisner, P., Kumar, M., Lance, K., Landers, T., Lara, M., Lee, W., Leger, J.P., Lennon, N., Leuper, L., LeVine, S., Liu, J., Liu, X., Lokyitsang, Y., Lokyitsang, T., Lui, A., Macdonald, J., Major, J., Marabella, R., Maru, K., Matthews, C., McDonough, S., Mehta, T., Meldrim, J., Melnikov, A., Meneus, L., Mihalev, A., Mihova, T., Miller, K., Mittelman, R., Mlenga, V., Mulrain, L., Munson, G., Navidi, A., Naylor, J., Nguyen, T., Nguyen, N., Nguyen, C., Nicol, R., Norbu, N.,

Norbu, C., Novod, N., Nyima, T., Olandt, P., O'Neill, B., O'Neill, K., Osman, S., Oyono, L., Patti, C., Perrin, D., Phunkhang, P., Pierre, F., Priest, M., Rachupka, A., Raghuraman, S., Rameau, R., Ray, V., Raymond, C., Rege, F., Rise, C., Rogers, J., Rogov, P., Sahalie, J., Settipalli, S., Sharpe, T., Shea, T., Sheehan, M., Sherpa, N., Shi, J., Shih, D., Sloan, J., Smith, C., Sparrow, T., Stalker, J., Stange-Thomann, N., Stavropoulos, S., Stone, C., Stone, S., Sykes, S., Tchuinga, P., Tenzing, P., Tesfaye, S., Thoulutsang, D., Thoulutsang, Y., Topham, K., Topping, I., Tsamla, T., Vassiliev, H., Venkataraman, V., Vo, A., Wangchuk, T., Wangdi, T., Weiand, M., Wilkinson, J., Wilson, A., Yadav, S., Yang, S., Yang, X., Young, G., Yu, Q., Zainoun, J., Zembek, L., Zimmer, A. & Lander, E.S. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438(7069), 803-19.

Louis, D.N. (2006). Molecular pathology of malignant gliomas. *Annual review of pathology* 1, 97-117.

Margulies, E.H., Blanchette, M., Haussler, D. & Green, E.D. (2003). Identification and characterization of multi-species conserved sequences. *Genome research* 13(12), 2507-18.

Mariani, S.M. (2004). Genes and autoimmune diseases - a complex inheritance. *MedGenMed : Medscape general medicine* 6(4), 18.

Merveille, A.-C., Davis, E.E., Becker-Heck, A., Legendre, M., Amirav, I., Bataille, G., Belmont, J., Beydon, N., Billen, F., Clement, A., Clercx, C., Coste, A., Crosbie, R., de Blic, J., Deleuze, S., Duquesnoy, P., Escalier, D., Escudier, E., Fliegauf, M., Horvath, J., Hill, K., Jorissen, M., Just, J., Kispert, A., Lathrop, M., Loges, N.T., Marthin, J.K., Momozawa, Y., Montantin, G., Nielsen, K.G., Olbrich, H., Papon, J.-F., Rayet, I., Roger, G., Schmidts, M., Tenreiro, H., Towbin, J.A., Zelenika, D., Zentgraf, H., Georges, M., Lequarre, A.-S., Katsanis, N., Omran, H. & Amselem, S. (2011). CCDC39 is required for assembly of inner dynein arms and the dynein regulatory complex and for normal ciliary motility in humans and dogs. *Nature genetics* 43(1), 72-78.

Ostrander, E.A. & Kruglyak, L. (2000). Unleashing the canine genome. *Genome research* 10(9), 1271-4.

Ostrander, E.A. & Wayne, R.K. (2005). The canine genome. *Genome research* 15(12), 1706-16.

Ott, J., Kamatani, Y. & Lathrop, M. (2011). Family-based designs for genome-wide association studies. *Nature reviews. Genetics* 12(7), 465-74.

Parker, H.G., VonHoldt, B.M., Quignon, P., Margulies, E.H., Shao, S., Mosher, D.S., Spady, T.C., Elkahloun, A., Cargill, M., Jones, P.G., Maslen, C.L., Acland, G.M., Sutter, N.B., Kuroki, K., Bustamante, C.D., Wayne, R.K. & Ostrander, E.A. (2009). An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science* 325(5943), 995-8.

Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. (2008). Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome research* 18(11), 1814-28.

Patterson, D.F. (2000). Companion animal medicine in the age of medical genetics. *Journal of veterinary internal medicine / American College of Veterinary Internal Medicine* 14(1), 1-9.

Patterson, D.F., Haskins, M.E., Jezyk, P.F., Giger, U., Meyers-Wallen, V.N., Aguirre, G., Fyfe, J.C. & Wolfe, J.H. (1988). Research on genetic diseases: reciprocal benefits to animals and man. *Journal of the American Veterinary Medical Association* 193(9), 1131-44.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. & Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 81(3), 559-75.

Ramsden, C.A., Bankier, A., Brown, T.J., Cowen, P.S., Frost, G.I., McCallum, D.D., Studdert, V.P. & Fraser, J.R. (2000). A new disorder of hyaluronan metabolism associated with generalized folding and thickening of the skin. *The Journal of pediatrics* 136(1), 62-8.

Redman, J. (2002). Steroid-responsive meningitis-arteritis in the Nova Scotia duck tolling retriever. *The Veterinary record* 151(23), 712.

Rivas, A.L., Tintle, L., Kimball, E.S., Scarlett, J. & Quimby, F.W. (1992). A canine febrile disorder associated with elevated interleukin-6. *Clinical immunology and immunopathology* 64(1), 36-45.

Salmon Hillbertz, N.H., Isaksson, M., Karlsson, E.K., Hellmen, E., Pielberg, G.R., Savolainen, P., Wade, C.M., von Euler, H., Gustafson, U., Hedhammar, A., Nilsson, M., Lindblad-Toh, K., Andersson, L. & Andersson, G. (2007). Duplication of FGF3, FGF4, FGF19 and ORAOV1 causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. *Nature genetics* 39(11), 1318-20.

Savolainen, P., Zhang, Y.P., Luo, J., Lundeberg, J. & Leitner, T. (2002). Genetic evidence for an East Asian origin of domestic dogs. *Science* 298(5598), 1610-3.

Schadt, E.E., Turner, S. & Kasarskis, A. (2010). A window into third-generation sequencing. *Human molecular genetics* 19(R2), R227-40.

Serfling, E., Klein-Hessling, S., Palmetshofer, A., Bopp, T., Stassen, M. & Schmitt, E. (2006). NFAT transcription factors in control of peripheral T cell tolerance. *European journal of immunology* 36(11), 2837-43.

Shendure, J. & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology* 26(10), 1135-45.

Shendure, J., Mitra, R.D., Varma, C. & Church, G.M. (2004). Advanced sequencing technologies: methods and goals. *Nature reviews. Genetics* 5(5), 335-44.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., Weinstock, G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W. & Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* 15(8), 1034-50.

Stern, R., Asari, A.A. & Sugahara, K.N. (2006). Hyaluronan fragments: an information-rich system. *European journal of cell biology* 85(8), 699-715.

Stojanov, S. & Kastner, D.L. (2005). Familial autoinflammatory diseases: genetics, pathogenesis and treatment. *Current opinion in rheumatology* 17(5), 586-99.

Strang, A. & MacMillan, G. (1996). *The Nova Scotia Duck Tolling Retriever*. Loveland, Colorado, USA: Alpine Publications.

Stuehler, B., Reichert, J., Stremmel, W. & Schaefer, M. (2004). Analysis of the human homologue of the canine copper toxicosis gene MURR1 in Wilson disease patients. *Journal of molecular medicine* 82(9), 629-34.

Sutter, N.B. & Ostrander, E.A. (2004). Dog star rising: the canine genetic system. *Nature reviews. Genetics* 5(12), 900-10.

Tan, E.M., Cohen, A.S., Fries, J.F., Masi, A.T., McShane, D.J., Rothfield, N.F., Schaller, J.G., Talal, N. & Winchester, R.J. (1982). The 1982 revised criteria for the classification of systemic lupus erythematosus. *Arthritis and rheumatism* 25(11), 1271-7.

Truvé, K., Eriksson, O., Norling, M., Wilbe, M., Mauceli, E., Lindblad-Toh, K. & Bongcam-Rudloff, E. (2011). SEQscoring: a tool to facilitate the interpretation of data generated with next generation sequencing technologies. *EMBnet journal* 17.1, 38-45.

van De Sluis, B., Rothuizen, J., Pearson, P.L., van Oost, B.A. & Wijmenga, C. (2002). Identification of a new copper metabolism gene by positional cloning in a purebred dog population. *Human molecular genetics* 11(2), 165-73.

Vaysse, A., Ratnakumar, A., Derrien, T., Axelsson, E., Rosengren Pielberg, G., Sigurdsson, S., Fall, T., Seppala, E.H., Hansen, M.S., Lawley, C.T., Karlsson, E.K., Bannasch, D., Vila, C., Lohi, H., Galibert, F., Fredholm, M., Haggstrom, J., Hedhammar, A., Andre, C., Lindblad-Toh, K., Hitte, C. & Webster, M.T. (2011). Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. *PLoS genetics* 7(10), e1002316.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C.,

Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. & Zhu, X. (2001). The sequence of the human genome. *Science* 291(5507), 1304-51.

Vila, C., Savolainen, P., Maldonado, J.E., Amorim, I.R., Rice, J.E., Honeycutt, R.L., Crandall, K.A., Lundeberg, J. & Wayne, R.K. (1997). Multiple and ancient origins of the domestic dog. *Science* 276(5319), 1687-9.

Voss, R.H. (1933). The Evolution of the Bulldog. In: *Our dogs*. pp. 810-811.

Weigel, P.H., Hascall, V.C. & Tammi, M. (1997). Hyaluronan synthases. *The Journal of biological chemistry* 272(22), 13997-4000.

Wilcox, B. & Walkowicz, C. (1995). *The atlas of dog breeds of the world*. 5th . ed. Neptune city: NJ:TFH publications.

Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K., Walter, K., Abnizova, I., Gilks, W., Edwards, Y.J., Cooke, J.E. & Elgar, G. (2005). Highly conserved non-coding sequences are associated with vertebrate development. *PLoS biology* 3(1), e7.

Zanna, G., Fondevila, D., Bardagi, M., Docampo, M.J., Bassols, A. & Ferrer, L. (2008). Cutaneous mucinosis in shar-pei dogs is due to hyaluronic acid deposition and is associated with high levels of hyaluronic acid in serum. *Veterinary dermatology* 19(5), 314-8.

9 Acknowledgements

Erik Bongcam Rudloff for choosing me as a PhD student and giving me the chance to solve genetic mysteries. Thank you for being encouraging, enthusiastic, and positive minded. Especially thanks for giving me the opportunity to teach in Kenya, which was a great experience.

Kerstin-Lindblad Toh for giving me the opportunity to take part in the most exciting projects. Thank you for introducing me to many collaborators and for all the time you spent on discussing my projects and to help me improve my work.

Göran Andersson, your passion for science and great knowledge in molecular biology has inspired me. Thanks for interesting discussions and valuable advise, and especially thanks for your time spent on proofreading this thesis.

Leif Andersson, it was great to have such an experienced scientist as you involved in the Shar-Pei project. You are an inspiration, having found the cause for so many genetic traits, making hard things seem within reach.

Cecilia Johansson for being a great coordinator of the dog group. I'm very thankful for all help with planning of meetings, and for your good ideas bringing my projects forward.

Åke Hedhammar, your dedication and great knowledge about dogs and their diseases is motivating and inspiring. You really care about both dogs and people. I am glad for your involvement in both the Shar-Pei and the brachycephaly/glioma project.

All participants in the dog group for invaluable sharing of experience, knowledge, and ideas or just nice chats :-). Special thanks to: Erik Axelsson, Izabella Baranowska, Jonas Berglund, Tomas Bergström, Susanne Björnerfeldt, Susanne Gustafsson, Lotta Lantz, Jennifer Meadows, Mia Olsson,

Abhi Ratnakumar, Gerli Rosengren Pielberg, Katarina Sundberg, Katarina Tengvall, Matt Webster, Maria Wilbe.
(Note this is not an author-list, but in alphabetical order ;-))

Past and present participants in the dog group at BROAD institute: Claire Wade, Michele Perlsoki, Ross Swofford, Mike Zody, Evan Mauceli, thanks for all help, advise and quick answers to any questions. Special thanks to: Elinor Karlsson for introducing me to the world of genome wide association mapping in dogs. Your help during my very first week as a PhD student was of the greatest value.

Ulla Gustafson and Eva Murén, thank you for guiding me in the lab. Your help has been of great value for me.

Danika Bannasch and Pete Dickinson for entrusting me with data from your glioma samples. Thanks for good collaboration.

Thanks to all other co-authors of the papers included in this thesis, not already mentioned.

Marie Ekerljung du är en underbar vän! Tack för allt stöd och alla trevliga pratstunder. Tack igen för den fina målningen som jag fick i förskott…nu får jag nog behålla den.

Hans-Henrik Fuxelius, Oscar Eriksson, och Martin Norling för att det har varit ett nöje att dela rum med er, och för att ni vid flera tillfällen ryckt in och hjälpt till i nödens stund som vid kraschad hårddisk, stulen plånbok, med flera stressande situationer.

Min svärmor Kerstin Truvé för all din ovärderliga hjälp hemma med barn och hundar. Du har hjälpt oss så mycket!

Mina föräldrar för att ni till slut efter mycket bönande och eget sparande lät mig och min tvillingsyster Agneta köpa vår första hund. Det måste ha varit där banan till hundgenetiker började :-). Tack för att ni funnits där under doktorandtiden, beredda att stötta och uppmuntra.

Min man Staffan Truvé för allt stöd och all hjälp, utan dig skulle jag aldrig bli doktor. For those of you who would have preferred another title of this work Staffan proposed “*Sick as a dog*” :-)

Mina barn Viktor, William och Theodor för att ni finns, och påminner mig om vad som är viktigt. Ni har haft en mamma som veckopendlat mellan Alingsås och Uppsala, med vad det innebär. Hoppas att det snart ska finnas mer tid för roliga stunder tillsammans!

Finally an acknowledgement to all wonderful dogs, and especially to my furry friends Buddy and Bamse.

"All knowledge,
the totality of all questions and all answers,
is contained in the dog."

Franz Kafka