UNIVERSITEIT
GENT

FACULTY OF PHARMACEUTICAL SCIENCES

Canine whole exome sequencing:

hip, hip, hurray?

The quest for genetic solutions for phenotypical problems

Bart Broeckx
Promotor: Prof. Dr. Apr. Dieter Deforce
Co-promotor: Prof. Dr. Drs. Frank Coopman
Co-promotor: Prof. Dr. Drs. Geert Verhoeven
Co-promotor: Prof. Dr. Apr. Filip Van Nieuwerburgh

Dissertation submitted in fulfillment of the requirements for the degree
of Doctor in Pharmaceutical Sciences
Academic year: 2015 - 2016

Members of the reading committee

Prof. Dieter Deforce （promotor， Ghent University）

Prof. Tom Coenye （chairman， Ghent University）

Dr. Wouter Coppieters （Université de Liège）

Prof. Paul Coucke （Ghent University）

Prof. Luc Duchateau （Ghent University）

Member of the examination committee

Prof. Frank Coopman （co-promotor， Ghent University）

Prof. Tim De Meyer （Ghent University）

Prof. Wim Van Hul （University of Antwerp）

Dr. Wim Van Haeringen （Dr. Van Haeringen Laboratorium）

List of abbreviations

ARSG	Arylsulfatase G
BVA/KC	British Veterinary Association/Kennel Club
CDS	coding DNA sequences
ddNTP	dideoxyribonucleotide triphosphate
DJD	Degenerative joint disease
DM	Degenerative myelopathy
DNM1	dynamin 1
dNTP	deoxyribonucleotide triphosphate
EIC	Exercise-induced collapse
FBN2	Fibrillin 2
FCI	Fédération Cynologique Internationale
FDR	False discovery rate
FRET	Fluorescence resonance energy transfer
FWER	Family-wise error rate
GM1	Gangliosidosis
GRMD	Golden Retriever muscular dystrophy
GWAS	genome-wide association study
HD	Hip dysplasia

Abbreviations

HMLR	centronuclear myopathy
HWE	Hardy-Weinberg equilibrium
KASP	Kompetitive Allele Specific PCR
LD	Linkage disequilibrium
LE	Linkage equilibrium
MAF	minor allele frequency
MG	myotonia congenita
MM	mismatch
MPS VII	mucopolysaccharidosis VII
mRNA	messenger RNA
NCISD	National Committee for Inherited Skeletal Disorders
NCL	Neuronal ceroid lipofuscinosis 4A
NGS	Next generation sequencing
OFA	Orthopedic Foundation for Animals
OMIA	Online Mendelian Inheritance in Animals
PM	Perfect match
PRA	progressive retinal atrophy
SNP	single nucleotide polymorphism
SOD1	Superoxide dismutase 1
UCSC	University of California Santa Cruz

Abbreviations

UTR	Untranslated region
VCF	variant call format
VD	ventrodorsal
WES	Whole exome sequencing
WGS	Whole genome sequencing
%GC	GC content per region

“If I have seen further, it is by standing on the shoulders of giants”

Preface

Although there might be some discussion to whom we can attribute this quote, we cannot argue that (big and small) advances in science are a consequence of the work of our predecessors. It seems no more than right to briefly sketch some of the breakthroughs in the field of genetics that made the research presented here, possible. One of these giants in genetics is, undoubtedly, Gregor Mendel, a scientist and monk that lived in the 19th century and that is currently considered to be the father of genetics. In his experiments on peas, he discovered that phenotypes (“how does it look?”) are transmitted in a predictable manner. In the 20th century, numerous discoveries revolutionized the field. In 1953, Watson, Crick, Wilkins and Franklin (the last one might be one of the most overlooked scientists) discovered that deoxyribonucleic acid (DNA) is organized in a double helix. Frederick Sanger proposed in 1977 a method that enables the accurate determination of the sequence of nucleotides (= the building blocks) of the DNA. This method is now known as “Sanger” sequencing. A downside at this point was the need for large amounts of starting material. This was resolved by Mullis in 1987 by the development of the polymerase chain reaction (PCR). This technique enables *in vitro* exponential amplification of specific DNA pieces (= DNA templates) and with some modern modifications, this is still one of the most widely used standard procedures in the lab.

These and other techniques were all necessary for what might be one of the milestones of the 21st century: the publication of the first draft of the human genome. Since then, the pace is ever increasing: newer sequencing technologies enable an enormous throughput, faster computing algorithms are developed and the genome of several thousands of organisms has been analysed.

Looking for the needle(s) in the haystack

1 Introduction

The dog, or *Canis familiaris*, has been accompanying us, humans, for several thousands of years and has a remarkable phenotypic diversity in terms of colour, size, skull shape, etc[1,2] (Figure 1.1). This is a consequence of the selective breeding of dogs with specific qualities and has resulted in well over 300 different breeds[3]. In modern times, the worldwide breeding of pedigree dogs is regulated by Kennel Club guidelines in order to meet the unique characteristics for each breed. To achieve this, intense inbreeding practices and a small number of popular sires have often been used[4]. It is generally accepted that, as a consequence of these breeding practices, the general health of pedigree dogs has been compromised compared to cross-breed dogs[4]. These concerns about the health of our pedigree dogs have also been raised in the documentary "Pedigree Dogs exposed", broadcasted by the BBC in 2008[5,6].

The Kennel Club's breeding standards are mainly focused on conformational characteristics rather than health characteristics. The hair ridge in the Rhodesian Ridgeback is associated with dermoid sinus. Breeding ridgeless dogs could eliminate dermoid sinuses[7]. Other examples are brachycephalic airway obstructive syndrome in the brachycephalic dog breeds or skin fold dermatitis in breeds with excessive skin folds[5]. Based on these examples, it is clear that some disorders are directly associated with these breeding guidelines. However, as other genetic disorders are also highly prevalent in the dog, health issues cannot be contributed to conformational guidelines alone[6,8].

In order to improve our dog's health, several propositions have been made[9]. This entails more than the mere exclusion of all dogs that carry a deleterious allele as every dog, just like humans, carries some disease-associated variants[4]. A good strategy starts with getting an overview of which disorders are important for which breed. After deciding which disorders to tackle first, proper screening tools have to be developed and implemented in breeding strategies. For heritable disorders, the best screening tools are DNA tests as they are not influenced by environmental variation. Continuous follow-up of the achieved breeding progress is required. Simultaneously, awareness regarding emerging disorders is important to guarantee a positive evolution.

According to the public Online Mendelian Inheritance in Animals (OMIA) database[10], a total of 653 genetic disorders are known in the dog. Two hundred and sixty three are Mendelian traits and for 193 of them, the causal variant is known. This implies that for the large majority of disorders, the causal variant(s) still need to be elucidated. In this chapter, the history and current state of the canine genome is discussed. Next, possible tools that can be used to link mutations to phenotypes, are described.

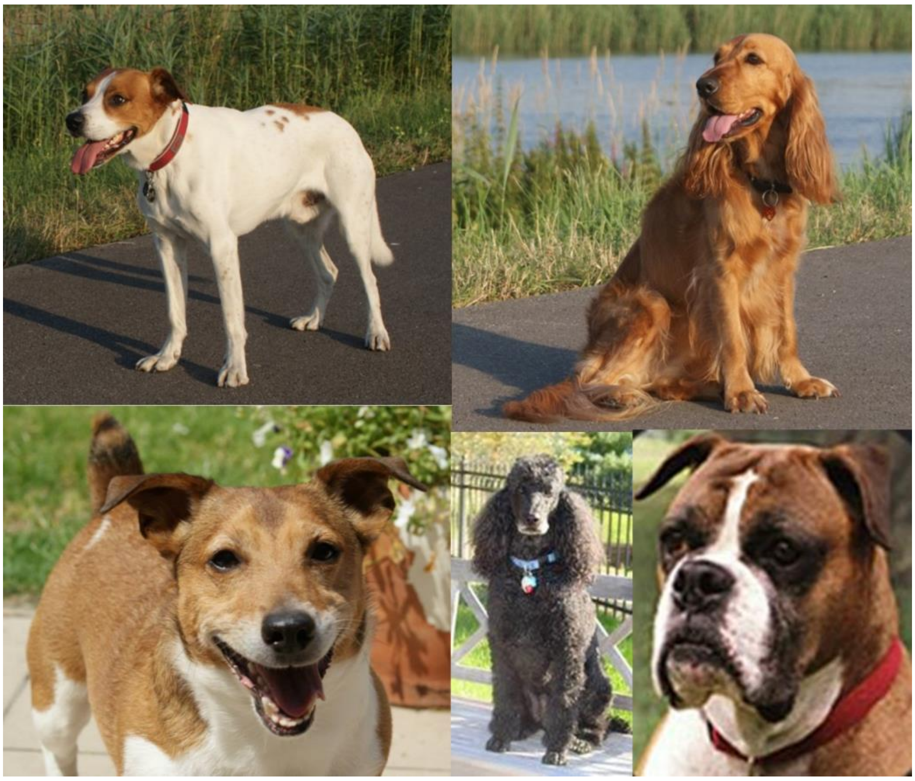


Figure 1.1. Examples of the phenotypic variation in the domesticated dog[11,12].

（From left to right: top row: crossbreed, Cocker Spaniel; bottom: Jack Russell Terrier, Poodle, Boxer.）

1.1 The dog and its genome

A genome is defined as the organism's complete set of DNA[13]. The dog genome consists of 39 pairs of chromosomes （38 pairs of autosomes and the X and Y sex chromosomes）[14] （Figure 1.2）. Only two years after the first draft of the human genome, the group of Venter published the first assembly of the canine genome （with a 1.5x coverage）[15]. The supplier of the DNA was Shadow, the poodle of Venter himself (Figure 1.1). This was followed by an updated version with a 7.5x coverage by sequencing Tasha, a female Boxer (Figure 1.1)[2]. By comparing these two genomes and low-pass shotgun sequencing of several dog breeds, wolves and a coyote, the first dense single nucleotide polymorphism (SNP) map was published at the same time[2]. In 2014, the present assembly (CanFam 3.1) was published, with reduced gaps and an improved annotation[16].

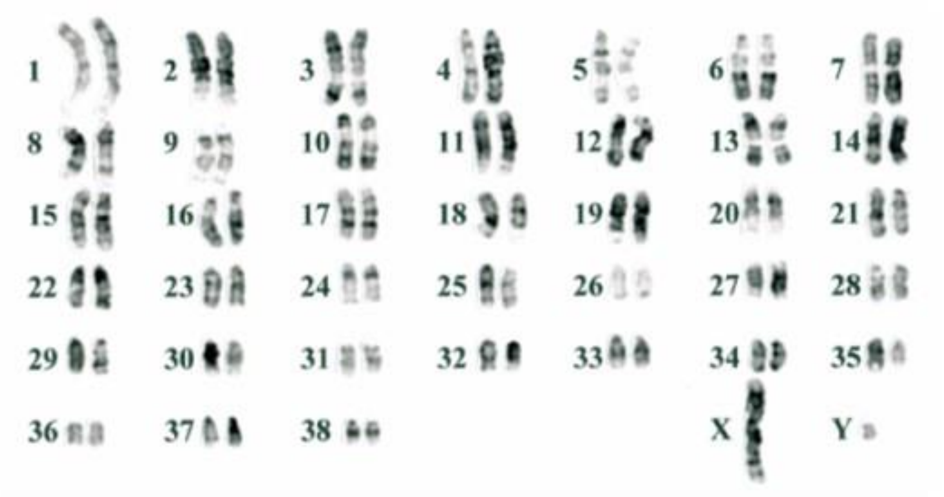


Figure 1.2. Karyogram of the dog[17].

In the process of domestication and breed creation, the dog population went through several bottlenecks[2,18,19]. Compared with the human genome, this resulted in large haplotype blocks within a breed and short blocks between breeds[2,20–22]. A haplotype is a set of DNA variants that tend to be inherited together[13]. In addition, the haplotype diversity within a breed is limited as well[21]. The length of the haplotype blocks varies between breeds and is in agreement with the population history of that breed[21,22].

These characteristics are favorable for genome-wide association studies (see 1.3.1).

1.2 Unravelling the link between a phenotype and a genotype

As already empirically found by Mendel in the 19th century: the way you look is at least partially heritable and linked to your DNA. With the discovery of several mutations responsible for all kinds of phenotypes, it has become feasible to predict the result of certain combinations by a priori performing some genetic tests. Aside from phenotypes, such as coat colour, the genetic cause for several diseases has been discovered as well[10,23]. The question is: how can you link a certain genetic mutation to a certain phenotype? In this era of fast advances in molecular biotechnology, several approaches have been developed that enable us to identify this link.

1.3 Theoretical Background

Simplified, phenotypes and disorders can be divided into two distinct groups[24-26]. The first group contains the rare Mendelian, monogenic or simple phenotypes, where the phenotype is caused by one genetic mutation or mutations in one gene and the inheritance pattern is dominant, recessive or sex-chromosome linked. In contrast, the complex or multifactorial traits are caused by a combination of genetic and environmental factors. Typically, these traits are genetically heterogeneous, making the identification of disease-causing genes very challenging. This heterogeneity is reflected in the genetic part of these common traits being the sum of a combination of relatively common variants and relatively rare variants, where the latter will increase the risk more than the former[27]. The difference between common and rare variants is important as it has its consequences in the process of linking them to a phenotype.

Some phenotypes develop through the action of newly formed mutations[25,26]. These so called *de novo* mutations (and the disease) are not present in the healthy parents, as they originate in the meiosis of the germline. However, they can be transmitted to the offspring and cause disease in these individuals. These mutations will be treated separately.

Overall, the methods that link causal genetic variation to a phenotype can be divided into two groups: the direct and the indirect methods. The direct methods try to identify the disease-causing mutations as such. Indirect methods use genetic markers nearby the actual disease-causing variant to identify the link with the phenotype (Figure 1.3). Examples of indirect methods are linkage analysis and genome-wide association studies (GWAS). Direct methods are whole genome sequencing (WGS) and whole exome sequencing (WES). Depending on the approach, candidate gene studies can be direct or indirect. In this introduction, we will focus on GWAS, candidate gene studies and WES. Due to the cost, WGS studies are currently not often performed. Due to the relatively low resolution, the need for large pedigrees and the advances in biotechnology, linkage analysis has been largely replaced by GWAS and will not be discussed in this introduction.

Irrespective of the method used (WGS, WES, GWAS or candidate gene studies), the power of a study will be dramatically influenced by misclassification of cases and controls[28]. Therefore, it is of utmost importance that the appropriate diagnostic tools and criteria are used.

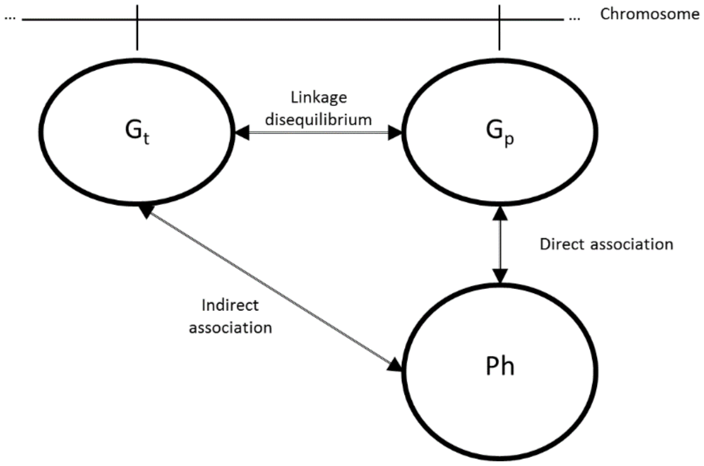


Figure 1.3. Direct (G_p) and indirect (G_t) association from genetic variation with a phenotype (Ph). G_t is a so-called tagSNP, whereas G_p represents causal variation.

1.3.1 GWAS

GWAS are conducted in several steps. The starting point is choosing an appropriate study design. Case-control designs are used most often, other designs are cohort studies and trio studies (i.e. an affected case with two unaffected parents). After genotyping these individuals with a GWAS array of choice, several checks can be performed. Overall, a sufficient number of SNPs should be called reliably in each sample, violations of Hardy-Weinberg equilibrium can be tested, only SNPs are retained that occur more frequently than a certain minor allele frequency

(MAF) and if duplicate samples are included, their SNP calls should be highly concordant[18,28,29]. Following these checks, a cleaned list of several thousands of SNP calls remains for the final association testing. This association testing is in essence no more than a comparison of the frequencies of alleles or genotypes in cases and controls and several statistical tests can be used (e.g. Fisher exact test, χ^2, logistic or linear regression). But what is the underlying principle of a GWAS?

A graphical representation of the principle of GWAS is presented in Figure 1.3. Essentially, it comes down to identifying a link between a SNP and a phenotype. Although it is unlikely, it might be that the SNP used in the GWAS, is directly causing the phenotype ($G_t = G_p$)[30]. This will seldom be the case. It is more likely that the link exists because a phenotype is associated with a causal genetic variant and, in turn, the causal genetic variant is associated with a tagSNP[31]. For a GWAS, this association between the tagSNP and the causal genetic variant is due to linkage disequilibrium (LD). LD is defined as the non-random association of alleles at two or more loci[32]. It is possible that LD exists between loci on different chromosomes, but it is usually defined in terms of loci on the same chromosome now[32]. In that sense, LD and linkage are "linked": LD is generally stronger when loci are linked more closely together[33]. Several formulas exist to quantify the LD between two loci, but the standard formula for two loci with each two alleles, is the following[31,32,34]:

(1) $D_{AB} = p_{AB} - p_A \cdot p_B$

With:

p_{AB} = the frequency of the AB haplotype

p_A = the frequency of allele A (at the first locus)

p_B = the frequency of allele B (at the second locus)

If D_{AB} is equal to zero, this is called linkage equilibrium (LE). If it ≠ 0, LD exists.

The amount of LD can also be expressed relative to its maximum possible value, given the allele frequencies[31,32,34]:

(2) $D' = D / \min\{p_A . (1 - p_B), p_B .(1 - p_A)\}$ if $D > 0$
$D' = D / \min\{p_A . p_B, (1 - p_A). (1 - p_B)\}$ if $D < 0$

To make a GWAS possible, three items are needed:

- a list of SNPs with their genomic location;
- LD has to be present between at least one SNP and the causal mutation;
- an easy and reliable method to genotype a large amount of SNPs.

The first prerequisite, a list of SNPs, became publicly available together with the second canine genome assembly[2]. The second prerequisite implies that disease-causing alleles in two individuals should be identical-by-descent and that the tagSNPs have to be close enough so that little recombination took place[35]. The gradual decay of LD can be quantified with the following formula[32]:

(3) $D_{AB}(t+1) = (1-c) . D_{AB}(t)$

With:

t = the time in generations

c = the recombination frequency

D_{AB} = the LD for haplotype AB

It is clear that after a sufficient number of generations, LE will be eventually reached, but at a slow rate for closely linked SNPs[32]. Due to the advances in biotechnology, the third prerequisite was fulfilled as well: at this point, four GWAS arrays are available for the dog[18,19,36,37]. Details on each array are discussed in section 1.4.2.

An important issue is the number of tagSNPs that need to be interrogated to ensure sufficient coverage of the genome. The aim is essentially to capture as much genetic variation as possible, with a minimum number of SNPs[35]. The number of SNPs needed, depends on the amount of LD. For the selection process, LD is usually quantified in terms of a correlation coefficient, using the following formula[32,34,38]:

$$(4) \quad r^2 = \frac{\mathrm{D}^2}{\mathrm{p}_A\,(1-\mathrm{p}_A)\,\mathrm{p}_B\,(1-\mathrm{p}_B)}$$

The idea is that, if several SNPs are highly correlated, it is not necessary to genotype them all, as genotyping one of them, gives a fairly good knowledge of the genotype of the ungenotyped SNPs as well[30]. So one has to choose the SNP(s) that reflect the LD landscape for a certain location in the genome.

The amount of LD varies between organisms and reflects the history of the population. Population bottlenecks, inbreeding and the use of popular sires can increase the LD and thus reduce the number of tagSNPs necessary to localize the region that harbors the disease-causing mutation[1,2,18,21]. This explains why LD is more extensive within a breed compared with the LD in the human genome[1,2,18,21,22]. The advantage is that, theoretically, far less SNPs would be needed to map a trait. However, using less SNPs, the putative region containing the causal variant will be larger as well, requiring more work further downstream in the analysis. As the LD between canine breeds is far less extensive, an option would be to do a two-stage mapping[18]. The first step is a GWAS within a breed, followed by a fine-mapping stage in more breeds. An important point is that this will only work if the same ancestral haplotype is shared between breeds[18].

An important issue when performing several thousands of statistical tests (called "multiple testing"), is the inflation of the type I-error, also known as the false positives. Assuming an α-threshold of 0.05 (which implies a chance of 5% of seeing at least the same result when H_0 is true) for one test, will result in ± 5 000 significant results occurring by chance alone when 100 000 tests (= 100 000 SNPs in this case) are performed. To correct this, several methods have been suggested. The first group of corrections, controls the family-wise error rate (FWER)[39]. As the name suggests, it controls the total number of false positives, for the entire "family" of tests. Mathematically, this is represented as $P(H_0 = \text{rejected} \mid H_0 = \text{true})$. One of the most known and used procedures, is the Bonferroni correction:

(5) $\alpha_t = \frac{\alpha_0}{n}$

With:

α_t = the α for each individual test

α_o = the overall α

n = the number of tests

So by increasing the stringency for each individual test, the overall type I error rate is controlled at a certain level. A downside of the Bonferroni correction, is that when a lot of tests are performed, the p-value for individual tests has to be extremely (unrealistically?) low. Thus, applying this correction affects the statistical power. A modified Bonferroni correction, called the Holm-Bonferroni correction, uses a stepwise correction to improve the power while maintaining a low error rate[40]. A second group of corrections, controls the False Discovery Rate (FDR)[39,41,42]. These methods give information on the probability that H_0 is true, given that H_0 was rejected ($P(H_0 = \text{true} \mid H_0 = \text{rejected})$). This explicitly tells you how many significant results are possibly incorrect, in contrast to the FWER. An example is the correction suggested by Benjamini and Hochberg[41]:

- Sort the *n* p-values in ascending orders; label these p_1, p_2, …, p_n
- Let *k* denote the largest index *i* for which $p_i \leq d \ x \ i/n$, for all *i*, with *d* the false discovery rate threshold of choice
- Declare all tests with p-values p_1 ,p_2, …, p_k significant

Permutations can also be used to estimate the FDR, as under H_0, the case-control labels can be permutated to determine the number of false positives[42,43].

A downside of GWAS is that they are sensitive for population stratification: it is difficult to distinguish whether a significantly associated SNP represents a true association or whether it is caused by differing allelic frequencies in subpopulations[22,28-30]. Although remedial methods have been developed to correct for stratification, careful selection of cases and controls should always be the first priority[44,45].

In the next three sections, the application of GWAS is discussed for Mendelian and common diseases and diseases caused by *de novo* mutations.

A. Rare Mendelian disorders

Two properties of genetic diseases are very important. The first one is detectance (i.e. P(genotype I phenotype)). If the detectance is 100%, there is no genetic heterogeneity and phenocopies are absent. A phenocopy is "an individual without the trait mutation that nonetheless has the trait due to environmental or other causes" [19]. The second feature is penetrance (i.e. P(phenotype I genotype)). If penetrance is 100%, every genetically affected individual is actually sick. Diseases with 100% detectance and 100% penetrance are highly amenable for disease-association studies. Throughout this introduction, we apply Fisher exact tests on 2x2 contingency tables and plot the obtained p-values to demonstrate the effect of sample size and reduced detectance and/or penetrance in association testing. These results can be used directly under the assumption of a dominant model and a recessive model[29]. Alternatively, allele or genotype frequencies in cases and control can be compared directly, without specifying an inheritance model[29]. In addition, the effect of testing with and without a correction for multiple testing is demonstrated. To be on the safe side for our estimations, we applied the conservative Bonferroni correction in the demonstration. As the number of SNPs genotyped varies between arrays (see 1.4.2), the following α_t-thresholds were applied:

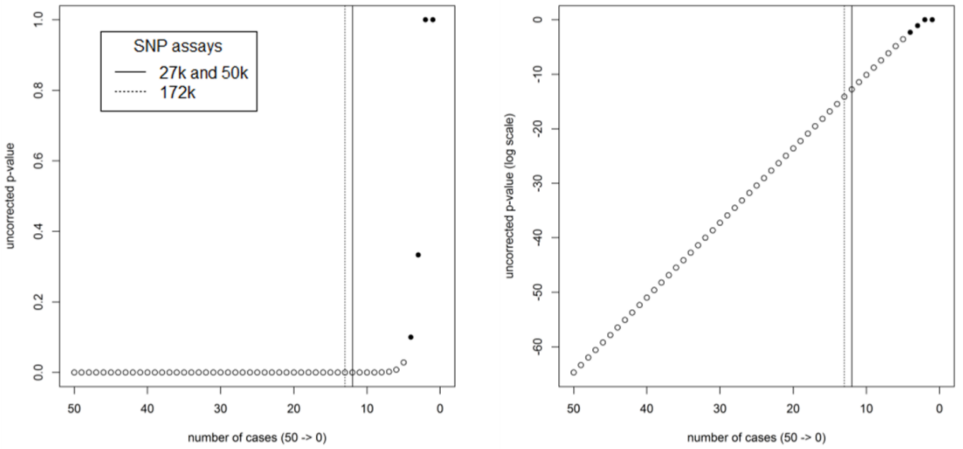
For the 27k array: $p \leq 1.851852 \ . \ 10^{-06}$

For the 50k array: $p \leq 1 \ . \ 10^{-06}$

For the 172k array: $p \leq 2.906977. \ 10^{-07}$

The effect of sample size was evaluated first. As demonstrated in Figure 1.4, significance can be reached quickly at very low sample sizes (total $n \geq 24$ or 26 (= sum of cases and controls)). Importantly, we assumed that the observed genotype/allele, is in perfect LD ($r^2 = 1$) with the causal mutation.

The situation is more difficult when the separation between cases and controls is not complete. Reduced penetrance, genetic heterogeneity, genotyping errors or phenotypical misclassification are quite common and can affect the results significantly. The effect of these factors is demonstrated in Figure 1.5 for a total sample size of 100. Although the total sample size remains constant and resulted in very low p-values in Figure 1.4, it is clear that depending on the distribution, significant values are not always reached. Phenotypical misclassification, genotyping error and a low LD can complicate results in every situation. Genetic heterogeneity and phenocopies will result in phenotypical cases with an incorrect genotype (see Figure 1.5, section A). Reduced penetrance results in individuals with an affected genotype but an incorrect phenotype (see Figure 1.5, section B).



		Genotype	
Phenotype	affected	healthy
Case	50 → 0	0
Control	0	50 → 0

Figure 1.4. Relation between the number of cases (x-axis), the uncorrected p-value (y-axis, left graph) and the uncorrected p-value in log scale (y-axis, right graph), using a Fisher-exact test. Results on the left hand side of the vertical lines indicate when the p-values become small enough to provide significant results when the Bonferroni correction is applied for the 27k, 50k and the 172k genome-wide association studies assays, respectively (α_0 = 0.05). Filled dots: uncorrected p-value > 0.05; open dots: uncorrected p-value ≤ 0.05.

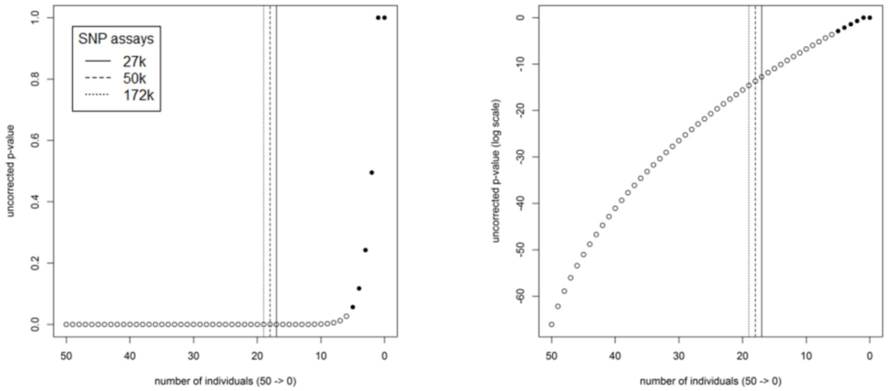
B. Common disorders

Common disorders are far more difficult to map compared to Mendelian disorders. At present, the theory is that common diseases are caused by common and rare variants[27]. A simplified representation of a situation that might occur, was already presented in Figure 1.5. It shows that rare variants (that, per definition, have a low allele frequency) will be difficult to detect, even with a total sample size of 100 individuals (Figure 1.5, part A). In addition, these variants typically add to the risk of developing disease and are not sufficient to cause disease themselves. This situation is shown in Figure 1.5, part B.

In Figure 1.6, the combined effect of these complicating factors that are typically associated with complex disorders is demonstrated again for a total sample size of 100. Significance is still reached but far less often. Based on these figures, it still seems possible to discover genetic variation linked to common disorders. However, in reality, the situation is even more complex. Common variants are not that common: depending on the source, they have been cited to occur at allelic frequencies of 1-5%, so to detect them, sample sizes have to be large ($n >> 1000$)[27,33]. For rare variants, the situation is even worse. Even though it is assumed that they contribute “more” to disease, their allele frequency is even lower. In addition, an important point of consideration relates to the link between LD, the allele frequency and the MAF-cutoff applied in the analysis steps. It has been demonstrated that as a prerequisite for a high

r^2 between the tagSNP and the untyped variant, the allelic frequencies of both have to be relatively similar[33,34].

Figure 1.5. Relation between the number of individuals (x-axis), the uncorrected p-value (y-axis, left graph) and the uncorrected p-value in log scale (y-axis, right graph), using a Fisher-exact test. Results on the left hand side of the vertical lines indicate when the p-values become small enough to provide significant results when the Bonferroni correction is applied for the 27k, 50k and the 172k genome-wide association studies assays, respectively (α_0 = 0.05). Filled dots: uncorrected p-value > 0.05; open dots: uncorrected p-value ≤ 0.05.



A

	Genotype	
Phenotype	affected	healthy
Case	50 → 0	0 → 50
Control	0	50

→

Within phenotypical cases, genotype changes

- Phenotypical misclassification
- Genotyping error
- r² (LD) low(er)
- Genetic heterogeneity
- Phenocopies

Example at extreme: rare risk allele (frequency affected genotype <<)

B

	Genotype	
Phenotype	affected	healthy
Case	50 → 0	0
Control	0 → 50	50

→

Within affected genotype, phenotype changes

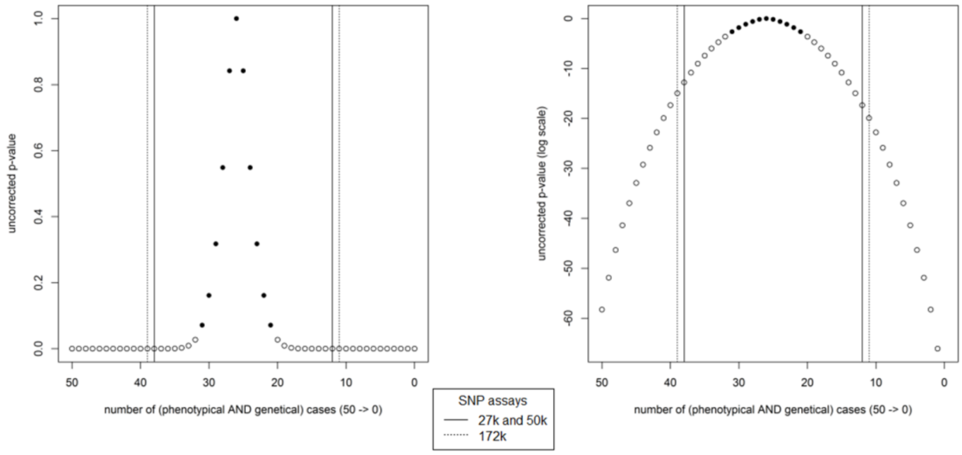
- Phenotypical misclassification
- Genotyping error
- r² (LD) low(er)
- Reduced penetrance

Example: low contribution to disease for common allele

For example, if the allele frequency of the tagSNP is 50% and if the r^2 has to be at least 0.8, the allelic frequency of the untagged variant has to be within ± 6%. The same situation, but with an allele frequency of the tagSNP of 5% (a frequently used cut-off for the MAF in the quality control in GWAS studies) requires the allelic frequency of the rare variant to be within 1%. A MAF of 4% for rare variants is simply too high, based on current criteria for human complex diseases[27].

It is not completely clear whether the applied allelic frequency thresholds for common and rare variants used in humans can be extrapolated directly to the canine population. Due to the population history of the dog, it is possible that the allele frequencies of rare variants are higher due to the reduced genetic diversity within breeds, but even if the allele frequency of rare variants would be sufficiently high, the sample size requirements would still be substantial[19]. Overall, it is very unlikely for rare variants associated with common disorders to be detected with GWAS.

Figure 1.6. Relation between the number of genetically and phenotypically affected individuals (x-axis), the uncorrected p-value (y-axis, left graph) and the uncorrected p-value in log scale (y-axis, right graph), using a Fisher-exact test. Results between the vertical lines indicate when the obtained p-values are higher than the Bonferroni correction threshold for the 27k, 50k and the 172k genome-wide association studies assays (α_0 = 0.05). Filled dots: uncorrected p-value > 0.05; open dots: uncorrected p-value ≤ 0.05.



	Genotype				Genotype	
Phenotype	affected	healthy	when	Phenotype	affected	healthy
Case	50 → 0	0 → 50	→	Case	≥ 39 or 38	≤ 11 or 12
Control	0 → 50	50 → 0	significant?	Control	≤ 11 or 12	≥ 39 or 38

C. *De novo* mutations

De novo mutations cannot be identified using GWAS. The reason is that GWAS relies on ancestral haplotypes being passed on, with the causal variant being in LD with the tagSNP. As *de novo* mutations arise in the germline, the haplotype is only created at that point. In addition, it is typically assumed that for *de novo* mutations, the reproductive fitness is affected, so the novel haplotype, created in the affected proband, will not be passed on to the following generations[26].

To conclude, GWAS has several downsides:

- it always requires further steps downstream, as it only identifies a region containing the causal variant;
- by design, it is very unlikely to discover rare variants in complex diseases and impossible to detect *de novo* mutations;
- it is sensitive for population stratification;
- not specific for GWAS, but a more general issue related to all indirect methods: they are always at best as efficient as direct methods, never better.

On the plus side:

- GWAS have been used successfully in disease-association studies for both Mendelian disorders and to identify common variants in complex disorders;
- it allows for an unbiased view of the entire genome and does not require prior assumptions in terms of biological knowledge (e.g. protein-coding mutations, regulatory mutations).

An example of a successful GWAS study in the dog is the identification of the causal mutation in the SOD1 gene (SOD1:c.118G>A) responsible for degenerative myelopathy in the dog, an autosomal recessive disorder with age-related incomplete penetrance[46].

1.3.2 Candidate gene studies

Depending on the applied method, the candidate gene methodology can be direct or indirect. The indirect methodology relies on the same principles as GWAS (indirect identification of causal variation due to LD with a tagSNP), the direct methodology relies on sequencing with the analysis being comparable to whole exome sequencing (see 1.3.3). Depending on the number and size of the candidate genes, the technical approach varies. Investigating a limited number of genes might be done with Sanger sequencing or individual SNP genotyping, whereas for larger projects, entire arrays or targeted enrichment designs might be more optimal. Irrespective of the indirect or direct methodology, the most important point is the appropriate selection of candidate genes[47]. Even though the selection might be perfectly sound, based on scientific knowledge at that time, it is not unlikely for a candidate gene study to fail. Often, when the causal mutation/gene is identified, it is a new gene that was never implicated before in the disease under investigation[37]. Success rates in candidate genes studies are thus rather low[48,49].

Overall, the downsides of both the indirect and direct candidate gene approach are the same as those for the GWAS and WES and are discussed in their respective sections. There are two additional disadvantages:

- the a priori assumptions that have to be made about the biological basis of the disease often turn out to be problematic;

- the wet lab tools developed for that project, can only be used for that specific project (e.g. primers).

On the plus side, it is likely the cheapest method, compared to GWAS and WES, although the price varies considerably with the project size.

An example of a successful candidate gene study, is the identification of a mutation in the *MC1R* gene that results in a truncated protein that is responsible for some of the coat colours seen in dogs[50].

1.3.3 Whole exome sequencing

The aim of WES is to selectively sequence all the exons throughout the genome. In contrast to GWAS, WES is a direct interrogation of the genome since the technique is based on sequencing. It is a fairly recent technique as it required several biotechnological breakthroughs. First of all, although the exome is much smaller compared with the genome (2 to 6% of the genome, depending on the design), it still comprises several millions of bases that need to be sequenced. It is possible to use Sanger sequencing to sequence even entire genomes. However, the development of second/next generation sequencing (also known as massively parallel sequencing) made it certainly more feasible[51]. A second issue is that most techniques relied on the polymerase chain reaction (PCR) to specifically target subsets of the genome[52]. This requires the design of large numbers of primers and numerous individual amplification reactions. Certainly for large resequencing projects, this is very time-consuming, error-prone and relatively expensive. The development of targeted enrichment techniques that allow reproducible resequencing of parts of the genome, fulfilled the second prerequisite to make WES feasible.

Selectively sequencing exons is less expensive compared to whole genome sequencing and most disease-causing variants have been observed to alter the amino-acid sequence for Mendelian disorders (= non-synonymous variants)[24,25]. In addition, most studies focus initially on non-synonymous variants as their effect is much easier to predict compared with synonymous or non-protein coding variants. A complicating factor for common disorders is the influence of regulatory mutations instead of protein-coding mutations alone[16]. These mutations will be missed in WES experiments.

The starting point of WES studies, is the choice of the "exome". The exact definition of the exome varies. For example, the ex- or inclusion of the 5' and 3' untranslated regions depends on the choice of the developer of WES enrichment designs and varies between commercial platforms[26,53]. With commercial platforms being able to target up to 200 Mb at this moment, the choice of which regions to include is nowadays less an issue of technical limitations but more of practical and theoretical considerations. It is important to consider that all designs are always based on the current knowledge of the annotation. Although the genome of several species has been thoroughly investigated with several techniques, it is likely that some of the coding regions are missed due to an incomplete annotation[26]. With new information becoming available, updates and extensions are thus required.

When variants have been called, they can subsequently be analyzed to assess their potential relationship with the phenotype under study. In general, the analysis methodologies can be divided into two major groups (that, however, are not mutually exclusive). The first group uses heuristic filters to filter sequencing variants[25]. The required sample size for these studies is typically low. Often sequencing only a couple of individuals (less than 10) is sufficient. In one of the first successful WES studies, only 4 individuals were sequenced[54]. Especially for Mendelian disorders and *de novo* mutations, heuristic filtering is highly amenable[24]. The second group of analysis uses a more statistical approach. This approach will be necessary when rare variants related to common disorders are studied. Although WES is a direct approach compared with GWAS, the same statistical tests can be used to identify putative causal variants.

A. Rare Mendelian disorders

The heuristic filtering approach is typically used to study rare Mendelian disorders. Often, these filters rely on several assumptions, as demonstrated in Figure 1.7. If, after filtering, only a few non-synonymous variants remain, the effect of every variant on its protein can be predicted with tools as PolyPhen or Provean to prioritize them further[25,55,56].

Often, one of the first steps in filtering consists of the removal of previously identified variants present in public databases such as dbSNP. This significantly reduces the number of putative variants. Depending on the disease studied, one can choose to use all the variants present in a database or to use only those variants that have a certain MAF. As the number of variants in these databases increases constantly, is it not unlikely for a database to become "contaminated" i.e. contain disease-causing variants[25]. Therefore, specifying a MAF might be a safer option.

B. Common disorders

As discussed in GWAS (section 1.3.1), several complicating factors arise when common disorders are studied. The big benefit of direct sequencing-based approaches is the direct interrogation of individual base pairs and not having to rely on LD (see Figure 1.5 and Figure 1.6)[25]. Both common and rare variants will thus be detected. However, associating these variants with a phenotype will still be difficult. For common variants, it will again be easier to do, in comparison with rare variants that will require enormous sample sizes when testing them one by one for an association.

To improve the power, different more statistical approaches have been proposed[25,57–60]. Although details vary between methods, several collapse variants together into one functional unit of choice and calculate differences in variant burden between cases and controls[25,57–60]. Some of them only include rare variants; others also include common variants and apply a weighting factor. Several other methods exist as well and focus for example on sequence similarity or use regression models[57]. Other propositions to increase the power involve family studies, extreme phenotype sampling and the use of population isolates[60]. Overall, it is clear that detecting these variants remains far from easy.

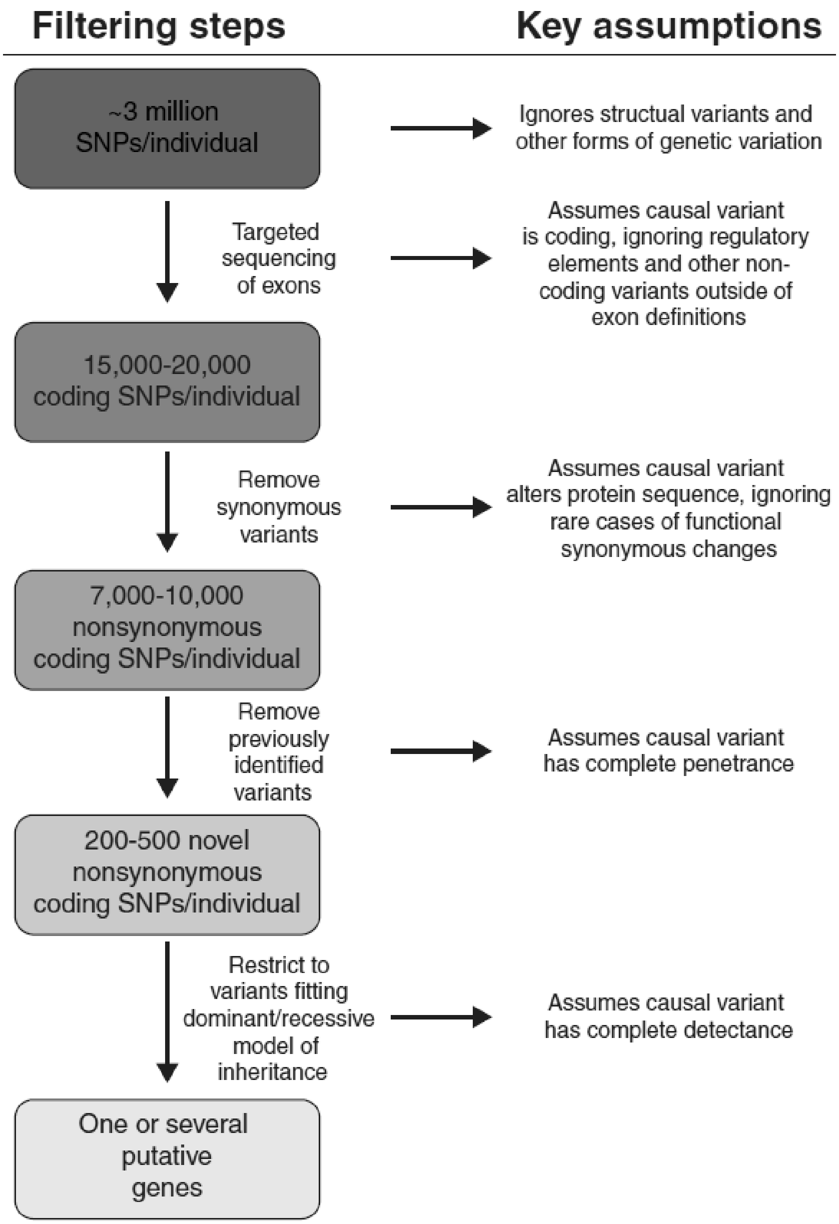

Filtering steps
Key assumptions
~3 million SNPs/individual
Ignores structual variants and other forms of genetic variation
Targeted sequencing of exons
Assumes causal variant is coding, ignoring regulatory elements and other non-coding variants outside of exon definitions
15,000-20,000 coding SNPs/individual
Remove synonymous variants
Assumes causal variant alters protein sequence, ignoring rare cases of functional synonymous changes
7,000-10,000 nonsynonymous coding SNPs/individual
Remove previously identified variants
Assumes causal variant has complete penetrance
200-500 novel nonsynonymous coding SNPs/individual
Restrict to variants fitting dominant/recessive model of inheritance
Assumes causal variant has complete detectance
One or several putative genes

Figure 1.7. Standard sequence of heuristic filters applied when filtering sequence variants[25].

C. *De novo* mutations

De novo mutations that affect the reproductive fitness (i.e. the affected individuals do not reproduce) can only be detected by direct sequencing methods as methods relying on LD require a shared ancestral haplotype to be passed on[26,61]. The study design of choice is the trio design mentioned earlier[26]. In this design, the variants in both parents are used to filter the variants in the affected individual. Theoretically, if no de novo mutations would occur and if sequencing would be perfect (no variable coverage of regions between any of the sequenced individuals and no errors in variant calling), no single variant would be retained in this design. In reality however, they all take place, and this results in a very limited number of variants being detected. This immediately points out one of the weaknesses: *de novo* mutations can easily be missed or be introduced incorrectly somewhere in the sequencing or computational processes. However, it is also clear that the power of this approach is substantial, as already demonstrated[61].

In conclusion, WES has some downsides:

- in general, the assumption for WES is that the putative causal variant lies in the protein-coding regions, although this assumption might be relaxed depending on the design;
- the power to detect rare variants for common disorders is still rather low;
- at this point, it is the most expensive method, compared to GWAS and candidate genes.

On the plus side:

- it can detect rare variants associated with common disorders;
- it allows for a more unbiased view compared with candidate genes;
- as a direct method, its efficiency is comparable with indirect methods (in the regions they both cover), or better.

At this point, WES has not been used without prior mapping in the dog to identify a disease-causing mutation. It has however been used in association with GWAS to identify an insertion deletion (indel) (c.2685delA2687_2688insTAGCTA) in the *CNGB1* gene that causes progressive retinal atrophy[62].

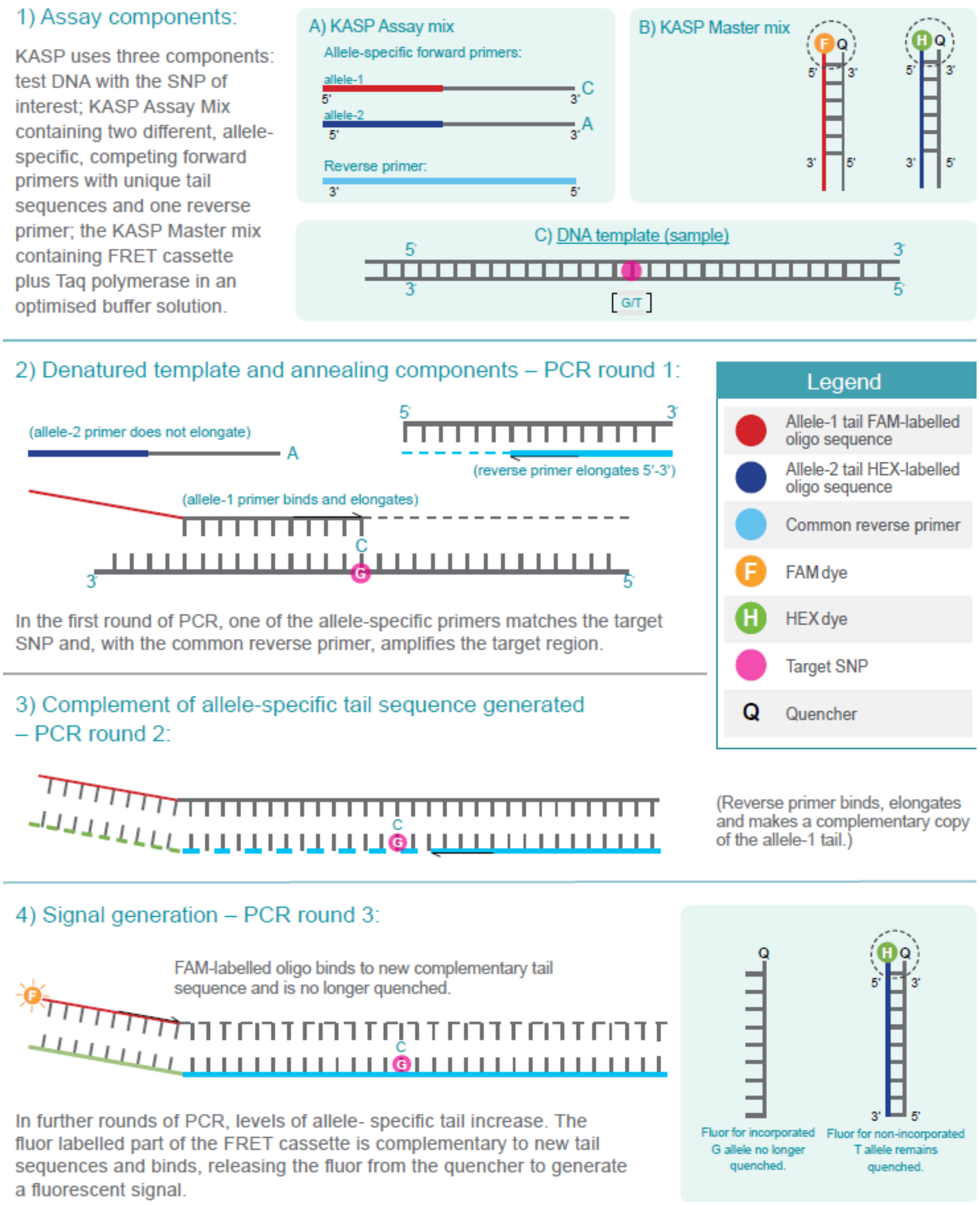
1.4 Technology

In this section, two groups of techniques are discussed. The first group requires prior knowledge of variants present in the genome. In this group, Kompetitive Allele Specific PCR (KASP) and GWAS arrays are discussed. The main difference between KASP and the GWAS arrays is that KASP is a singleplex technique whereas GWAS arrays allow for the simultaneous interrogation of several thousands of variants. The second group of techniques is based on sequencing and, in this group, the focus will be on Sanger sequencing and Illumina sequencing.

1.4.1 Probes: KASP

KASP is a fluorescence-based singleplex genotyping technology that allows for the detection of both SNPs and indels. It is based on the competition between two allele-specific forward primers with unique tail sequences that are each complementary to one of two different fluorescence resonance energy transfer (FRET) cassettes[63,64]. Because KASP cassettes are not variant specific, but primer-tail specific, the same cassettes can be used for all assays whereas the primers have to ensure the specificity for a certain variant. This decreases the cost as primers are much cheaper compared with quencher-reporter assays. Overall, KASP has proven to be a reliable, cost-effective and flexible genotyping technique[63]. An overview of the KASP technology is presented in Figure 1.8.

Figure 1.8. Overview of the Kompetitive Allele Specific PCR (KASP) genotyping technology[64].



1) Assay components:

KASP uses three components: test DNA with the SNP of interest; KASP Assay Mix containing two different, allele-specific, competing forward primers with unique tail sequences and one reverse primer; the KASP Master mix containing FRET cassette plus Taq polymerase in an optimised buffer solution.

2) Denatured template and annealing components – PCR round 1:

(allele-2 primer does not elongate)
A
5'
3'
(reverse primer elongates 5'-3')
(allele-1 primer binds and elongates)
C
G
3'
5'

In the first round of PCR, one of the allele-specific primers matches the target SNP and, with the common reverse primer, amplifies the target region.

Legend

- ● Allele-1 tail FAM-labelled oligo sequence
- ● Allele-2 tail HEX-labelled oligo sequence
- ● Common reverse primer
- F FAM dye
- H HEX dye
- ● Target SNP
- Q Quencher

3) Complement of allele-specific tail sequence generated – PCR round 2:

(Reverse primer binds, elongates and makes a complementary copy of the allele-1 tail.)

4) Signal generation – PCR round 3:

FAM-labelled oligo binds to new complementary tail sequence and is no longer quenched.

Fluor for incorporated G allele no longer quenched.

Fluor for non-incorporated T allele remains quenched.

In further rounds of PCR, levels of allele- specific tail increase. The fluor labelled part of the FRET cassette is complementary to new tail sequences and binds, releasing the fluor from the quencher to generate a fluorescent signal.

1.4.2 Large scale GWAS arrays

At this point, four different SNP microarrays are available for GWAS in the dog. They differ by the number of SNPs interrogated, the manufacturer and technology, but all arrays are based on the CanFam 2.0. The first GWAS array was developed in 2007 and allows for the interrogation of ± 27 000 SNPs[18]. A second array was designed to interrogate ± 50 000 SNPs. Both arrays were commercialized by Affymetrix with the first one combining perfect match and mismatch (PM/MM) probes and the second one containing PM probes only. The combination of PM and MM probes allows for correction of background noise[65]. In both arrays, the genotype is derived by allele-specific hybridization of DNA fragments to 25-mer probes.

The other two arrays, the CanineSNP20 BeadChip and the CanineHD BeadChip, are commercialized by Illumina and interrogate ± 22 000 SNPs and ± 172 000 SNPs, respectively[36,37]. In these arrays, after hybridization of DNA fragments to 50-mer probes on beads, a single labeled nucleotide that is complementary to the allele in the DNA is added ("single base extension") and, after signal amplification, the genotype is derived[30,65–67]. An overview of both the Affymetrix and Illumina genotyping is shown in Figure 1.9.

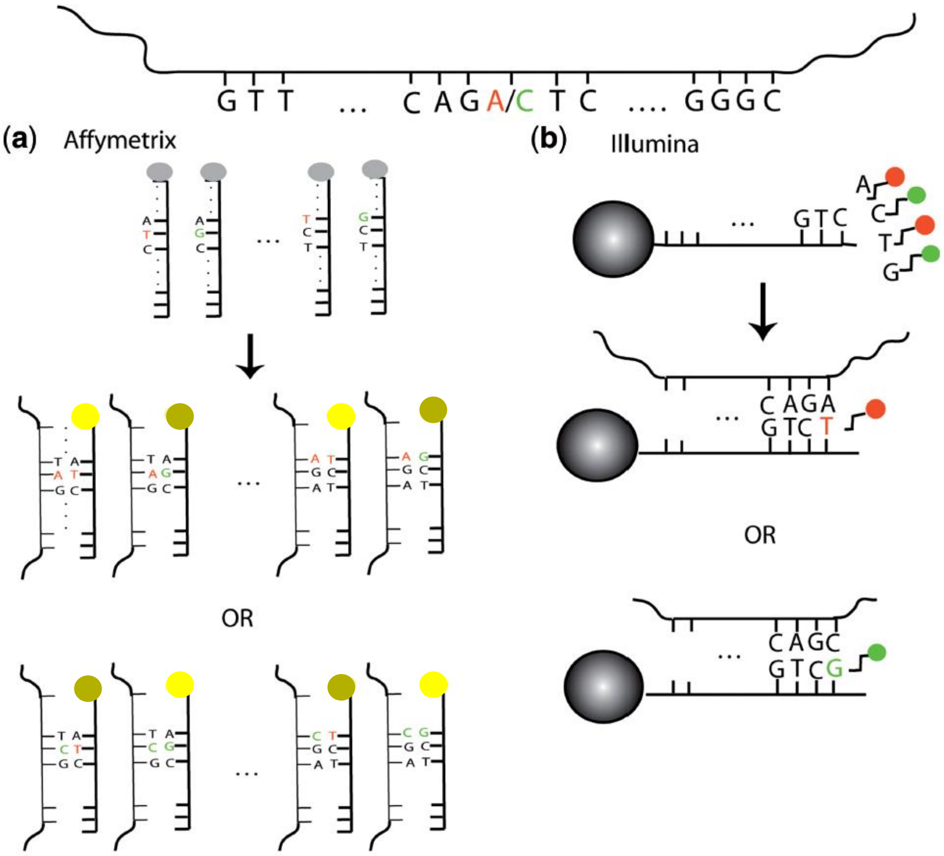


Figure 1.9. Principles of Affymetrix and Illumina genotyping technology used in the genome-wide association studies assays[65]**.** At the top is the fragment of DNA harboring an A/C SNP to be interrogated by the probes shown. (a) In the Affymetrix assay, the DNA binds to both the PM and MM probes regardless of the allele it carries, but it does so more efficiently when it is complementary to all 25 bases (bright yellow) rather than mismatching the SNP site (dimmer yellow). (b) Attached to each Illumina bead is a 50-mer sequence complementary to the sequence adjacent to the SNP site. A single-base complementary to the allele carried by the DNA is added and results in the appropriately-coloured signal (red or green, respectively).

1.4.3 Sanger Sequencing

Sanger sequencing or chain-termination sequencing has been the dominant sequencing technique for several decades. Since its development in 1977, it has evolved in several ways, but the key principles remained[68–73]. It is based on the random incorporation of deoxyribonucleotide triphosphates (dNTPs) and dideoxyribonucleotide triphosphates (ddNTPs) by a DNA polymerase. After incorporation of a ddNTP, no nucleotides can be added further due to the lack of a 3' OH group in the ddNTP. By adding these ddNTPs and dNTPs together with a DNA polymerase and a primer to a DNA template, a mixture of complementary DNA strands of variable length are synthesized. These strands of variable length are size-separated using capillary or gel electrophoresis. Initially, the ddNTPs were labelled radioactively. Nowadays, the four different ddNTPs are linked to unique fluorochromes, enabling the combination of all sequencing reactions in one. An overview of the technique is provided in Figure 1.10. Sanger sequencing allows high quality and long read (> 1000 bp) sequencing, but in terms of large-scale projects, it cannot compete with second generation sequencing techniques[69,70,74].

1.4.4 Next generation sequencing

Next generation sequencing (NGS), second generation sequencing or massively parallel sequencing are three synonyms for novel technologies that revolutionized genomics about a decade ago. As depicted by the name, they enable relatively cheap generation of enormous amounts of sequencing data. Although several competitors entered initially, Illumina dominates the market nowadays.

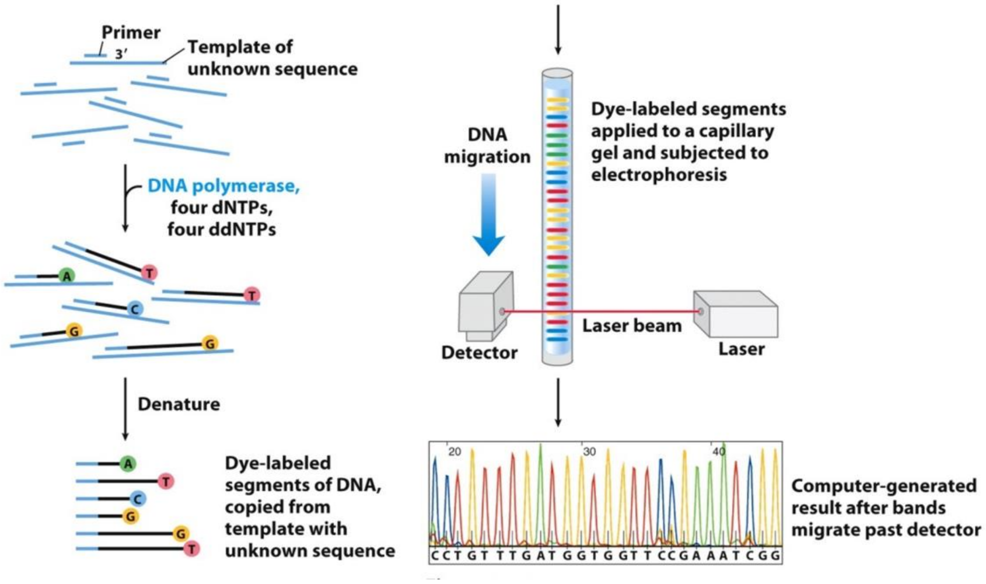
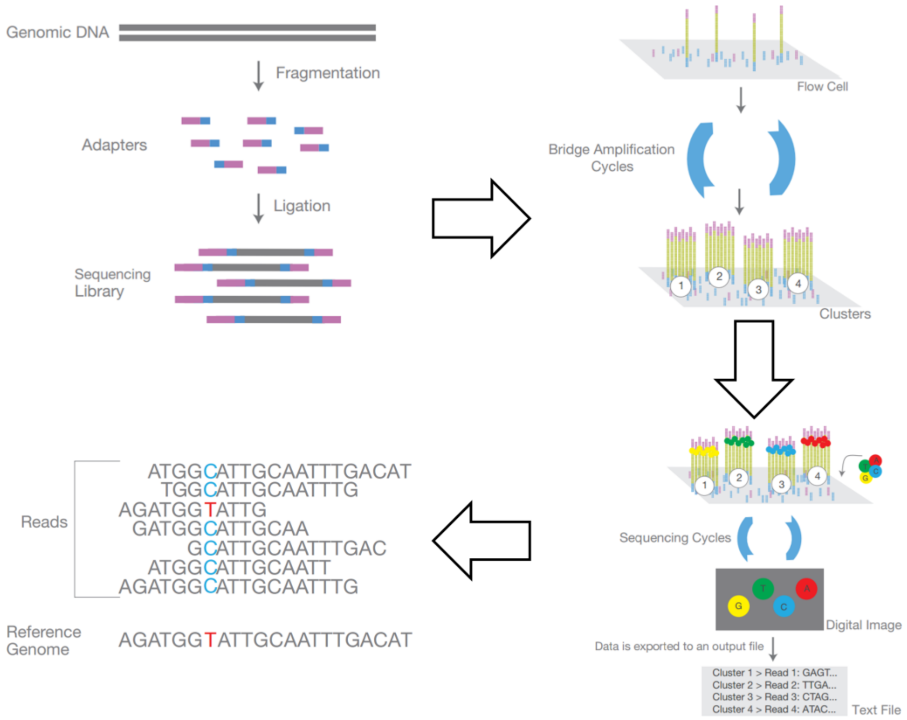

Primer
3′
Template of unknown sequence
DNA polymerase, four dNTPs, four ddNTPs
A
T
T
C
G
G
Denature
A
T
C
G
G
T
Dye-labeled segments of DNA, copied from template with unknown sequence
DNA migration
Dye-labeled segments applied to a capillary gel and subjected to electrophoresis
Detector
Laser beam
Laser
20
30
40
C C T G T T T G A T G G T G G T T C C G A A A T C G G
Computer-generated result after bands migrate past detector

Figure 1.10. Principles of Sanger sequencing[73].

The standard workflow for Illumina sequencing, can be divided in three parts: a library preparation, followed by clonal amplification and sequencing (Figure 1.11)[51,70,71]. The library prep involves DNA fragmentation, followed by adapter ligation and enrichment of adapter-ligated DNA fragments. The adapter-ligated fragments are clonally amplified by an isothermal bridge PCR, resulting in clusters. Contrary to the irreversible chain-termination Sanger sequencing, Illumina uses reversible chain-termination sequencing: after incorporation and excitation of a fluorescently labelled nucleotide with a blocking group, the blocking group and fluorescent dye are chemically cleaved and additional nucleotides can be incorporated. As with modern

Sanger sequencing, four differentially labelled nucleotides are added simultaneously and compete for incorporation in each cycle. Depending on the applications and the choice of the user, the number of cycles and thus the read lengths can vary, but at this point, the maximum read length is limited to 300 bp[75].

Figure 1.11. Basic principles of Illumina sequencing and data analysis[76].



The standard workflow allows for sequencing of entire genomes, but this comes at a price, literally. To increase cost-efficiency, several techniques have been developed to reduce sequencing to only those regions that are of interest. For a limited number of regions/a small target size, PCR generated amplicons can be sequenced[77]. However, at a certain point, this becomes a laborious, complex and expensive way of working. To solve this problem, several solutions have been developed. One of them is commercialized by Roche Nimblegen and is an in-solution based capturing method[53]. Following standard fragmentation and adapter ligation steps, it uses biotinylated oligonucleotide baits complementary to the genomic targets to hybridize to genomic DNA. Making use of the streptavidin-biotin non-covalent binding, the bound DNA is recovered by magnetic streptavidin beads later on, discarding the unwanted genomic DNA fragments, and then sequenced.

1.5 A short overview of canine hip dysplasia (HD)

One of the diseases studied in this thesis is HD. In this section, a short overview is provided. HD is a common orthopedic disorder in the dog. Literally, the term “hip dysplasia” means an abnormal development of the hip joint. However, a clearer definition includes the etiopathogenesis of HD[78]:

“ *... varying degree of laxity of the hip joint permitting subluxation during early life, giving rise to varying degrees of shallow acetabulum and flattening of the femoral head, and finally inevitably leading to osteoarthritis.*”

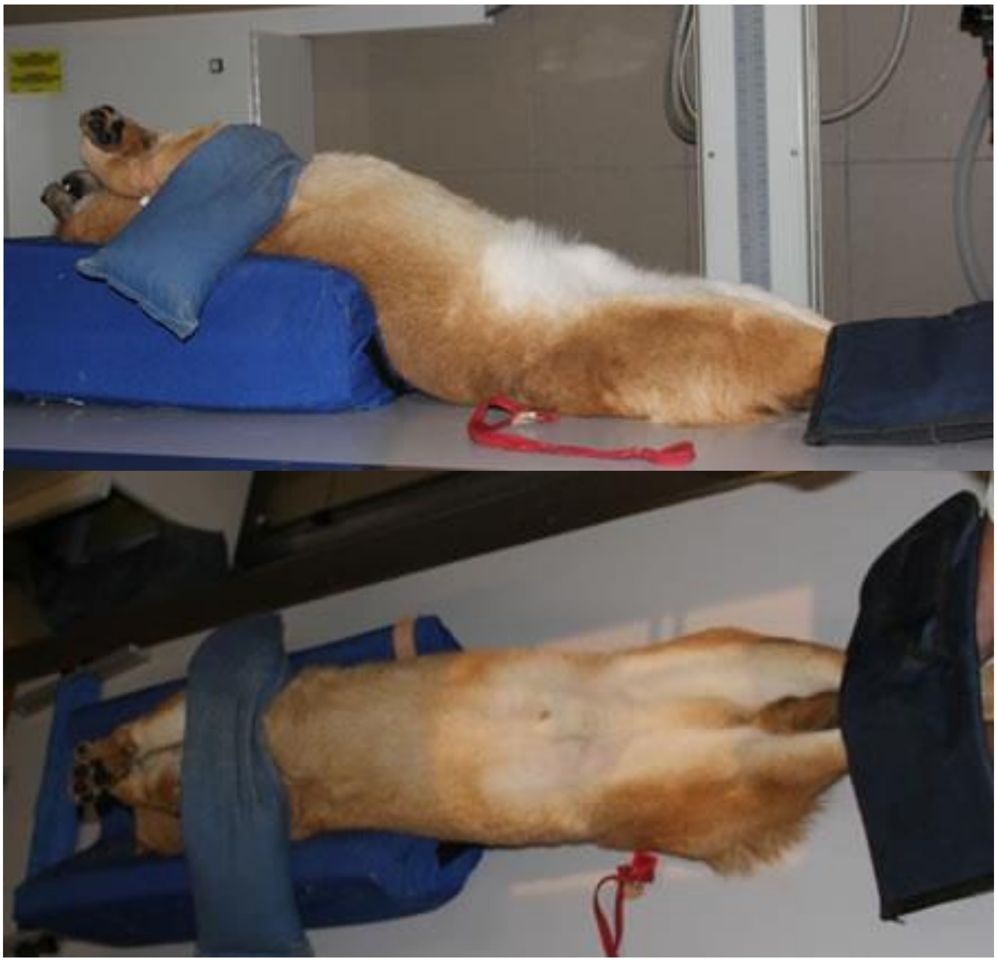
Although the exact etiology of HD is unknown, two primary causes are cited:

- an abnormal degree of joint laxity[78]
- a delayed ossification of the bones[79,80]

Both result eventually in the development of osteoarthritis.

The clinical presentation of dogs with hip dysplasia is highly variable. Typically, it affects large dogs, but rare cases in smaller breeds and even cats have been reported[81,82]. Clinical symptoms include but are not limited to an abnormal gait, bunny hopping and an excessive pelvic swaying movement[83,84]. Hip dysplasia can be the presumptive diagnosis based on clinical symptoms, but the definitive diagnose is based on radiographs. The most frequently applied technique is called the standard ventrodorsal (VD) hip-extended radiograph (Figure 1.12). For this technique, the dog is positioned in dorsal recumbency, with the hind limbs in full extension, the femora have to be positioned parallel and are endorotated to ensure the patella is projected in the middle of the femur[83].

Figure 1.12. Positioning of a dog for the standard ventrodorsal hip-extended radiograph.



The VD is often used in patients with clinical complaints to confirm that the symptoms originate from the hip joint. Frequently, it is also used for screening of potential breeding dogs. For that purpose, several classification schemes are used worldwide. In Belgium, the scoring system of the Fédération Cynologique Internationale (FCI) was adopted. This system classifies each dog in five distinct categories (from A to E) based on the appearance of the worst of both hips on the radiograph[85]. Dogs with A and B hips are considered non-dysplastic, dogs classified from C to E are considered dysplastic[85].

The main aim of the screening of breeding dogs is to reduce the prevalence of HD in the population. Unfortunately, even though screening has been going on for almost five decades, even the most optimistic reports show a very limited improvement[86]. This can be attributed to several factors. First of all, the position of the dog for the VD results in a low sensitivity to diagnose laxity of the hip joint and it is that laxity which is considered to be the most important primary cause of HD[87]. This would not be that important if the secondary changes (= the osteoarthritis) would be readily diagnosed. Unfortunately, most screening programs check the hips at an age of 24 months (United States) or even younger. A long term follow-up study showed that from all dogs that developed osteoarthritis, 78% developed it after two years of age[88]. The screening is thus performed too young to diagnose the secondary changes in the majority of the affected animals. Additional complicating factors are the unstandardized anesthesia protocols and the low interobserver agreement for the diagnosis[89-92].

These issues can be (at least partially) resolved by applying different radiographic techniques that allow an accurate assessment of the laxity in the hip joint. The most frequently used technique is PennHIP. The PennHIP procedure requires three radiographs:

- the first one is the standard VD: on this radiograph, only the secondary changes are evaluated

- a compression radiograph: to evaluate the congruency of the hip joint

- a distraction radiograph: to evaluate the laxity of the hip joint (more specific, the lateral displacement of the hip joint).

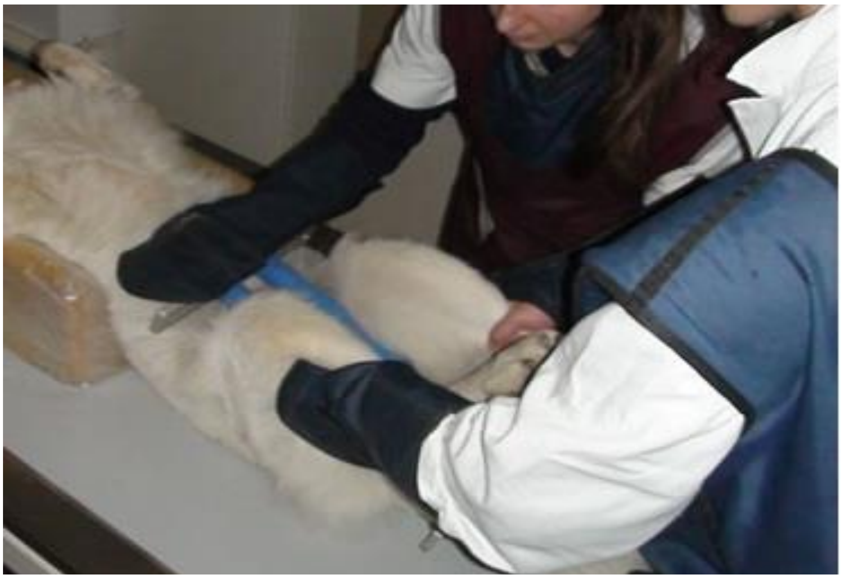


Figure 1.13. Positioning of a dog for the distraction radiograph.

For both the compression and distraction radiograph, the dog is also positioned in dorsal recumbency, but with the hip joints perpendicular to the table (called the “neutral” or “standing” position) (Figure 1.13). This position allows the accurate diagnosis of hip joint laxity from an age of 4 months[93,94]. An overview of the three radiographs obtained with PennHIP is presented in Figure 1.14. Disadvantages of PennHIP are a higher exposure to radiation and increased costs due to the two additional radiographs. PennHIP requires all veterinarians to go through a certification process to ensure the technique is applied correctly. This is of course positive, but a downside of this certification process is that it limits the accessibility for clients as the number of PennHIP certified veterinarians is rather limited. Unfortunately, PennHIP or similar techniques have not been adopted in any of the screening programs of the three major pedigree dogs organizations worldwide.

Even though several radiographic techniques exist, the fastest progress is expected from selection based on genetic information instead of selection based on phenotypes[86]. Our research currently focuses on the identification of genetic variation linked to hip dysplasia.

Screening programs try to reduce the incidence of HD in the next generations, but what if a dog is diagnosed with HD? Luckily at that point, several treatment options can be chosen from[95]. The first choice to be made is whether conservative or surgical procedures will be used. When conservative treatment is chosen, it often involves weight control, controlled exercise and when necessary, medication to alleviate pain (e.g. NSAIDS). If conservative treatment is not successful or if the indications are right, surgical procedures might be preferred. A wide range of surgical procedures can be chosen from, each with their specific indications, advantages and disadvantages. Irrespectively of the treatment option, physio- and hydrotherapy can be used to additionally support the dog and improve recovery post-surgery.

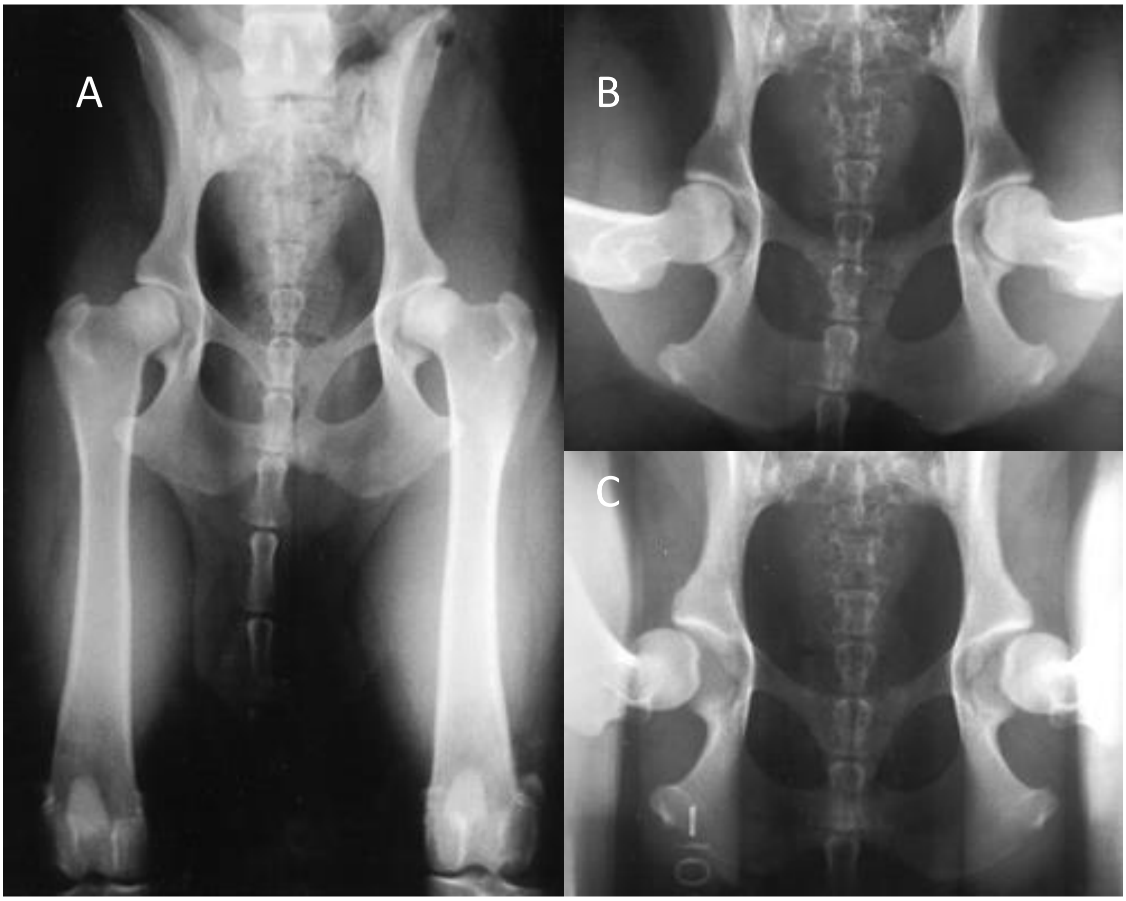


Figure 1.14. An overview of the three radiographs obtained with PennHIP. A. The standard ventrodorsal radiographic view. B. The compression radiograph. C. The distraction radiograph.

1.6 References

1. Ostrander, E. a. & Wayne, R. K. The canine genome. *Genome Res.* **15,** 1706–1716 (2005).

2. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438,** 803–819 (2005).

3. Fédération Cynologique International. at <http://www.fci.be/en/>

4. Mellersh, C. DNA testing and domestic dogs. *Mamm. Genome* **23,** 109–123 (2012).

5. Asher, L., Diesel, G., Summers, J. F., McGreevy, P. D. & Collins, L. M. Inherited defects in pedigree dogs. Part 1: Disorders related to breed standards. *Vet. J.* **182,** 402–411 (2009).

6. Summers, J. F., Diesel, G., Asher, L., McGreevy, P. D. & Collins, L. M. Inherited defects in pedigree dogs. Part 2: Disorders that are not related to breed standards. *Vet. J.* **183,** 39–45 (2010).

7. Salmon Hillbertz, N. H. C. *et al.* Duplication of FGF3, FGF4, FGF19 and ORAOV1 causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. *Nat. Genet.* **39,** 1318–1320 (2007).

8. Coopman, F. *et al.* Combined prevalence of inherited skeletal disorders in dog breeds in Belgium. *Vet. Comp. Orthop. Traumatol.* **27,** 12–14 (2014).

9. Collins, L. M., Asher, L., Summers, J. & McGreevy, P. Getting priorities straight: Risk assessment and decision-making in the improvement of inherited disorders in pedigree dogs. *Vet. J.* **189,** 147–154 (2011).

10. Nicholas, F. W. Online Mendelian Inheritance in Animals (OMIA): A comparative knowledgebase of genetic disorders and other familial traits in non-laboratory animals. *Nucleic Acids Res.* **31,** 275–277 (2003).

11. Broad Institute. Dog genome project. at <https://www.broadinstitute.org/mammals/dog>

12. Witfield, J. Dog genome unveiled. at

<http://www.nature.com/news/2003/030926/full/news030922-17.html>

13. National Institutes of Health. Talking Glossary of Genetic Terms. *National Human Genome Research Institute* at <http://www.genome.gov/glossary/>

14. Breen, M. Canine cytogenetics - From band to basepair. *Cytogenet. Genome Res.* **120,** 50–60 (2008).

15. Kirkness, E. F. *et al.* The dog genome: survey sequencing and comparative analysis. *Science* **301,** 1898–1903 (2003).

16. Hoeppner, M. P. *et al.* An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. *PLoS One* **9,** e91172 (2014).

17. The Breen Lab at North Carolina State University. The dog karyogram. at <http://www.breenlab.org/karotype.html>

18. Karlsson, E. K. *et al.* Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat. Genet.* **39,** 1321–1328 (2007).

19. Karlsson, E. K. & Lindblad-Toh, K. Leader of the pack: gene mapping in dogs and other model organisms. *Nat. Rev. Genet.* **9,** 713–725 (2008).

20. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296,** 2225–2229 (2002).

21. Sutter, N. B. *et al.* Extensive and breed-specific linkage disequilibrium in Canis familiaris. *Genome Res.* **14,** 2388–2396 (2004).

22. Quignon, P. *et al.* Canine population structure: Assessment and impact of intra-breed stratification on SNP-based association studies. *PLoS One* **2,** e1324 (2007).

23. Nicholas, F. W., Crook, A. & Sargan, D. R. Internet resources cataloguing inherited disorders in dogs. *Vet. J.* **189,** 132–135 (2011).

24. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12,** 745–755 (2011).

25. Stitziel, N. O., Kiezun, A. & Sunyaev, S. Computational and statistical approaches to analyzing variants identified by exome

sequencing. *Genome Biol.* **12,** 227 (2011).

26. Goh, G. & Choi, M. Application of whole exome sequencing to identify disease-causing variants in inherited human diseases. *Genomics Inf.* **10,** 214–219 (2012).

27. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* **40,** 695–701 (2008).

28. Pearson, T. a & Manolio, T. a. How to interpret a genome-wide association study. *JAMA* **299,** 1335–1344 (2008).

29. Lewis, C. M. Genetic association studies: design, analysis and interpretation. *Brief. Bioinform.* **3,** 146–153 (2002).

30. Bush, W. S. & Moore, J. H. Chapter 11: Genome-Wide Association Studies. *PLoS Comput. Biol.* **8,** e1002822 (2012).

31. Weiss, K. M. & Terwilliger, J. D. How many diseases does it take to map a gene with SNPs? *Nat. Genet.* **26,** 151–157 (2000).

32. Slatkin, M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9,** 477–485 (2008).

33. Visscher, P. M., Brown, M. a., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90,** 7–24 (2012).

34. Wray, N. R. Allele frequencies and the r2 measure of linkage disequilibrium: impact on design and interpretation of association studies. *Twin Res. Hum. Genet.* **8,** 87–94 (2005).

35. Terwilliger, J. D. & Hiekkalinna, T. An utter refutation of the ‘fundamental theorem of the HapMap’. *Eur. J. Hum. Genet.* **14,** 426–437 (2006).

36. Vaysse, A. *et al.* Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. *PLoS Genet.* **7,** e1002316 (2011).

37. Lequarré, A. S. *et al.* LUPA: A European initiative taking advantage of the canine genome architecture for unravelling complex disorders in both human and dogs. *Vet. J.* **189,** 155–159 (2011).

38. Carlson, C. S. *et al.* Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74,** 106–120 (2004).

39. Glickman, M. E., Rao, S. R. & Schultz, M. R. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *J. Clin. Epidemiol.* **67,** 850–857 (2014).

40. Aickin, M. & Gensler, H. Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods. *Am. J. Public Health* **86,** 726–728 (1996).

41. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc.* **57,** 289–300 (1995).

42. Xie, Y., Pan, W. & Khodursky, A. B. A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics* **21,** 4280–4288 (2005).

43. Backes, C. *et al.* Systematic permutation testing in GWAS pathway analyses: identification of genetic networks in dilated cardiomyopathy and ulcerative colitis. *BMC Genomics* **15,** 622 (2014).

44. Devlin, B. & Roeder, K. Genomic control for Association Studies. *Biometrics* **55,** 997–1004 (1999).

45. Bacanu, S. a, Devlin, B. & Roeder, K. The power of genomic control. *Am. J. Hum. Genet.* **66,** 1933–1944 (2000).

46. Awano, T. *et al.* Genome-wide association analysis reveals a SOD1 mutation in canine degenerative myelopathy that resembles amyotrophic lateral sclerosis. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 2794–2799 (2009).

47. Daly, A. K. & Day, C. P. Candidate gene case-control association studies: advantages and potential pitfalls. *Br. J. Clin. Pharmacol.* **52,** 489–499 (2001).

48. Clements, D. N. *et al.* A candidate gene study of canine joint diseases. *J. Hered.* **101,** 54–60 (2010).

49. Miyadera, K., Acland, G. M. & Aguirre, G. D. Genetic and phenotypic variations of inherited retinal diseases in dogs: The power of within- and across-breed studies. *Mamm. Genome* **23,** 40–61 (2012).

50. Newton, J. M. *et al.* Melanocortin 1 receptor variation in the

domestic dog. *Mamm. Genome* **11,** 24–30 (2000).

51. Mardis, E. R. A decade's perspective on DNA sequencing technology. *Nature* **470,** 198–203 (2011).

52. Hodges, E. *et al.* Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* **39,** 1522–1527 (2007).

53. Clark, M. J. *et al.* Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.* **29,** 908–914 (2011).

54. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42,** 30–35 (2010).

55. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS One* **7,** e46688 (2012).

56. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* 1–41 (2013). doi:10.1002/0471142905.hg0720s76

57. Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.* **11,** 773–785 (2010).

58. Luo, L., Boerwinkle, E. & Xiong, M. Association studies for next-generation sequencing. *Genome Res.* **21,** 1099–1108 (2011).

59. Neale, B. M. *et al.* Testing for an unusual distribution of rare variants. *PLoS Genet.* **7,** e1001322 (2011).

60. Auer, P. L. & Lettre, G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med.* **7,** 1–11 (2015).

61. Helsmoortel, C. *et al.* Challenges and opportunities in the investigation of unexplained intellectual disability using family based whole exome sequencing. *Clin. Genet.* **88,** 140–148 (2015).

62. Ahonen, S. J., Arumilli, M. & Lohi, H. A CNGB1 Frameshift Mutation in Papillon and Phalène Dogs with Progressive Retinal Atrophy. *PLoS One* **8,** e72122 (2013).

63. Semagn, K., Babu, R., Hearne, S. & Olsen, M. Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): Overview of the technology and its application in crop

improvement. *Mol. Breed.* **33,** 1–14 (2014).

64. LGC. KASP Genotyping. at <http://www.lgcgroup.com/LGCGroup/media/PDFs/Products/Genotyping/kasp-explanation-fact-sheet.pdf>

65. LaFramboise, T. Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances. *Nucleic Acids Res.* **37,** 4181–4193 (2009).

66. Syvänen, A.-C. Toward genome-wide SNP genotyping. *Nat. Genet.* **37 Suppl,** S5–S10 (2005).

67. Illumina. Infinium Chemistry. at <http://support.illumina.com/content/dam/illumina-support/courses/Infinium_Chemistry/>

68. Sanger, F., Nicklen, S. & Coulson, a R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–5467 (1977).

69. Hutchison, C. a. DNA sequencing: Bench to bedside and beyond. *Nucleic Acids Res.* **35,** 6227–6237 (2007).

70. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26,** 1135–1145 (2008).

71. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).

72. Nicholas, F. W. *Introduction to veterinary genetics*. (Wiley-Blackwell, 2010).

73. Nelson, D. L. & Cox, M. M. *Lehninger Principles of Biochemistry*. (W.H. Freeman, 2004).

74. Morozova, O. & Marra, M. a. Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92,** 255–264 (2008).

75. Illumina. Sequencing power for every scale. at <http://www.illumina.com/content/dam/illumina-marketing/documents/products/brochures/brochure_sequencing_systems_portfolio.pdf>

76. Illumina. An introduction to second generation sequencing techology. at <http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf>

77. Kiialainen, A. *et al.* Performance of microarray and liquid based capture methods for target enrichment for massively parallel sequencing and SNP discovery. *PLoS One* **6,** e16486 (2011).

78. Henricson, B., Norberg, I. & Olsson, S.-E. On the etiology and pathogenesis of hip dysplasia: a comparative review. *J. Small Anim. Pract.* **7,** 673–688 (1966).

79. Madsen, J. S., Reimann, I. & Svalastoga, E. Delayed ossification of the femoral head in dogs with hip dysplasia. *J. Small Anim. Pract.* **32,** 351–354 (1991).

80. Todhunter, R. *et al.* Onset of epiphyseal mineralization and growth plate closure in radiographically normal and dysplastic Labrador Retrievers. *J. Am. Vet. Med. Assoc.* **210,** 1458–& (1997).

81. Keller, G. G., Reed, a L., Lattimer, J. C. & Corley, E. a. Hip dysplasia: a feline population study. *Vet. Radiol. Ultrasound* **40,** 460–464 (1999).

82. Todhunter, R. J. & Lust, G. in *Textbook of Small Animal Surgery, 3th Edition* (ed. Slatter, D. H.) 2009–2019 (Elsevier Science, 2003).

83. Fries, C. L. & Remedios, a M. The pathogenesis and diagnosis of canine hip dysplasia: a review. *Can. Vet. J.* **36,** 494–502 (1995).

84. Ginja, M. M. D., Silvestre, A. M., Gonzalo-Orden, J. M. & Ferreira, A. J. A. Diagnosis, genetic control and preventive management of canine hip dysplasia: A review. *Vet. J.* **184,** 269–276 (2010).

85. Flückiger, M. Scoring radiographs for canine hip dysplasia - the big three organizations in the world. *Eur J Companion Anim Pr.* **17,** 135–140. (2007).

86. Wilson, B., Nicholas, F. W. & Thomson, P. C. Selection against canine hip dysplasia: Success or failure? *Vet. J.* **189,** 160–168 (2011).

87. Heyman, S. J., Smith, G. K. & Cofone, M. A. Biomechanical study of the effect of coxofemoral positioning on passive hip joint laxity in dogs. *Am. J. Vet. Res.* **54,** 210–215 (1993).

88. Smith, G. K. *et al.* Lifelong diet restriction and radiographic

evidence of osteoarthritis of the hip joint in dogs. *J. Am. Vet. Med. Assoc.* **229,** 690–693 (2006).

89. Genevois, J.-P. *et al.* Influence of anaesthesia on canine hip dysplasia score. *J. Vet. Med. A. Physiol. Pathol. Clin. Med.* **53,** 415–417 (2006).

90. Malm, S. *et al.* Impact of sedation method on the diagnosis of hip and elbow dysplasia in Swedish dogs. *Prev. Vet. Med.* **78,** 196–209 (2007).

91. Verhoeven, G. *et al.* Interobserver agreement in the diagnosis of canine hip dysplasia using the standard ventrodorsal hip-extended radiographic method: Paper. *J. Small Anim. Pract.* **48,** 387–393 (2007).

92. Verhoeven, G. E. C. *et al.* The effect of a technical quality assessment of hip-extended radiographs on interobserver agreement in the diagnosis of canine hip dysplasia. *Vet. Radiol. Ultrasound* **51,** 498–503 (2010).

93. Lust, G. *et al.* Joint laxity and its association with hip dysplasia in labrador retrievers. *Am. J. Vet. Res.* **54,** 1990–1999 (1993).

94. Smith, G. K., Gregor, T. P., Rhodes, W. H. & Biery, D. N. Coxofemoral joint laxity from distraction radiography and its contemporaneous and prospective correlation with laxity, subjective score, and evidence of degenerative joint disease from conventional hip-extended radiography in dogs. *Am. J. Vet. Res.* **54,** 1021–42 (1993).

95. Anderson, A. Treatment of hip dysplasia. *J. Small Anim. Pract.* **52,** 182–189 (2011).

2 Aims and overview

Dogs take an important place in our society. Whether they are kept as working dogs or purely act as companion animals, good health is of paramount importance. Unfortunately, the processes that created the dog as we know it, inadvertently resulted in genetic diseases in the dog being far from rare. Luckily, public awareness has led to mentality changes and with the enormous advances in biotechnology the last decade, the prerequisites for improvement are available. As discussed in the introduction (Chapter 1), health improvement requires a stepwise approach that involves recognition, characterization and prioritization of diseases, the planning of remedial measures, execution of these plans and evaluation. This dissertation focuses on the first aspects involved in this process.

The study of genetic diseases starts with the characterization of the phenotype. This involves the development of diagnostic criteria, but also gaining knowledge on the importance of the disease. As a demonstration of some of the difficulties that might arise, we focus in chapter 3 on canine hip dysplasia (HD). Typically associated with HD are low agreements between assessors and diagnostic tools and prevalence estimates that vary widely.

Deciding which approach should be used to study genetic diseases, depends on several factors: the tools that are available, the disease characteristics (both within the species studied and in other species), previous experiences and the financial resources. Based on the success in human medicine, the attention was drawn towards WES. In chapter 4 and 5, the development of three WES enrichment designs is discussed and their performance is compared. The development of novel wet lab techniques often requires the development or optimization of software tools to analyze the data as well. In chapter 6, the R-package "Mendelian" is discussed. This package was specifically designed to enable heuristic filtering of sequencing variants obtained in WES experiments in the dog in search for causal genetic variation. As a demonstration of the power of WES and "Mendelian", two coat colour loci in the Labrador Retriever were reconfirmed. Non optimal sample selection does not necessarily impede the discovery of causal mutations, but optimizing sample selection can increase the efficiency tremendously. To provide guidelines on which combination of samples is likely to be the most efficient and what can be expected in terms of variant filtering, a variety of case-control study designs were evaluated and discussed in chapter 7.

Closely associated with the phenotypical prevalence estimates are the DNA tests that can be used to determine the allele frequencies of disease-causing mutations in the population. The latter however, are more accurate as carriers can be recognized and the blurring environmental effects are omitted. This is an important step in disease prioritization and in defining the optimal strategy for health improvement. Chapter 8 details on the results when DNA tests for nine genetic disorders were performed in a population of Belgian, Dutch and German dogs. The breed-specific differences in allelic frequencies stress the need for health programs that not only focus on the global dog population, but also consider the health issues at the individual breed level.

The final chapter of this dissertation, chapter 9, provides a discussion on the present state of WES in the dog and future prospects. While acknowledging the pitfalls, the strength of WES and the potential of studying canine diseases for both the dog and their human counterparts are discussed in detail.

3 The effects of positioning, reason for screening and the referring veterinarian on prevalence estimates of canine hip dysplasia

B.J.G. Broeckx[a], G. Verhoeven[b], F. Coopman[c], W. Van Haeringen[d], T. Bosmans[e], I. Gielen[b], S. Henckens[a], J.H. Saunders[b], H. van Bree[b], B. Van Ryssen[b], V. Verbeke[a], K. Van Steendam[a], F. Van Nieuwerburgh[a], D. Deforce[a]

a Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Ottergemsesteenweg 460, 9000 Ghent, Belgium

b Department of Medical Imaging and Small Animal Orthopaedics, Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium

c Department of Applied Biosciences, Faculty of Bioscience Engineering, University College Ghent, Valentin Vaerwyckweg 1, 9000 Ghent, Belgium

d Dr. Van Haeringen Laboratorium b.v., AgroBusinessPark 100, 6708 Wageningen, The Netherlands

e Department of Medicine and Clinical Biology of Small Animals, Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium

The Veterinary Journal 201 (2014), 378-384

3.1 Abstract

Although the prevalence of canine hip dysplasia (HD) has been the subject of a number of published studies, estimates vary widely. This study evaluated several possible causes for these differences. Sixty Belgian, Dutch and German veterinarians were asked to submit all hip radiographs obtained for screening purposes (irrespective of HD status) over a 2-year period, resulting in a database of 583 dogs. Each set of radiographs was accompanied by information on the reason for screening (breeding soundness examination, clinical complaint, assistance dogs, or other reasons), and dog breed, date of birth and age.

Dog positioning exerted an effect at multiple levels. The agreement among different observers regarding correct or incorrect positioning was limited and incorrect positioning itself reduced the interobserver agreement for radiographic hip conformation. Dysplastic dogs were more commonly positioned incorrectly than non-dysplastic dogs. The clinical complaint population had a high prevalence of dysplastic dogs (> 70%) compared with the breeding population (11%) and the assistance dogs (6%). The prevalence of dysplastic dogs varied widely between breeds (16.7 – 71.4%). Dogs diagnosed with dysplasia were significantly older than dogs considered healthy (p = 0.001) and dogs classified as borderline dysplastic (p = 0.035). Interobserver agreement for hip conformation was moderately low, resulting in > 7% variation in prevalence estimates for dysplasia.

To assess the potential effect of the referring veterinarian performing the radiographic procedure, a second database (the database of the National Committee for Inherited Skeletal Disorders) was used. In this database, there was a significantly lower prevalence of HD among cases referred by veterinarians who frequently submitted hip-extended radiographs for evaluation ($p = 0.002$) compared to those who refer less frequently. However, this was likely to be selection bias, as radiographs that were from dogs suspected to be dysplastic were not submitted by frequent senders.

3.2 Introduction

Canine hip dysplasia (HD), first described in 1935, is a multifactorial, polygenetic disorder mainly characterized by hip joint laxity, which eventually leads to degenerative joint disease (DJD)[1]. This debilitating disorder is a common reason for euthanasia in dogs[2].

A broad spectrum of clinical and radiographic techniques can be used to diagnose HD[3]. The most frequently applied technique is the standard ventrodorsal (VD) hip-extended radiograph. Three other radiographic methods used to identify laxity are the PennHIP distraction index, the subluxation index and the dorsolateral subluxation score[4–6].

To reduce the prevalence of this disease, three major pedigree dog organizations, the Fédération Cynologique Internationale (FCI), the Orthopedic Foundation for Animals (OFA) and the British Veterinary Association/Kennel Club (BVA/KC), use VD hip-extended radiographs to grade the hips of potential breeding dogs[7]. In Belgium, canine pelvic radiographs are evaluated by the National Committee for Inherited Skeletal Disorders (NCISD). For certain breeds, screening for HD is obligatory and affected dogs are restricted or prohibited from breeding. To assess whether screening has beneficial effects, prevalence must be estimated. However, several studies reported variable prevalences across and within breeds[8–11].

Radiographic positioning has been shown to affect the appearance of anatomical structures[12], so it follows that incorrect positioning could perhaps reduce interobserver agreement. Based on clinical experience, we hypothesized that dogs with HD would be positioned incorrectly more frequently than those without HD. We also hypothesized that the prevalence of HD would change depending on the reason for screening (breeding soundness examination, clinical complaint, assistance dogs, or other reasons), with the highest prevalence found in those dogs presented with clinical signs of hip disease. Additionally, selection bias has been reported to affect the prevalence of HD when radiographs submitted for official evaluation are investigated[9], and we aimed to determine whether the number of radiographs sent by each veterinarian would be an independent risk factor for the diagnosis of HD. Dysplastic dogs are older than their healthy counterparts[13] and there are breed differences in prevalence[10]. Positioning can also affect diagnostic outcome and VD radiographs are typically associated with a low interobserver agreement on the presence or absence of HD[14–16].

The aim of this study was to evaluate the effect of the following parameters, which could potentially influence estimates of HD prevalence: (1) radiographic positioning; (2) the reason for screening; and (3) the referring veterinarian.

3.3 Materials and methods

Dogs. For the purposes of this study, 60 veterinarians were asked to send in every hip radiograph obtained for screening purposes (irrespective of HD status) during a 2-year period. This resulted in a sample set of 583 Belgian, Dutch and German dogs.

Approval from the local ethical (Faculty of Veterinary Medicine, Ghent University, Belgium) and deontological (Federal Public Service Health, Food Chain Safety and Environment, Brussels, Belgium) committees was granted (EC2010_171, 28 January 2011 and EC2011_193, 20 January 2012).

Radiographic evaluation. Standard VD radiographs (n = 583) were independently evaluated by two veterinarians experienced in the field of HD and film reading. The following questions were answered: (1) is the dog correctly positioned to assess hip conformation (yes or no)?; (2) based on the presence of laxity, incongruency, bony remodelling and/ or other degenerative changes on the more severely affected hip, would you consider the dog healthy, borderline or dysplastic?[17,18]; (3) if HD has been diagnosed, was the diagnosis based on the presence of degenerative joint disease (DJD), laxity (based on sub/luxation) or a combination of both, assessed separately for both hips[17,18]? Positioning was assessed according to the OFA, BVA/KC and FCI guidelines[a], which required that the pelvis was not tilted, the femurs were parallel and the patellae were centered on each femur. Radiographic examples of each subjective assessment are provided in Figure 3.1.

[a] See: http://www.offa.org/hd_procedures.html, http://www.bva.co.uk/uploadedFiles/Content/Canine_Health_Schemes/hip-dysplasia-scheme-procedure-notes-july-2015.pdf, http://www.dkk.dk/xdoc/120/46-2009-annex1.pdf (accessed 27 July 2015)

Reason for screening. For each radiograph, veterinarians were asked to provide details of the reason for screening (breeding purpose, clinical complaint, assistance dogs, other reasons), breed, date of birth and age.

Referring veterinarians. To assess the potential effect of the referring veterinarian performing the radiographic procedure, a second database (the database of the National Committee for Inherited Skeletal Disorders) was used. This database contains the radiographic results from breeding dogs evaluated between January and September 2012 ($n = 876$). Based on the frequency with which radiographs were submitted, two groups were created. Frequent senders submitted > 20 radiographs during this period, while less frequent senders submitted ≥ 20 radiographs.

Statistical analysis. Agreement between observers regarding positioning (correct/incorrect) and hip conformation (healthy/borderline/dysplastic) was evaluated for each radiograph ($n = 583$) using Cohen's kappa (κ), applying quadratic weighting for hip conformation[b]. Cut-offs were used as initially reported[19]. Group comparisons were made using chi-square tests ($\chi 2$).

[b] See: http://vassarstats.net/kappa.html (accessed 27 July 2015).

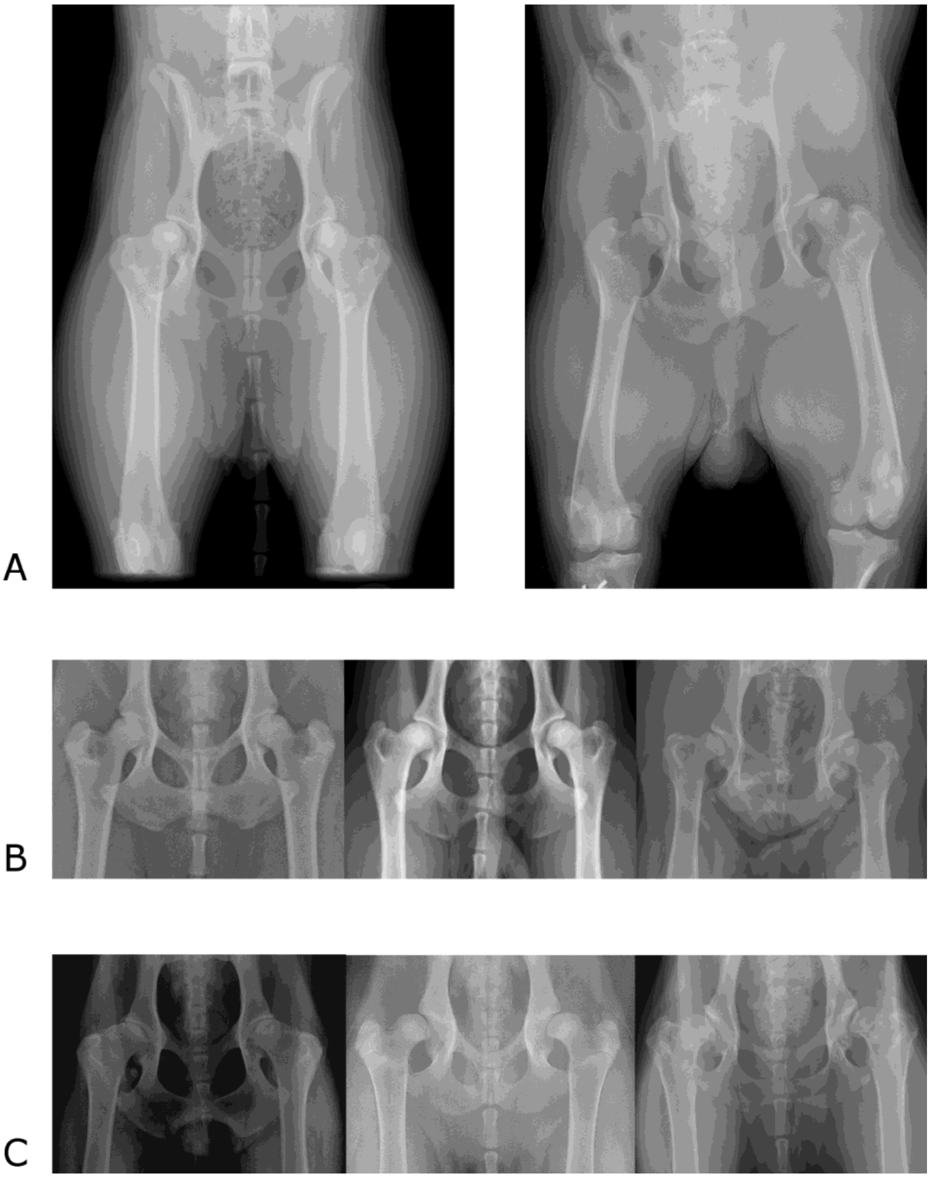


Figure 3.1. Radiographic assessment of positioning. （A） Left, correct positioning; right, incorrect positioning. （B） Hip conformation. Left, healthy; centre, borderline; right, dysplastic. （C） Reasons for the diagnosis of dysplasia. Left, DJD; centre, laxity; right, both.

To investigate the effects of variables rather than observers, only those radiographs where both assessors were in agreement were used. Further details, including sample sizes, are provided in Figure 3.2. The effect of positioning on interobserver agreement for conformation was assessed using Cohen's κ with quadratic weighting (n = 427). The effects of conformation were analyzed (n = 341), and the reasons for the diagnosis of HD, stratified by positioning, were assessed (n = 323 for DJD, n = 321 for laxity, n = 318 for both; χ2). In correctly positioned dogs, the effect of the reason for screening was assessed (n = 215).

In the NCISD population (n = 876), the effect of the frequent and less frequent senders was assessed (χ2). To assess the possible effects of positioning, the difference between the right and left Norberg angles was calculated and a comparison between the groups of frequent and infrequent senders was made (independent Student's t test).

Additionally, in correctly positioned dogs from the original population (n = 583), the effect of breed was assessed in the five breeds with the highest sample size (n = 161). Age distribution was compared using Kruskal–Wallis tests in correctly positioned dogs (n = 268), in the reason for screening subgroup (n = 211) and in the breed subgroup (n = 145). Post-hoc comparisons were performed using Mann–Whitney U tests. For normally distributed data, mean ± standard deviation was calculated and for nonparametric data, median and range were calculated. Statistical significance was set at $p < 0.05$ using a commercially available software package (SPSS version 21, IBM).

3.4 Results

The distributions of hip conformations, as independently assessed by each observer, were significantly different (p < 0.001; Figure 3.3). The general agreement between observers was approximately 80% for conformation and approximately 70% for positioning (Table 3.1). The interobserver agreement on conformation was higher in correctly positioned dogs than in incorrectly positioned dogs (Table 2). A significant difference in hip conformation was demonstrated when correctly and incorrectly positioned dogs were compared (p = 0.003). The prevalence of HD in the incorrectly positioned group was 47.2% and 24.3% in correctly positioned dogs (Figure 3.4). Only dysplastic dogs with DJD were more frequently malpositioned than dogs without DJD (p = 0.014; Table 3). No significant differences in the frequency of malpositioning were found in dogs with DJD and laxity (p = 0.114) or in dogs with laxity alone (p = 0.292; Table 3).

Table 3.1

Overall interobserver agreement.

	Agreement	κ ± SE	95% CI	Strength of agreement
Conformation –all[a,b]	0.789	0.827 ± 0.014	0.800–0.854	Almost perfect
Conformation – agreed[a,c]	0.799	0.833 ± 0.015	0.804–0.863	Almost perfect
Positioning	0.732	0.318 ± 0.043	0.233–0.403	Fair

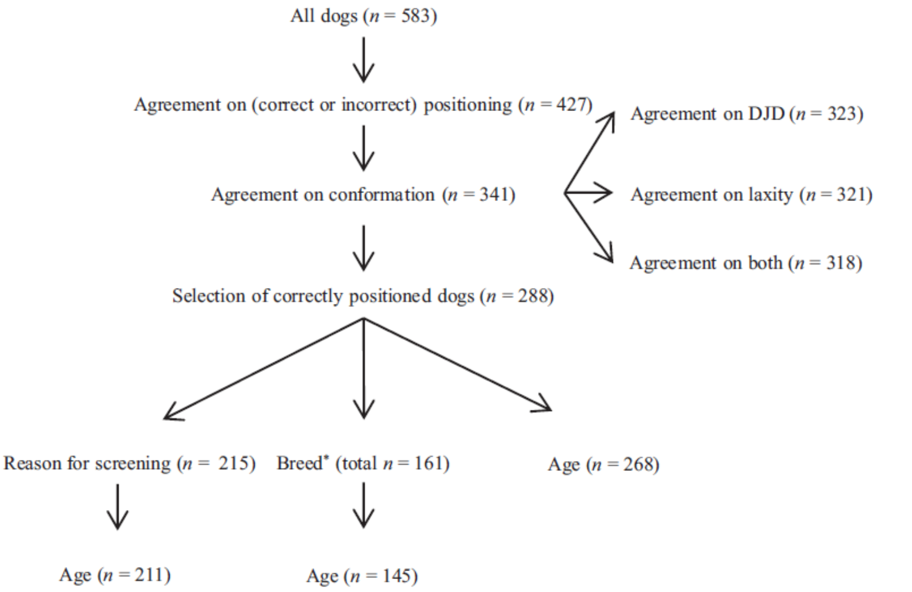
SE, standard error; CI, confidence interval.

[a] κ with quadratic weighting.

[b] All dogs (*n* = 583).

[c] Dogs where both assessors agreed on correct or incorrect positioning (*n* = 427).

Based on the reason for screening, the highest prevalence of HD was found in the clinical complaint population (n = 42/58, 72%), followed by breeding dogs (n = 10/94, 11%) and assistance dogs (n = 4/63, 6%; Figure 3.5).


All dogs ($n = 583$)
Agreement on (correct or incorrect) positioning ($n = 427$)
Agreement on DJD ($n = 323$)
Agreement on conformation ($n = 341$)
Agreement on laxity ($n = 321$)
Agreement on both ($n = 318$)
Selection of correctly positioned dogs ($n = 288$)
Reason for screening ($n = 215$)
Breed* (total $n = 161$)
Age ($n = 268$)
Age ($n = 211$)
Age ($n = 145$)

*Bernese mountain dog ($n = 7$), Border collie ($n = 14$), German shepherd dog ($n = 48$), Golden retriever ($n = 42$), Labrador retriever ($n = 50$)

Figure 3.2. Sample sizes used for statistical analysis.

Table 3.2

The effect of radiographic positioning on interobserver agreement.

	Correct positioning			Incorrect positioning		
	Agreement	κ ± SE	95% CI	Agreement	κ ± SE	95% CI
Conformation[a,b]	0.825	0.859 ± .008	0.843 - 0.876	0.680	0.718 ± 0.063	0.596 – 841

SE, standard error; CI, confidence interval.

[a] κ with quadratic weighting.

[b] Dogs where both assessors agreed on correct or incorrect positioning ($n = 427$).

Table 3.3

Effect of degenerative joint disease (DJD), laxity or both on radiographic positioning among dogs for which both observers agreed on positioning and radiographic assessment.

		Positioning		Total	p
		Incorrect n (%)	Correct n (%)		
DJD	No	42 (87.5)	266 (96.7)	308	
	Yes	6 (12.5)	9 (3.3)	15	0.014[1]
	Total	48	275	323	
Laxity	No	38 (84.4)	249 (90.2)	287	
	Yes	7 (15.6)	27 (9.8)	34	> 0.05[2]
	Total	45	276	321	
Both	No	37 (86.0)	257 (93.5)	294	
	Yes	6 (14.0)	18 (6.5)	24	> 0.05[2]
	Total	43	275	318	

[1] For dogs with DJD, the proportion positioned incorrectly is significantly higher than the proportion positioned correctly.

[2] For dogs with laxity alone or laxity and DJD, there was not a statistically significant difference in the proportions positioned incorrectly and correctly.

In the NCISD population, the prevalence of HD was almost twice as high in radiographs from veterinarians who submitted radiographs less frequently than those who were frequent senders (p = 0.002; Figure 3.6). When the difference between left and right Norberg angles was compared between the less frequent senders and frequent senders, no significant differences were found when only healthy and borderline dogs were considered (p = 0.171, frequent senders = 1.06 ± 3.34, infrequent senders = 0.71 ± 3.66) and when all dogs were considered (p = 0.122, frequent senders = 1.20 ± 4.28, infrequent senders = 0.74 ± 4.44), respectively.

Figure 3.3. Comparison of hip conformation for both observers (p < 0.001). Of 583 dogs, assessor 1 determined that 335 dogs were healthy, 62 were borderline and 186 were dysplastic. Assessor 2 determined that 341 dogs were healthy, 100 were borderline and 142 were dysplastic.

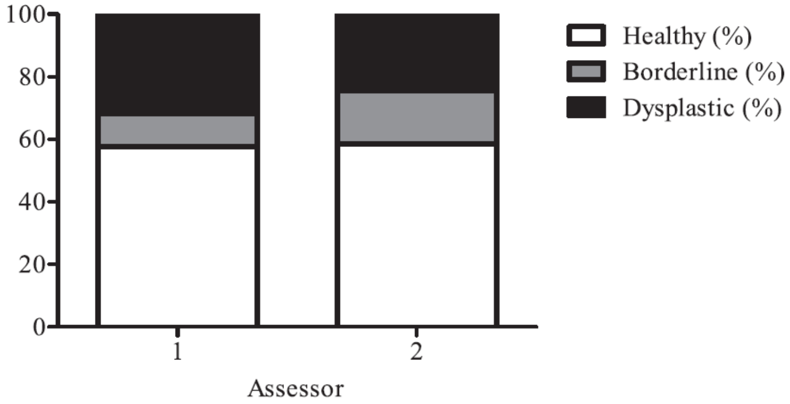
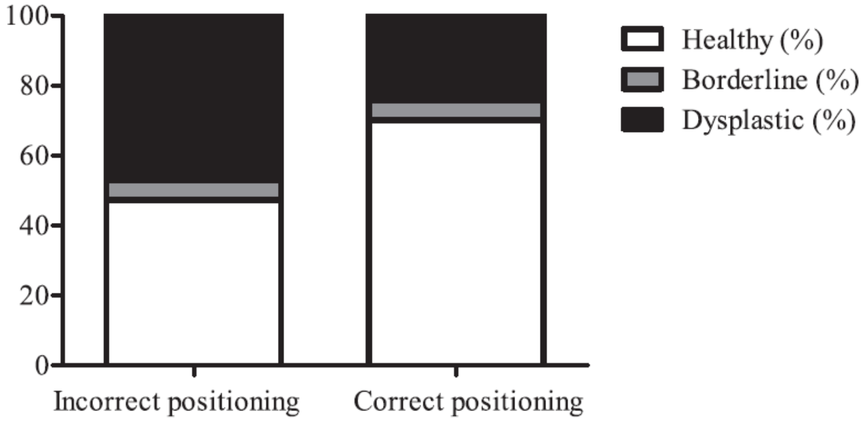


Figure 3.4. The effect of radiographic positioning on hip conformation (p = 0.003) among dogs for which both observers agreed on positioning (n = 341).



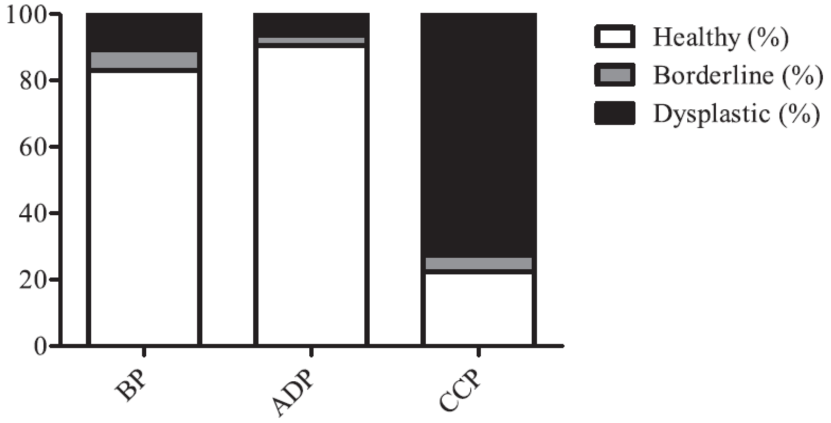


Figure 3.5. Distribution of healthy, borderline and dysplastic hips in the breeding population (BP), assistance dog population (ADP) and clinical complaint population (CCP; p < 0.001).

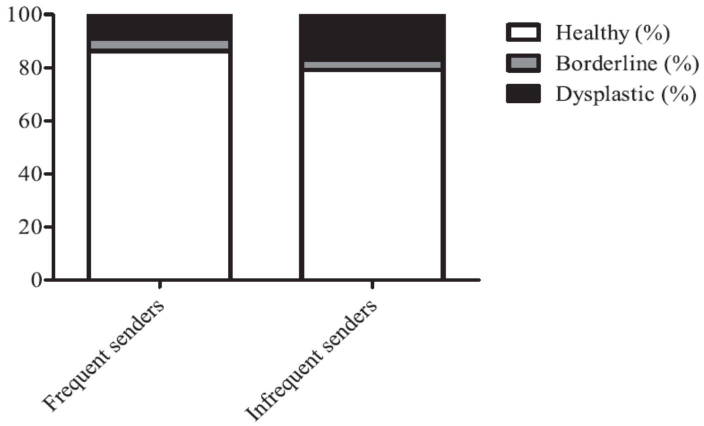


Figure 3.6. Comparison of the group of frequent senders with the group of less frequent senders with respect to hip conformation (p = 0.002).

Based on breed, prevalence estimates of HD ranged from 16.7% to 71.4% (Figure 3.7). The highest prevalence was in Bernese mountain dogs, while the lowest prevalences were in Golden retrievers and German shepherds (Figure 3.7). Dysplastic dogs (median, 2 years 1.6 months; range 5.9 months – 11 years 8.8 months) were significantly older than healthy dogs (median, 1 year 2.3 months; range 3.7 months – 8 years 2.3 months; $p < 0.001$) or borderline dogs (median, 12.9 months; range 7.2 months – 5 years 3.5 months; $p = 0.035$). Assistance dogs (median, 12.7 months; range 5.9 months – 2 years 8.7 months) were significantly younger than the breeding population (median, 1 year 6.1 months; range 5.2 months – 6 years 9.2 months; $p = 0.001$) and the clinic population (median, 2 years 4.8 months; range 4.2 months – 11 years 8.7 months; $p = 0.006$). There was no significant difference in age between the most common five breeds ($p = 0.227$). The median age (range) was 1 year 11.8 months (10.4 months – 6 years 1.8 months) for the Bernese mountain dog, 1 year 2.7 months (4.5 months – 6 years 0.4 months) for the Border collie, 1 year 0.6 months (3.7 months – 11 years 6.5 months) for the German shepherd, 1 year 1.2 months (5.9 months – 6 years 11.4 months) for the Golden retriever and 1 year 0.4 months (8.2 months – 8 years 3.6 months) for the Labrador retriever.

3.5 Discussion

The three major organizations responsible worldwide for canine HD screening (the OFA, BVA/KC and FCI) require that correct positioning is used for all radiographs submitted for evaluation. This means that the pelvis should not be tilted, the femurs should be parallel and the patellae should be centered on each femur. Although these criteria are quite clear, the decision to accept or refuse a radiograph is subjective. Good positioning is important, since it can affect the radiographic assessment of pelvic anatomical structures[12].

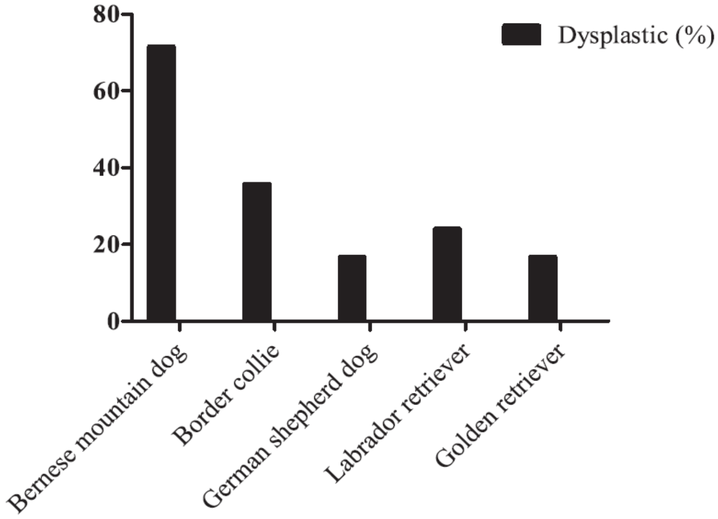


Figure 3.7. Distribution of dysplastic dogs in five different breeds.

Our study evaluated this subjective assessment and demonstrated that positioning affected prevalence estimates of HD in a number of ways. Firstly, correct positioning increased interobserver agreement (Table 3.2), confirming the findings of an earlier report[16]. Secondly, dogs with HD were malpositioned more frequently than dogs without HD (Figure 3.4) and this effect was particularly apparent in dogs with DJD (Table 3.3). It is possible that if there are clear signs of HD, veterinarians might not consider that correct positioning is necessary to make the diagnosis. However, our clinical experience supports the claim that dogs with HD and especially those with DJD are difficult to position correctly, as they have a reduced ability to extend the hip joint completely. A third observation was that there was limited interobserver agreement (κ = 0.32) as to whether or not a dog was correctly positioned (Table 3.1).

This finding is comparable with an earlier report, where agreement was unanimous in only 24% of radiographs[15] and is probably attributable to the subjective nature of the evaluation. The second and third observations affect widely quoted HD prevalence estimates reported by NCISD and similar organizations worldwide[10,11]. However, one of the prerequisites for the submission of radiographs is correct positioning. Although this is a reasonable requirement and should increase interobserver agreement, it is a subjective measure that varies between individual observers. This could lead to the exclusion of some dysplastic dogs and therefore underestimates of the prevalence of HD. We suggest that assessments should be performed by at least two experienced observers and an objective assessment of positioning should be used, such as the one suggested by Verhoeven et al. (2010). Since only correctly positioned radiographs should be evaluated, veterinarians should aim for optimal positioning, even for dogs with clear signs of HD.

Our study demonstrated an effect on HD prevalence exerted by the reason for screening, which agreed with our expectations. In the clinical complaint population, over 70% of dogs had HD, while in breeding and assistance dogs HD prevalence was 11% and 6%, respectively (Figure 3.5). The difference in prevalence between assistance dogs and breeding dogs might be attributable to age, since assistance dogs were significantly younger. However, it could also represent a true difference in prevalence. As assistance dogs are screened for performance, the selection of breeding stock for assistance dogs might be subject to stringent orthopaedic screening criteria, resulting in a lower prevalence of HD in the progeny.

NCISD and similar committees worldwide often provide data that are used to estimate prevalence, although some reports suggest they might underestimate the true prevalence of HD[9]. One possible reason is the prerequisite for correct positioning, but a second reason is selection bias, as veterinarians tend to withhold radiographs from affected dogs from official screening[9]. In our study, we found twice as many dogs in the group submitted by less frequent senders were dysplastic, even though there were no significant differences in positioning between frequent and less frequent senders (Figure 3.6). One reason for this could be that frequent senders perform pre-screening and withdraw radiographs from breeding animals from official screening. As this adversely affects prevalence estimates, the selective withdrawal of radiographs should be discontinued.

This study examined the effects of age and breed on HD prevalence. Older dogs had a significantly higher prevalence of HD, which has been previously reported and is in agreement with clinical expectations[13]. In our breed analysis, we noticed that the prevalence of HD in German shepherd dogs was lower than or equal to that in Labrador retrievers and Golden retrievers, respectively (Figure 3.7). This might indicate that German shepherd dogs are less commonly affected by HD, although this contradicts earlier reports that the German shepherd dog had the highest or second highest prevalence of dysplasia[10,20]. However, our results should be interpreted with care, as they could have been affected by the relatively low sample size. Although this study focused on purebred dogs, two other studies identified dysplastic dogs in equal proportions among mixed breeds[20,21], emphasizing the importance of HD as a problem in the general canine population.

There was at least 7% difference between observers in the number of dogs diagnosed as dysplastic in this study (Figure 3.3). This is also reflected in the moderately low interobserver agreement of 79%, since if interobserver agreement was achieved purely by chance, it would occur in approximately 33% of cases. In one study of FCI-classification, an even lower interobserver agreement of 43.6% was found[14].

All results in this study are based on the standard VD view used worldwide for official HD screening. However, the technique lacks sensitivity in the identification of laxity in the hip joints, because of the positioning necessary for VD views to be obtained[22]. The dorsolateral subluxation score, the subluxation index and the PennHIP distraction index are diagnostic techniques that can be used to identify passive hip joint laxity. The PennHIP distraction index has been studied extensively and is used frequently. Comparisons with both OFA scores and the Norberg angle, a measure of laxity on VD radiographs, show that the diagnosis of laxity can easily be missed on this view[23,24]. This suggests that the prevalence of HD, based on the standard VD view, could underestimate true prevalence.

The use of chemical restraint and the drug protocol used can influence hip laxity and thus hip grading[25,26]. Seasonal variation has also been reported to influence hip score[27]. When the presence of caudolateral curvilinear osteophyte (Morgan Line, ML) was included as an additional sign of osteoarthritis in the reading protocol, the prevalence of HD increased from 53% to 73% and 41% to 69% in Golden retrievers and Rottweilers, respectively[9]. In this study, in agreement with OFA and FCI criteria (G.G. Keller and A.C.C. Criel, personal communication), ML alone was not considered to be sufficient to declare a dog dysplastic. Only the BVA/KC includes the ML in their scoring systems[28]. This results in an effect of scoring system used on the prevalence of HD reported. A recent publication compared the OFA scoring system with a Canadian system[29] (the Ontario Veterinary College Hip Certification Program), and although agreement between systems was acceptable overall, there were variations in categorizations for 4/37 (11%) dogs that underwent evaluation, depending on the scoring system used.

3.6 Conclusions

The prevalence of HD was difficult to estimate, as numerous factors, including the reason for screening, breed and age of dogs, influenced our results. Efforts should be made to correctly position every dog for radiographic evaluation and only correctly positioned dogs should be included in studies of HD prevalence. Since interobserver agreement in this study was only fair, film reading by several experienced observers is encouraged. There was a significantly lower prevalence of HD among cases referred by veterinarians who frequently submitted hip-extended radiographs for evaluation compared with those who submitted radiographs less frequently.

3.7 References

1. Schnelle, G. B. Some new diseases in the dog. *Am. Kennel Gaz.* **52,** 25–26 (1935).

2. Bonnett, B. N., Egenvall, a, Hedhammar, a & Olson, P. Mortality in over 350,000 insured Swedish dogs from 1995-2000: I. Breed-, gender-, age- and cause-specific rates. *Acta Vet. Scand.* **46,** 105–120 (2005).

3. Fries, C. L. & Remedios, a M. The pathogenesis and diagnosis of canine hip dysplasia: a review. *Can. Vet. J.* **36,** 494–502 (1995).

4. Smith, G. K., Biery, D. N. & Gregor, T. P. New-concepts of coxofemoral joint stability and the development of a clinical stress-radiographic method for quantitating hip-joint laxity in the dog. *J. Am. Vet. Med. Assoc.* **196,** 59–70 (1990).

5. Farese, J. P., Lust, G., Williams, a. J., Dykes, N. L. & Todhunter, R. J. Comparison of measurements of dorsolateral subluxation of the femoral head and maximal passive laxity for evaluation of the coxofemoral joint in dogs. *Am. J. Vet. Res.* **60,** 1571–1576 (1999).

6. Flückiger, M. a, Friedrich, G. a & Binder, H. A radiographic stress technique for evaluation of coxofemoral joint laxity in dogs. *Vet. Surg.* **28,** 1–9 (1999).

7. Verhoeven, G., Fortrie, R., Van Ryssen, B. & Coopman, F. Worldwide Screening for Canine Hip Dysplasia: Where Are We Now? *Vet. Surg.* **41,** 10–19 (2012).

8. Leppänen, M. & Saloniemi, H. Controlling canine hip dysplasia in Finland. *Prev. Vet. Med.* **42,** 121–131 (1999).

9. Paster, E. R. *et al.* Estimates of prevalence of hip dysplasia in Golden Retrievers and Rottweilers and the influence of bias on

published prevalence figures. *J. Am. Vet. Med. Assoc.* **226,** 387–392 (2005).

10. Coopman, F., Verhoeven, G., Saunders, J., Duchateau, L. & van Bree, H. Prevalence of hip dysplasia, elbow dysplasia and humeral head osteochondrosis in dog breeds in Belgium. *Vet. Rec.* **163,** 654–658 (2008).

11. Genevois, J. P. *et al.* Prevalence of hip dysplasia according to official radiographic screening, among 31 breeds of dogs in France. *Vet. Comp. Orthop. Traumatol.* **21,** 21–24 (2008).

12. Thompson, R., Roe, S. C. & Robertson, I. D. Effects of pelvic positioning and simulated dorsal acetabular rim remodeling on the radiographic shape of the dorsal acetabular edge. *Vet. Radiol. Ultrasound* **48,** 8–13 (2007).

13. Smith, G. K. *et al.* Lifelong diet restriction and radiographic evidence of osteoarthritis of the hip joint in dogs. *J. Am. Vet. Med. Assoc.* **229,** 690–693 (2006).

14. Verhoeven, G. *et al.* Interobserver agreement in the diagnosis of canine hip dysplasia using the standard ventrodorsal hip-extended radiographic method: Paper. *J. Small Anim. Pract.* **48,** 387–393 (2007).

15. Verhoeven, G. E. C. *et al.* Interobserver agreement on the assessability of standard ventrodorsal hip-extended radiographs and its effect on agreement in the diagnosis of canine hip dysplasia and on routine fci scoring. *Vet. Radiol. Ultrasound* **50,** 259–263 (2009).

16. Verhoeven, G. E. C. *et al.* The effect of a technical quality assessment of hip-extended radiographs on interobserver agreement in the diagnosis of canine hip dysplasia. *Vet. Radiol. Ultrasound* **51,** 498–503 (2010).

17. Smith, G. K. Advances in diagnosing canine hip dysplasia. *J. Am. Vet. Med. Assoc.* **210,** 1451–1457 (1997).

18. Dassler, C. L. in *Textbook of Animal Surgery* (ed. Slatter, D.) 2019–2029 (Saunders, 2003).

19. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33,** 159–174 (1977).

20. Rettenmaier, J. L., Keller, G. G., Lattimer, J. C., Corley, E. a & Ellersieck, M. R. Prevalence of canine hip dysplasia in a veterinary teaching hospital population. *Vet. Radiol. ultrasound* **43,** 313–318 (2002).

21. Bellumori, T. P., Famula, T. R., Bannasch, D. L., Belanger, J. M. & Oberbauer, A. M. Prevalence of inherited disorders among Mixed-Breed and Purebred Dogs : 27,254 cases (1995 - 2010). *J. Am. Vet. Med. Assoc.* **242,** 1549–1555 (2013).

22. Heyman, S. J., Smith, G. K. & Cofone, M. A. Biomechanical study of the effect of coxofemoral positioning on passive hip joint laxity in dogs. *Am. J. Vet. Res.* **54,** 210–215 (1993).

23. Culp, W. T. N. *et al.* Evaluation of the Norberg angle threshold: A comparison of Norberg angle and distraction index as measures of coxofemoral degenerative joint disease susceptibility in seven breeds of dogs. *Vet. Surg.* **35,** 453–459 (2006).

24. Powers, M. Y. *et al.* Evaluation of the relationship between Orthopedic Foundation for Animals' hip joint scores and PennHIP distraction index values in dogs. *J. Am. Vet. Med. Assoc.* **237,** 532–541 (2010).

25. Genevois, J.-P. *et al.* Influence of anaesthesia on canine hip dysplasia score. *J. Vet. Med. A. Physiol. Pathol. Clin. Med.* **53,** 415–417 (2006).

26. Malm, S. *et al.* Impact of sedation method on the diagnosis of hip and elbow dysplasia in Swedish dogs. *Prev. Vet. Med.* **78,** 196–209 (2007).

27. Worth, A., Bridges, J., Cave, N. & Jones, G. Seasonal variation in the hip score of dogs as assessed by the New Zealand

Veterinary Association Hip Dysplasia scheme. *N. Z. Vet. J.* **60,** 110–114 (2012).

28. Dennis, R. Interpretation and use of BCA/KC hip scores in dogs. *In Pract.* **34,** 178–194 (2012).

29. Chalmers, H. J., Nykamp, S. & Lerer, A. The Ontario Veterinary College Hip Certification Program - Assessing inter- and intra-observer repeatability and comparison of findings to those of the Orthopedic Foundation for Animals. *Can. Vet. J.* **54,** 42–46 (2013).

4 Development and performance of a targeted whole exome sequencing enrichment kit for the dog (Canis Familiaris Build 3.1)

B.J.G. Broeckx[a], F. Coopman[b], G.E.C. Verhoeven[c], V. Bavegems[d], S. De Keulenaer[a], E. De Meester[a], F. Van Nieuwerburgh[a*], Dieter Deforce[a*]

a Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Ottergemsesteenweg 460, 9000 Ghent, Belgium

b Department of Applied Biosciences, Faculty of Bioscience Engineering, University College Ghent, Valentin Vaerwyckweg 1, 9000 Ghent, Belgium

c Department of Medical Imaging and Small Animal Orthopaedics, Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium

d Department of Medicine and Clinical Biology of Small Animals, Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium

* These authors contributed equally to this work.

Scientific Reports 4 (2014), 5597

4.1 Abstract

Whole exome sequencing is a technique that aims to selectively sequence all exons of protein-coding genes. A canine whole exome sequencing enrichment kit was designed based on the latest canine reference genome (build 3.1.72). Its performance was tested by sequencing 2 exome captures, each consisting of 4 pre-capture pooled, barcoded Illumina libraries on an Illumina HiSeq 2500. At an average sequencing depth of 102x, 83 to 86% of the target regions were completely sequenced with a minimum coverage of five and 90% of the reads mapped on the target regions. Additionally, it is shown that the reproducibility within and between captures is high and that pooling four samples per capture is a valid option. Overall, we have demonstrated the strong performance of this WES enrichment kit and are confident it will be a valuable tool in future disease association studies.

4.2 Introduction

Since the first reported study on Whole Exome Sequencing (WES) in 2007[11], well over 2000 papers have been published applying this technique (PubMed search: "exome sequencing"). WES is a cost-efficient approach to selectively sequence the coding regions of the genome. This approach allows to identify most functional variation without the high costs associated with whole genome sequencing (WGS). Unfortunately, predesigned and validated kits are only commercially available for human and mouse. Scientific reports of WES on other animals are scarce[2–4]. This is unfortunate, as for example the dog is an excellent animal model for comparative disease genetics[5]. To fill the gap, we designed a canine exome sequencing kit (based on build 3.1.72) and tested its performance using Illumina Sequencing.

4.3 Results

Design. We designed a canine whole exome sequencing enrichment kit based on the latest canine reference genome (Broad CanFam 3.1.72)[6]. This design was based on the combination of the Ensembl Genes, the RefSeq Genes and the mRNA annotation. Additionally, known microRNAs were added from miRBase. After merging overlapping regions, the total size of the design was 52,876,195 bp (≈ 2% of the genome) divided over 203,059 regions. Based on our design, capturing baits were developed by Roche Nimblegen to target the specific regions. To avoid too much off-target sequencing, the most stringent setting was chosen for the baits design, allowing only unique matches from each bait to the reference genome.

Performance: coverage and specificity. To assess the performance of the WES enrichment kit, two exome captures were done, each consisting of four pooled samples. Sequencing depth, coverage of targeted regions and targeted bases and specificity were assessed for every sample. The results are reported for a minimum coverage of one and five (as five was the threshold used for variant calling). Each capture library was sequenced in one Illumina HiSeq 2500 lane. The number of raw reads generated per sample varied between 74,657,388 and 111,624,766. After quality trimming, mapping and duplicate read removal, between 87% and 90% of the reads were retained (Table 4.1). The average sequencing depth overall was 102x and ranged from 82.6x to 125.1x (Table 4.1).

Table 4.1

Statistics for exome sequencing eight dogs.

	Sequencing reads					
Sample	Total	Mapped	Duplicate	Remaining	Remaining (%)	Sequencing depth (x)
1	82,574,410	77,392,469	4,820,648	72,571,821	87.9	93.0
2	74,657,388	69,542,653	4,518,820	65,023,833	87.1	82.6
3	90,534,096	83,841,822	4,680,806	79,161,016	87.4	102.0
4	77,786,110	72,147,586	4,457,341	67,690,245	87.0	87.1
5	111,624,766	108,781,536	9,882,797	98,898,739	88.6	125.1
6	96,041,166	93,261,066	8,278,081	84,982,985	88.5	106.9
7	103,290,412	100,440,603	8,653,736	91,786,867	88.9	116.7
8	86,094,438	83,226,207	5,926,249	77,299,958	89.8	99.3

Overall, an average of 92% of the regions were covered by at least one read and 90% by at least five reads. At a minimum coverage of one, 89 to 90% of the regions in our design, were completely covered.

83 to 86% of the regions were completely covered when a minimum coverage of five was applied (Figure 4.1). A clear relationship exists between the percentage of each region being sequenced and the proportion of total regions being sequenced (Figure 4.1). On average, a minimum coverage of five was not consistently reached throughout the entire region for 15% of the regions (Table 4.2). However, for only 8% of the regions on average, the maximum coverage never reached five (Table 4.2).

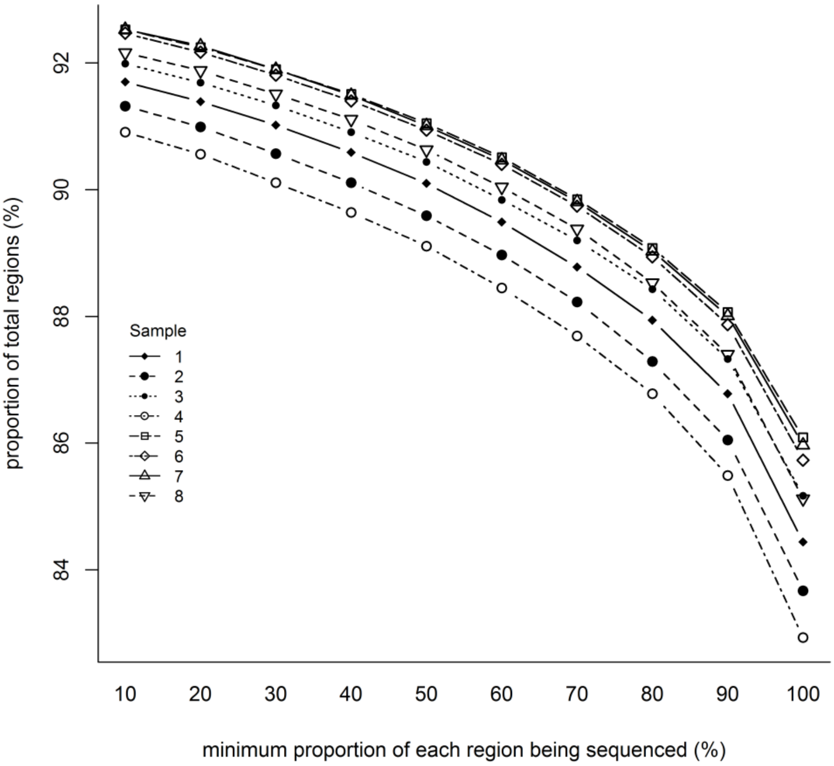


Figure 4.1. Relation between the proportion of each region being sequenced and the total amount of regions sequenced (%). For each individual region per sample,

the percentage of the region being sequenced at a minimum coverage of five, was calculated. On average 85% of the regions were completely sequenced. This number increased to 87% of the regions being sequenced for at least 90%. Around 90% of the regions were being sequenced for at least 60%.

Table 4.2

Regions with a coverage below 5.

Sample	Regionswithminimum coverage < 5x (%)	Regionswithmaximum coverage < 5x (%)
1	31,604(15.56)	16,330 (8.04)
2	33,167(16.33)	17,042 (8.39)
3	30,122(14.83)	15,800 (7.78)
4	34,655(17.07)	17,831 (8.78)
5	28,250(13.91)	14,733 (7.26)
6	28,979(14.27)	14,824 (7.30)
7	28,487(14.03)	14,696 (7.24)
8	30,224(14.88)	15,465 (7.62)

Table 4.3

Coverage of targeted base pairs.

Sample	% of target bp covered (> 1x)	% of target bp covered (> 5x)
1	93.15	89.96
2	92.90	89.54
3	93.22	90.24
4	92.82	89.15
5	93.52	90.63
6	93.66	90.71
7	93.53	90.63
8	93.56	90.53

The second and third column show the percentage of base pairs from the design of 52,876,195 basepairs with a coverage of at least one and five, respectively, within each sample.

When looking at the coverage of targeted bases instead of regions, 93 to 94% (≈ 49 Mb) of the targeted bases (≈ 53 Mb) were covered at least once and 89 to 91% were covered at least five times (Table 4.3). We also assessed the specificity (reads on target/total number of reads). With an average overall specificity of 90%, off-target sequencing is rather small and comparable with earlier reports[7]. The specificity was also assessed in every single sample per chromosome. For all eight samples, results were similar, with the highest specificity on average found on chromosome nine (94.05%) and the lowest specificity found on average on chromosome 22 (84.09%). Per chromosome specificity is available in Supplementary Table S4.1.

Performance: reproducibility. The reproducibility within and between captures was checked by comparing the amount of targeted bases and regions that are sequenced at least once and five times in every single sample. Overall, from the ≈ 53 Mb target base pairs, 48,141,464 base pairs (91.0%) were sequenced at least once in all eight samples. We also assessed how many base pairs were never sequenced. Overall, 2,313,892 base pairs (4.4%) were never covered. Overall, the remaining 4.6% of the total target base pairs are being sequenced variably. Comparing the four samples within each capture, we found that 48,333,432 (91.4%) or 48,663,244 (92.0%) base pairs were common and 2,816,548 (5.3%) or 2,553,759 (4.8%) base pairs were never sequenced. A similar analysis was conducted for a coverage of five. For all eight samples, 46,236,131 base pairs (87.4%) were sequenced consistently with a minimum coverage of five and 4,078,886 base pairs (7.7%) never reached a coverage of five. 46,439,217 (87.8%) or 47,102,104 (89.1%) and 4,572,396 (8.6%) or 4,216,408 (8.0%) base pairs were common within each pool reaching a coverage of at least five or never reaching a coverage of five, respectively.

The regions in common were also assessed for a coverage of 1 and five. From the 203,059 regions, overall, 4,791 (2.4%) regions were never sequenced and 176,645 (87,0%) were consistently covered at least once. Within each pool, 177,664 (87.5%) or 179,463 (88.4%) regions were common and 6,620 (3.3%) or 5,722 (2.8%) regions were never sequenced. For a coverage of five, 11,691 regions (5.8%) were consistently not sequenced sufficiently and 160,366 (79.0%) regions were. This results in 31,002 (15.3%) of the regions being variably sequenced. Within each pool for a coverage of five, 162,312 (79.9%) or 167,830 (82.7%) reqions were sequenced and 13,484 (6.6%) or 12192 (6.0%) regions were not. The non-covered base pairs and regions are probably a consequence of the chosen stringency when baits were designed as only unique matches were allowed. A table containing these annotated 11,691 regions is available on request.

Sample pooling. Pooling several samples together prior to capturing is common practice, mainly to reduce cost. Of course, pre-capture pooling should only be done when it does not significantly decrease the enrichment performance. To check the effect of pre-capture pooling, we created subsets containing 25% randomly chosen reads out of the total number of reads in the combined output of the four samples per capture. The rationale is to simulate samples as if they were not barcoded and as if the DNA strands presented to the capture baits are from one sample. A random subset of 25% needs to be taken to reduce the total number of reads to a number comparable to the number of reads in the individual barcoded samples. Per pool, ten subsets were created, resulting in a total of twenty new samples. Comparing the number of regions that are completely covered at least once in the subsets and the original samples, an average of 687 and 684 additional regions (≈ 0.3%) were covered in the subsets of pool one and two, respectively. An average of 119,097 and 131,774 additional target base pairs were covered at least once in the respective subsets, which represents 0.2% of the ≈ 53 Mb design. At a minimum sequencing depth of five, an average of 1,705 (0.8%) and 1,468 (0.7%) additional completely covered regions and an additional 174,070 (0.3%) and 164,522 (0.3%) base pairs were covered for pool one and two, respectively. The average specificity increased from 90.85% to 91.48% and from 89.86% to 90.73% in both pools, respectively. Based on these results, we conclude that pre-capture pooling of samples is a valid option as the effect on the different performance parameters is minimal. The exact number of samples that can be pooled, depends on a cost-benefit assessment.

Variant calling. Finally, we also called variants using a probabilistic variant caller. The number of non-reference variants called per sample, ranged from 55,683 to 60,576 and from 62,117 to 67,890 with the "require presence in both forward and reverse reads" setting being applied or not, respectively (Supplementary Table S4.2). Applying this setting might exclude variants at the boundaries of the targeted regions, however it decreases the amount of erroneous variants[8].

4.4 Discussion

This study is the first to report on the performance of an exome kit designed for the dog on the latest annotation (CanFam 3.1.72). With on average 90% of the regions and 90% of the bases covered five times, a high amount of the targeted regions is captured, without too much off-target sequencing as the specificity is 90%. The reproducibility within and between captures is high. Additionally, the results indicate that pooling four samples per capture is a valid option as it has only a very limited effect on the performance, but substantially reduces the costs. This makes WES even more affordable (compared with WGS). Finally, it is demonstrated that WES is capable to detect variants within the coding regions. Overall, we have demonstrated the strong performance of this WES enrichment kit and are confident it will be a valuable tool in future disease association studies.

4.5 Methods

Sample collection. Eight blood samples were obtained from a canine blood bank available at Ghent University to study genetic disorders[9]. Approval was granted by the local ethical (Faculty of Veterinary Medicine, Ghent University, Belgium) and deontological (Federal Public Service Health, Food Chain Safety and Environment, Brussels, Belgium) committees (EC2013_193).

Design. The data needed to design the exome kit was downloaded from the University Of California Santa Cruz (UCSC) (http://genome.ucsc.edu/) table browser (Dog, CanFam3.1)[10]. From the Genes and Gene prediction tracks, RefSeq Genes and Ensembl Genes were selected. The output format was a BED file with the setting "exons (plus 0 bases at each end)". From the mRNA and EST Tracks, Dog mRNAs and all_mrna were selected respectively. The output format was also a BED file with the "blocks plus 0 bases at each end" setting. MicroRNA sequence positions were downloaded from miRBase[11]. Regions were merged using bedtools version v2.17.0. The total size of the design was 52,876,195 Mb (≈ 2% of the genome) divided over 203,059 regions. The BED file is available on request.

Roche Nimblegen WES enrichment kit. Our design was processed by the Roche Nimblegen custom design group (Madison, USA). Using an SSAHA algorithm, capturing baits were developed based on our design and the reference genome of the dog (Canis Familiaris 3.1). Design settings for the baits allowed five or fewer single-base insertions, deletions or substitutions between the baits and the genome. Each bait itself was only allowed to match one location in the genome to avoid too much off target sequencing. Regions under 100 bp were padded to 100 bp to increase capturing efficiency. After approval, the baits were generated and provided as SeqCap Developer Library.

DNA extraction. Genomic DNA was extracted with the DNeasy Blood & Tissue Kit (QIAGEN) with 100 µl of blood as input. The standard protocol was followed with the exception of the final elution step: instead of using 200 µl of Buffer AE, only 50 µl was used. The eluate was used again to elute a second and third time to increase the concentration. The DNA yield was measured with Quant-iT^{TM} Picogreen® dsDNA Assay (Life Technologies).

Sample preparation and sequencing. Extracted DNA was fragmented on a Covaris S2 System in a 50µl volume (aim: 300 bp fragments, settings: duty cycle: 10%, intensity: 5, cycles per burst: 200, time: 50s). After shearing, another picogreen assay was performed. Around one µg of the fragmented DNA was used as input for the library preparation. Samples were end repaired, A-tailed and ligated with TruSeq adapters using the reagents from the NEBNext DNA Library Prep Master mix set for Illumina (New England Biolabs) according to the manufacturer's protocol. AMPure XP beads (Beckman Coulter) were used for selection of fragments with an insert size around 300 bp. One µl of the ligated product was subsequently amplified in an enrichment PCR (10 cycles) for library quality assessment as recommended in the 'SeqCap EZ Library SR

User's Guide' (Nimblegen, Roche). Thereafter, the pre-capture LM-PCR was performed on the samples for 8 cycles as prescribed in the SeqCap EZ library protocol. The concentration of each PCR product was determined using Quant-iT^{TM} Picogreen® dsDNA Assay (Life Technologies). Two times four samples were equimolarly pooled to obtain a total DNA input of 1250 ng. The pooled library was hybridized for 67-68 hours with the baits (SeqCap Developer Library). The hybridized library was washed and the captured and pooled DNA was recovered. After a final amplification (LM-PCR, 18 cycles), the quality of the library was checked using the High Sensitivity DNA chip (Agilent).

QPCR. To check the fold enrichment after capturing, a qPCR is performed as a final quality control step before sequencing. We chose to test five loci. Primer one is the standard primer provided by Roche Nimblegen (NSC-0237). The other four primers were designed using NCBI Primer-BLAST (http://www.ncbi.nlm.nih.gov/tools/primer-blast/)[12]. Sequences are available in Supplementary Table S4.3. The amplification efficiency of each primer was determined by qPCR. One µl of the following template DNA quantities were added to each reaction: 20 ng, 10 ng, 5 ng, 2.5 ng and 1.25 ng. Each reaction was performed in triplicate. Efficiencies E were calculated with the following formula: E = $10^{(-1/\text{slope of standard curve})}$ and are mentioned in Supplementary Table S4.3. To assess the fold enrichment for both pools of four samples, a qPCR was performed according to the instructions from Roche Nimblegen. Fold enrichment was calculated using the following formula: ($E^{\text{delta-Ct}}$) with delta-Ct being the difference in threshold cycle between the library prior and post capturing. The average fold enrichment was well over the tenfold threshold suggested by Roche Nimblegen.

Sequencing. The two pools were sequenced on two different lanes in two different runs on a HiSeq 2500 PE 100 bp.

Data-analysis. Data-analysis was performed using the CLC Genomics Workbench (Version 6.5.1, CLC Bio, Aarhus, Denmark). Data was trimmed with the following settings: ambiguous trim = no, quality trim = yes, quality limit = 0.05, use colourspace = no, create report = yes, also search on reversed sequence = yes, save discarded sequences = yes, remove 5' terminal nucleotides = no, discard short reads = no, remove 3' terminal nucleotides = no, trim adapter list = adapter list Illumina, discard long reads = no, save broken pairs = yes. The reference genome was downloaded from the UCSC genome browser[6]. For read mapping, the following parameters were used: mismatch cost = 2, insertion and deletion cost = 3, length fraction: 0.5, similarity fraction = 0.8, global alignment = no, auto-detect paired distances = yes, non-specific match handling = ignore, output mode = create reads track, create report = yes, collect un-mapped reads = yes. Duplicated reads were removed with the Duplicate Mapped Reads Removal (Version 1.0 beta 5) plugin (setting: maximum representation of minority sequence (percent) to 20.0). Reads were locally realigned with the following settings: realign unaligned ends = yes, multi-pass realignment = 3, guidance-variant track = not set, output mode = create reads track, output track of realigned regions = yes. Variants were called twice using probability variant detection with the following settings: ignore non-specific matches = yes, ignore broken pairs = yes, minimum coverage = 5, variant probability = 90.0, required variant count = 2, ignore variants in non-specific regions = yes, filter 454/Ion homopolymer indels = no, maximum expected variants = 2, genetic code = 1 standard , create track = yes, create annotated table = yes. The first variants were called with the "require presence in both forward and reverse reads = yes", the second call was run without this setting.

Effect of pooling. Forward and reverse reads from each pool were combined in two large pools. From each pool, ten random subsets were created using seqtk version 1.0-r31. Data was analysed using the same settings as the real samples.

4.6 References

1. Hodges, E. *et al.* Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* **39,** 1522–1527 (2007).

2. Ahonen, S. J., Arumilli, M. & Lohi, H. A CNGB1 Frameshift Mutation in Papillon and Phalène Dogs with Progressive Retinal Atrophy. *PLoS One* **8,** e72122 (2013).

3. Cosart, T. *et al.* Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics* **12,** 347 (2011).

4. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42,** 30–35 (2010).

5. Lequarré, A. S. *et al.* LUPA: A European initiative taking advantage of the canine genome architecture for unravelling complex disorders in both human and dogs. *Vet. J.* **189,** 155–159 (2011).

6. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438,** 803–819 (2005).

7. Clark, M. J. *et al.* Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.* **29,** 908–914 (2011).

8. Nguyen, P. *et al.* Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. *BMC Genomics* **12,** 106 (2011).

9. Broeckx, B. J. G. *et al.* The Prevalence of Nine Genetic Disorders in a Dog Population from Belgium, the Netherlands and Germany. *PLoS One* **8,** e74811 (2013).

10. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32,** D493–D496 (2004).

11. Kozomara, A. & Griffiths-Jones, S. MiRBase: Integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* **39,** 152–157 (2011).

12. Ye, J. *et al.* Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13,** 134 (2012).

4.7 Supplementary material

Supplementary Table S4.1

Per sample per chromosome specificity (calculated as mapped reads on target/total number of mapped reads).

	Specificity per sample (%)							
Chr	1	2	3	4	5	6	7	8
1	92.23	91.96	92.57	92.26	91.65	91.3	91.20	91.79
2	91.59	91.29	92.02	91.71	91.01	90.66	90.52	91.10
3	89.48	89.41	89.96	89.78	88.43	88.17	87.83	88.59
4	90.26	90.12	90.64	90.53	89.43	89.05	88.75	89.46
5	92.82	92.55	93.17	92.8	92.32	91.91	91.82	92.38
6	92.68	92.46	93.13	92.74	92.26	91.8	91.75	92.44
7	91.95	91.67	92.24	92.13	91.37	90.92	90.82	91.39
8	90.28	90.10	90.80	90.44	89.63	89.31	89.12	89.86
9	94.19	93.88	94.53	94.17	94.22	93.66	93.66	94.11
10	91.45	91.26	91.88	91.61	90.84	90.49	90.36	90.96
11	91.06	90.88	91.36	91.19	90.43	90.07	89.88	90.46
12	91.73	91.53	92.16	91.89	91.17	90.86	90.69	91.35
13	89.63	89.63	90.18	90.14	88.60	88.50	88.09	88.80
14	89.98	89.81	90.37	90.29	88.99	88.75	88.42	89.05
15	91.03	90.83	91.47	91.28	90.31	90.01	89.80	90.40
16	89.82	89.33	90.31	89.89	88.66	88.59	88.71	89.20
17	91.99	91.74	92.38	92.12	91.50	91.07	91.00	91.50
18	92.91	92.63	93.28	92.80	92.44	92.12	92.07	92.66
19	85.92	86.31	86.69	86.73	84.87	84.63	83.91	84.78
20	93.36	93.01	93.69	93.12	93.12	92.70	92.74	93.30
21	91.15	90.93	91.46	91.12	90.43	90.09	90.03	90.54
22	84.69	84.86	85.38	85.35	83.18	83.21	82.59	83.44
23	90.41	90.34	90.78	90.77	89.64	89.27	88.96	89.58
24	92.08	91.89	92.44	92.05	91.55	91.14	91.06	91.68
25	90.81	90.68	91.29	91.11	90.07	89.85	89.62	90.25
26	92.22	91.96	92.55	92.19	91.95	91.53	91.34	92.02
27	92.33	92.01	92.65	92.37	91.98	91.54	91.50	92.06
28	91.64	91.47	92.02	91.93	91.26	90.77	90.60	91.03
29	88.01	87.98	88.57	88.75	86.75	86.59	85.97	86.70
30	92.91	92.66	93.22	93.06	92.70	92.28	92.03	92.57
31	86.13	86.18	86.75	86.23	84.99	84.57	84.36	85.40
32	88.97	88.81	89.37	89.57	88.30	88.08	87.72	88.19
33	91.77	91.67	92.05	92.13	91.20	90.76	90.61	91.05
34	88.89	88.71	89.31	89.08	87.74	87.52	87.12	87.96
35	89.14	89.03	89.53	89.41	88.07	87.84	87.42	88.21
36	92.80	92.67	93.20	93.30	92.20	91.95	91.73	92.04
37	91.20	91.10	91.62	91.57	90.66	90.26	90.00	90.58
38	91.20	90.98	91.70	91.16	90.46	90.16	89.97	90.63
X	87.85	87.88	88.53	87.99	87.74	87.62	87.26	87.89

Supplementary Table S4.2

Non-reference variants detected on target per sample.

Sample	Variants (bidirectional)	Variants
1	57,827	65,060
2	56,401	63,491
3	58,769	66,639
4	57,623	64,793
5	60,444	67,222
6	60,576	67,890
7	55,683	62,117
8	58,615	65,967

The second and third column show the number of variants being called with and without "the "require presence in both forward and reverse reads" setting being applied, respectively.

Supplementary Table S4.3

Primers (and their efficiencies) used to assess fold enrichment.

Primer	Forward	Reverse	E	Chr
1	5'-CGCATTCCTCATCCCAGTATG-3'	5'-AAAGGACTTGGTGCAGAGTTCAG-3'	1.60	12
2	5'-GTAGTGAGGCGAGTGGCTTT-3'	5'-CCGACAGCACTACATGGGTT-3'	2.06	36
3	5'-CTCCTGGGGCACAAATGAGT-3'	5'-AGGGAGAATATGGCCCACCT-3'	1.81	30
4	5'-TCTGTGAGGGTGGCTTTTCC-3'	5'-TCTCTGGGGCATCTGTGAGA-3'	1.75	17
5	5'-TCGCTGACGTGTTCAAAGGA-3'	5'-AGAACCCACGCCTGAAGATG-3'	1.67	3

The second and third column contain the sequences of all five primers. Primer one is the standard primer provided by Nimblegen (NSC-0237). The fourth column contains the efficiency of amplification (E) as calculated with the following formula: $E = 10^{(-1/\text{slope of standard curve})}$. The chromosome on which the control locus is located is mentioned in the fifth column.

5 Improved canine exome designs, featuring ncRNAs and increased coverage of protein coding genes

B.J.G. Broeckx[a], C. Hitte[b], F. Coopman[c], G.E.C. Verhoeven[d], S. De Keulenaer[a], E. De Meester[a], T. Derrien[b], J. Alfoldi[e], K. Lindblad-Toh[e,g], T. Bosmans[f], I. Gielen[d], H. Van Bree[d], B. Van Ryssen[d], J.H. Saunders[d], F. Van Nieuwerburgh[a*], D. Deforce[a*]

a Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Ottergemsesteenweg 460, 9000 Ghent, Belgium

b Institut de Génétique et Développement de Rennes, CNRS-URM6290, Université Rennes1, avenue du Professeur Léon Bernard 2, 35043 Rennes-Cedex, France

c Department of Applied Biosciences, Faculty of Bioscience Engineering, University College Ghent, Valentin Vaerwyckweg 1, 9000 Ghent, Belgium

d Department of Medical Imaging and Small Animal Orthopaedics, Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium

e Broad Institute of MIT and Harvard, Cambridge, Main Street 415, Cambridge, 02142 Massachusetts, USA

f Department of Medicine and Clinical Biology of Small Animals, Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium

g Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, 751 23 Uppsala, Sweden

* These authors contributed equally to this work.

Scientific Reports 5 (2015), 12810

5.1 Abstract

By limiting sequencing to those sequences transcribed as mRNA, whole exome sequencing is a cost-efficient technique often used in disease-association studies. We developed two target enrichment designs based on the recently released annotation of the canine genome: the exome-plus design and the exome-CDS design. The exome-plus design combines the exons of the CanFam 3.1 Ensembl annotation, more recently discovered protein-coding exons and a variety of non-coding RNA regions (microRNAs, long non-coding RNAs and antisense transcripts), leading to a total size of ≈ 152 Mb. The exome-CDS was designed as a subset of the exome-plus by omitting all 3' and 5' untranslated regions. This reduced the size of the exome-CDS to ≈ 71 Mb. To test the capturing performance, four exome-plus captures were sequenced on a NextSeq 500 with each capture containing four pre-capture pooled, barcoded samples. At an average sequencing depth of 68.3x, 80% of the regions and well over 90% of the targeted base pairs were completely covered at least five times with high reproducibility. Based on the performance of the exome-plus, we estimated the performance of the exome-CDS. Overall, these designs provide flexible solutions for a variety of research questions and are likely to be reliable tools in disease studies.

5.2 Introduction

In 2014, the first report detailing the design and performance of a whole exome sequencing (WES) enrichment assay for the dog was published by our group[1]. Aiming to selectively sequence all the regions that are transcribed to mRNA, WES is a reliable tool used to identify disease-causing or predisposing mutations at a fraction of the price of whole genome sequencing (WGS) studies. A limitation of WES is that it is based on our current knowledge of the annotation of the genome and that many disease causing mutations are likely to fall outside protein-coding regions. With new information becoming available, updates and extensions are required. Recently, an improved annotation for the dog genome has been published and new data on non-protein coding genes has been obtained[2]. Based on this data, two new target enrichment designs for dogs, called the exome-plus and the exome-CDS, were developed. The exome-plus offers the most comprehensive design. The exome-CDS is a subset of the exome-plus, focusing on the coding DNA sequences (CDS) by excluding the 3' and 5' untranslated regions (UTRs). These two designs offer flexible solutions for a variety of research questions associated with targeted dog exome resequencing. Our current study describes the development of the new designs and the performance of the exome-plus. Based on the results of the exome-plus, we estimate the performance of the exome-CDS. In addition, we provide an in-depth comparison with the previously published exome-1.0[1].

5.3 Results and discussion

Design. Commercially available target enrichment technologies are able to capture up to 200 Mb, which is around 10% of the dog genome. The choice of which regions to include can therefore be based on practical and theoretical considerations, instead of technical limitations. A smaller design does not necessarily result in a cheaper target capture assay, as most commercial custom target capture design tools will increase the tiling density of the baits when the target region size decreases and thus a similar amount of baits are produced. This increased tiling density might increase the capture efficiency. The main cost difference between smaller and bigger designs lies in the increased sequencing cost: more sequence reads will need to be generated to achieve the same sequencing depth on a bigger target compared to a smaller captured region of interest. With these considerations in mind, two separate designs were developed.

The first design, called the exome-plus, has a total size of 151,698,592 bp (≈ 6% of the genome) divided over 242,914 regions. The exome-plus contains both protein-coding genes and their UTRs and specific non-coding genes. The protein-coding regions contain the exons from the Ensembl annotation (*Canis familiaris*, CanFam 3.1) and newly discovered protein-coding exons recently identified by RNA-sequencing[2]. The non-coding genes are a combination of the microRNAs from miRBase[3], experimentally characterized long non-coding RNA[2] and antisense transcripts[2].

The second design, the exome-CDS, was designed to be a subset of the exome-plus, containing only the CDS from both the Ensembl annotation and the newly discovered protein-coding genes. The 3' and 5' UTRs were thus excluded. Candidate CDS within transcript sequences were identified through TransDecoder[4]. Interestingly, this bioinformatics tool discovered a small number of additional exons and CDS, adding a total of 115,044 bp (0.16% of the size exome-CDS) that were not shared with the exome-plus. Overall, the exome-CDS targets 71,254,801 bp (≈ 3% of the genome) spread over 244,543 regions.

Based on these designs, capturing baits were developed by Roche Nimblegen to target specific regions. When the baits were designed, the regions on the mitochondrial DNA were omitted to avoid overcapturing and oversequencing of the mitochondrial DNA compared with the nuclear DNA[5]. If mitochondrial DNA sequencing is required, one of the options would be to design baits separately and to spike them in at a low concentration.

Sequencing. In total, 16 canine Labrador Retriever DNA samples were sequenced using the exome-plus design. To assess performance, four separate captures were performed, each consisting of four different pre-capture pooled and indexed samples. Each pool was sequenced in a separate run on a NextSeq 500 Illumina sequencing system. These results were also used to estimate the coverage performance for the exome-CDS, which is a subset of the exome-plus. On average, 243 million reads were generated per sample (Table 5.1). Following quality trimming, mapping and duplicate reads removal, 87.2% of the reads were retained on average. This result is comparable with previous reports[1,6].

Performance of the exome-plus: coverage. The on target sequencing depth for the exome-plus varied from 42.6x to 93.9x and was on average 68.3x (Table 5.1). To assess the regions and base pairs covered, a cut-off sequencing depth of 5x was used as this is the threshold applied usually for variant calling.

Table 5.1
Statistics for exome sequencing sixteen dogs.

Sample	Pool	Total reads	Mapped reads	Duplicate reads	Remaining reads	Remaining (%)	Sequencing depth (x)
1	1	284,357,886	264,735,195	9,414,179	255,321,016	89.8	85.8
2	1	281,522,490	261,170,320	9,366,318	251,804,002	89.4	84.3
3	1	249,659,670	231,433,861	7,819,714	223,614,147	89.6	75.4
4	1	181,728,820	168,679,105	4,382,284	164,296,821	90.4	55.5
5	2	266,996,086	251,028,902	17,002,907	234,025,995	87.7	75.3
6	2	187,857,302	176,207,940	12,226,544	163,981,396	87.3	53.9
7	2	233,403,500	216,361,182	13,330,685	203,030,497	87.0	65.0
8	2	314,005,584	289,641,450	23,514,154	266,127,296	84.8	82.9
9	3	262,726,150	246,019,167	14,919,187	231,099,980	88.0	74.8
10	3	181,120,464	169,294,819	10,140,076	159,154,743	87.9	51.6
11	3	269,017,896	247,287,291	16,377,215	230,910,076	85.8	73.1
12	3	243,350,554	227,421,662	12,820,659	214,601,003	88.2	69.5
13	4	154,004,914	142,631,944	9,095,086	133,536,858	86.7	42.6
14	4	193,942,804	175,552,484	13,670,936	161,881,548	83.5	50.1
15	4	221,094,842	204,380,382	15,086,795	189,293,587	85.6	59.5
16	4	364,079,702	337,983,155	31,679,889	306,303,266	84.1	93.9

From the total of 242,914 targeted regions of the exome-plus, on average 193,722 regions (79.7%) were completely covered with a depth of at least five reads. The number of partially sequenced target regions with a minimal percentage covered, increases when the minimal required percentage of coverage is lowered: e.g. 88% of the regions are covered at least 90%. For on average 9192 regions (3.8%), the maximum sequencing depth reached was four reads. An overview can be found in Table 5.2. The relation between the number of regions with a minimal coverage and the percentage minimally covered is visualized in Figure 5.1.

Table 5.2

Regions with a sequencing depth below 5x.

Sample	Pool	Regions with minimum sequencing depth < 5x (%)	Regions with maximum sequencing depth < 5x (%)
1	1	42,705 (17.6)	6,977 (2.9)
2	1	42,749 (17.6)	6,980 (2.9)
3	1	45,739 (18.8)	7,307 (3.0)
4	1	55,643 (22.9)	9,346 (3.8)
5	2	41,798 (17.2)	8,502 (3.5)
6	2	54,390 (22.4)	10,953 (4.5)
7	2	50,032 (20.6)	10,439 (4.3)
8	2	40,793 (16.8)	8,238 (3.4)
9	3	44,312 (18.2)	8,884 (3.7)
10	3	57,615 (23.7)	11,185 (4.6)
11	3	44,956 (18.5)	8,698 (3.6)
12	3	46,948 (19.3)	9,125 (3.8)
13	4	66,381 (27.3)	12,380 (5.1)
14	4	57,903 (23.8)	10,446 (4.3)
15	4	41,031 (16.9)	7,457 (3.1)
16	4	54,075 (22.3)	10,156 (4.2)

In terms of covered base pairs, on average 95.1% of the targeted bases pairs reach a minimum sequencing depth of five (Table 5.3). Overall, these results are similar to commercially available human exome sequencing kits[6].

Performance of the exome-plus: percentage reads on target. The percentage (%) reads on target is calculated as the number of reads on target, divided by the total number of reads. This parameter informs us of the enrichment efficiency. Overall, the average % reads on target (for all chromosomes and samples) is 75.8%. The lowest average chromosome

% reads on target is 63.3% for chromosome X, the highest is 82.2% for chromosome 9. Only a small difference on the % reads on target for chromosome X was noticed when the two sexes were compared: the average % reads on target was 62.6% for males (n = 8) and 63.8% for females (n = 8). Detailed results are provided in Supplementary Table S5.1. The obtained percentages were similar in the sequenced samples.

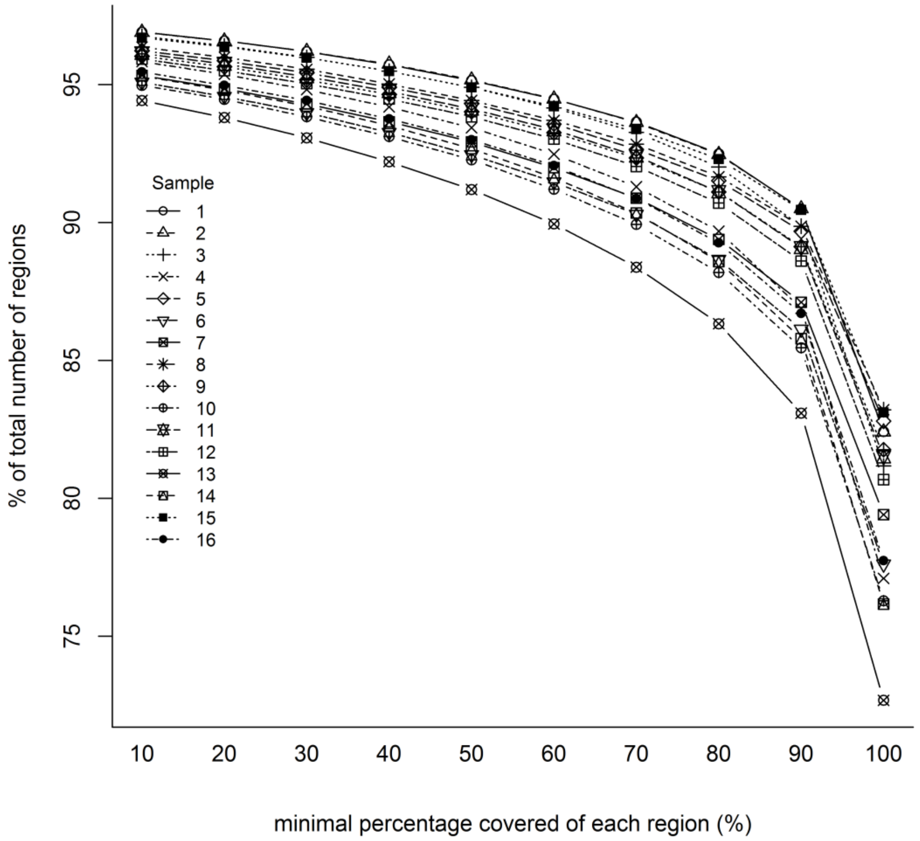


Figure 5.1. Relation between the minimal percentage covered of each region (%) and the percentage of the total number of regions (%). For each individual region,

the proportion of the region covered at a minimum sequencing depth of 5x, was calculated.

Table 5.3

Coverage of targeted base pairs (≥ 5x).

Sample	Pool	base pairs exome-plus (%)	base pairs exome-CDS (%)
1	1	145,066,553 (95.6)	67,225,626 (94.3)
2	1	145,081,909 (95.6)	67,250,020 (94.4)
3	1	144,607,207 (95.3)	67,007,231 (94.0)
4	1	143,036,351 (94.3)	66,060,290 (92.7)
5	2	145,147,282 (95.7)	67,070,797 (94.1)
6	2	143,617,018 (94.7)	66,121,333 (92.8)
7	2	144,051,293 (95.0)	66,351,108 (93.1)
8	2	145,267,122 (95.8)	67,097,056 (94.2)
9	3	144,802,496 (95.5)	66,904,165 (93.9)
10	3	143,126,270 (94.3)	65,865,667 (92.4)
11	3	144,861,681 (95.5)	66,904,165 (93.9)
12	3	144,585,713 (95.3)	66,786,409 (93.7)
13	4	142,170,547 (93.7)	65,328,731 (91.7)
14	4	143,170,286 (94.4)	66,018,207 (92.7)
15	4	145,360,509 (95.8)	67,273,067 (94.4)
16	4	143,449,544 (94.6)	66,120,048 (92.8)

Performance of the exome-plus: reproducibility. We determined the overall reproducibility for both the targeted regions and the targeted base pairs. For all 16 samples, 154,318 regions (63.5%) were completely covered at a minimum sequencing depth of 5x in every single sample. For 4,220 (1.7%) of the regions the maximum sequencing depth reached was four reads for all 16 samples. In terms of base pairs, 137,071,014 base pairs (90.4%) are consistently sequenced at least five times and 3,642,390 base pairs (2.4%) never reach a sequencing depth of 5x.

Assessment of possible reasons for differences in sequencing depth between regions. It seems that the regions can be divided in three categories based on their sequencing performance. Group 1 contains the 154,318 regions that were completely covered in all sixteen samples at a minimum sequencing depth of 5x. Group 2 contains the 84,376 regions that at least partially did not reach a minimum sequencing depth of 5x in all 16 samples. The last group, group 3, contains the 4,220 regions with a maximum sequencing depth of 4x for all sixteen samples. We evaluated whether differences in GC content and bait design could be linked to the obtained sequencing performance of these regions. It has been reported that a low sequencing depth can be caused by a high or low GC content[6–8]. Initially, we compared the GC content per region (% GC) for all three groups, resulting in median % GC values of 46.7%, 54.9% and 76.8% for group 1, 2 and 3, respectively (Figure 5.2, yellow boxes). Sharp drops of sequencing depth have been reported with % GC above 60.0% and below 40.0%[6]. As all the regions in group 1 were completely covered, we considered that group to be a reliable reference and used the 2.5th and 97.5th percentile of that group as cut-off values for a sequenceable % GC. Based on these cut-offs (which were 32.0% and 64.5%, respectively), we determined for each group the proportion of regions with a more extreme % GC. The proportion of regions with a more extreme value were 4.9%, 24.4% and 75.2% for group 1, group 2 and group 3, respectively. Based on these results, especially group 3 has a relatively large group of extreme % GCs.

Roche Nimblegen does not design baits in regions with low complexity or regions that are highly repetitive to avoid off-target sequencing (called the "repeats" from now on). In addition, for a small number of regions in the canine genome, the exact nucleotide composition is unknown (called the "Ns" from now on), making bait design difficult. Our next step was to determine if any group contained more of these regions. Upon request, Roche Nimblegen provided us with two BED files containing the regions with no baits directly designed for due to repeats and due to Ns, respectively. For group 3, 275 regions (6.5%) contained Ns and 90 regions (2.1%) repeats. Overall, this results in 364 of the regions (8.6%) being (at least partially) excluded for bait design. For group 2, 1,210 regions (1.4%) have Ns and 17,119 regions (20.3%) contained repeats. Overall, this results in 18,136 of the regions (21.5%) being (at least partially) excluded for bait design. For group 1, no regions contained Ns and 1,798 (1.2%) contained repeats. Overall, group 2 contains the largest proportion of regions (partially) excluded from bait design. As some regions in group 1 were consistently sequenced but were at least partially excluded from bait design, it seems that some regions could still be sequenced efficiently due to the presence of neighboring baits. During the designing process, Roche Nimblegen tries to predict this as well and provided an additional BED file that identifies regions that are predicted not to be sequenced. These regions are a subset of the repeat and Ns regions. We compared their estimates with our results and this showed that 0.1%, 7.2% and 1.6% of the regions in group 1, group 2 and group 3, respectively were predicted not to be sequenced by Roche. For group 1, this result seems to be close to correct. Group 2 contained again the largest proportion of difficult regions.

We also compared the % GC of the remaining regions after 1) exclusion of the regions that Roche Nimblegen predicted not to be sequenced and 2) exclusion of all regions that were at least partially excluded from bait design due to Ns and/or repeats (Figure 5.2, green and red boxes, respectively). This allows us to check whether the % GC of the remaining regions in each group differs from the overall % GC in each group. For group 1, the median % GC of 46.7% remained identical and the proportion of regions with an extreme % GC remained nearly the same (from 4.9% over 4.9% to 5.0%). For group 2, the median % GC increased from 54.9% over 56.0% to 57.9%. The proportion of regions with an extreme % GC increased likewise from 24.4% over 26.0% to 29.5%. For group 3, the median value and proportion of regions with an extreme % GC only increased slightly (from 76.8% over 77.0% to 77.3% and 75.2% over 75.4% to 76.0%, respectively). It seems that in the second group, the remaining regions tend to have slightly higher % GCs, which might negatively influence sequencing.

In the end, we combined the criteria for the % GC (with 32.0% and 64.5% as cut-offs) and the bait design results to determine the total proportion of regions in each group considered to be at risk for reduced sequencing. Due to extreme % GC and/or regions excluded due to Ns/repeats, 6.1%, 44.6% and 78.1% of the regions for group 1, group 2 and group 3, respectively, were identified to be at risk for reduced performance. Due to extreme % GC and/or regions that were predicted not to be sequenced, 5.0%, 31.4% and 75.8% of the regions for group 1, group 2 and group 3, respectively, were identified to be at risk.

Overall, our criteria seem to be relatively correct as they classify the largest proportion of regions at risk in group 3, the second largest in group 2 and only a small amount in group 1. Specifically for group 3, the majority of the regions seems to be insufficiently covered due to extreme % GC. For group 2, the results seem to be a more balanced combination of extreme % GC and bait design.

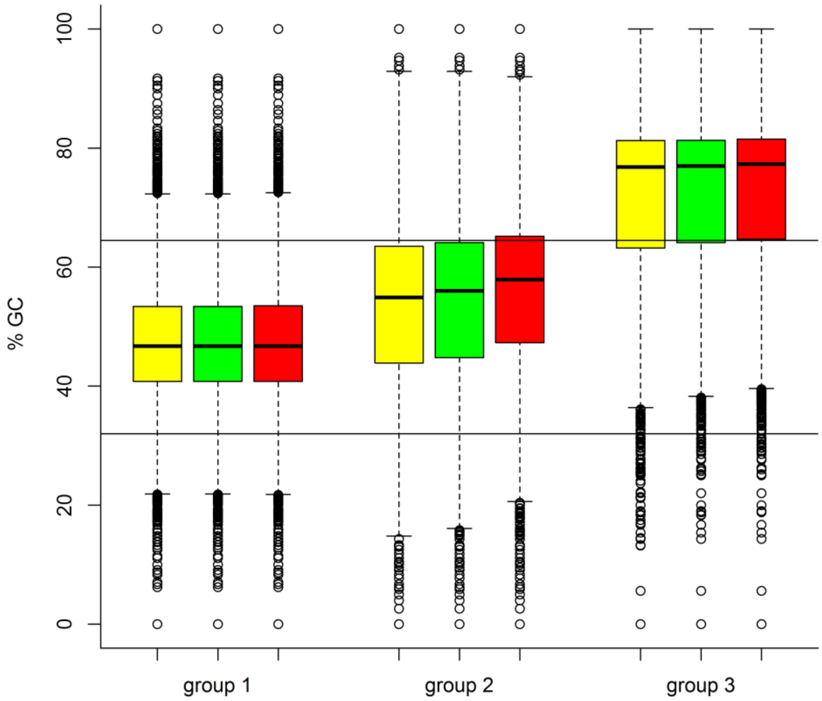


Figure 5.2. Comparison of the GC content per region (% GC) for the completely covered regions with a minimum sequencing depth of 5x (= group 1), the regions with a varying sequencing depth (= group 2) and the regions with a maximum sequencing depth of 4x (= group 3). Each box represents the 25th (Q1), median (Q2) and the 75th (Q3) quartile, the whiskers represent 1.5 times the interquartile range (Q3-Q1). Outliers are represented as circles. Vertical lines represent the cutoffs at 32.0% GC and 64.5% GC. The yellow boxes represent the values for all the regions in a group. The green boxes represent the values for the remaining regions after exclusion of the regions that Roche Nimblegen predicted to not being sequenced. The red boxes represent the values for the remaining regions after exclusion of all the regions with repeats and Ns.

Estimating the coverage for the exome-CDS. Although all samples were sequenced with the exome-plus, we believe we can reliably estimate the performance of the exome-CDS with respect to coverage and reproducibility. This is due to the fact that the exome-CDS is (almost entirely) a subset of the exome-plus. We might underestimate the performance a little bit for the exome-CDS due to the constant number of target baits in each capturing assay. Each target enrichment sequencing assay contains 2.1 million baits. As the exome-CDS is half the size of the exome-plus, twice the number of baits can be used per region. Taking these considerations into account, the following coverage results might be conservative. From the 244,543 regions in the exome-CDS, on average 208,950 regions (85.4%) are estimated to have a sequencing depth of at least 5x throughout the entire region. For on average 9,031 regions (3.7%), the maximum sequencing depth estimated was 4x. In terms of base pairs, we estimate that on average 93.4% (66,586,495 base pairs) of the targeted base pairs are sequenced at least five times. As for the reproducibility, we estimated that in all 16 samples 174,667 regions (71.4%) would be completely sequenced at a minimal sequencing depth of 5x and for 4138 regions (1.7%) the maximum sequencing depth reached would be 4x. In terms of base pairs, 62,455,013 (87.7%) of the base pairs were estimated to be covered consistently in all samples and 2,438,559 (3.4%) base pairs consistently not. The % on target of the exome-CDS was not assessed as a part of the off-target reads for the exome-CDS would actually be on-target reads based on the exome-plus. Including these in the calculation would lead to an underestimation of the % on target.

Variant calling. As WES is often used in disease-association studies, variant detection is an essential part[9]. Overall, between 250,196 and 278,688 non-reference variants were detected inside the targeted regions of the exome-plus (Table 5.4). Filtering for those variants inside the exome-CDS, reduces this number to 110,047 to 122,429 variants (Table 5.4).

Table 5.4

Variants called inside the target regions.

Sample	Pool	Exome-plus (n)	Exome-CDS (n)
1	1	266,334	118,686
2	1	267,695	119,322
3	1	259,499	115,874
4	1	250,196	110,047
5	2	271,834	118,987
6	2	269,445	117,882
7	2	269,995	118,085
8	2	273,081	119,495
9	3	278,462	122,429
10	3	274,222	119,880
11	3	278,688	122,285
12	3	262,793	116,038
13	4	254,919	111,805
14	4	260,454	114,874
15	4	269,313	118,527
16	4	262,786	114,849

Comparison with the exome-1.0: design. A visual comparison between the exome-plus, the exome-CDS and the exome-1.0 can be found in Figure 5.3. Overall, 34.77 Mb are shared between all three designs. Although the vast majority is targeted by the exome-plus, a small number of base pairs is targeted uniquely by the exome-1.0 (0.09 Mb) and the exome-CDS (0.12 Mb).

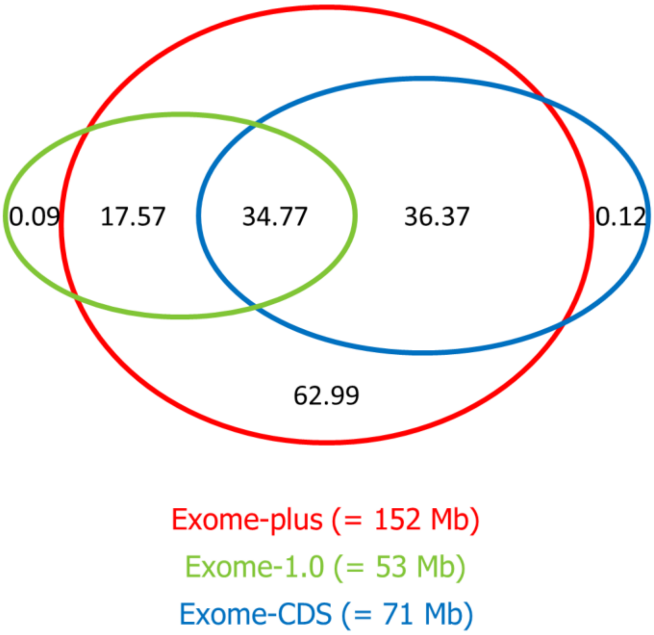


Figure 5.3. Venn diagram showing the overlap between the exome-1.0 (= 53 Mb), exome-CDS (= 71 Mb) and the exome-plus (= 152 Mb). The depicted numbers represent the size in Mb for the various intersections. Overall, 34.77 Mb is shared by all designs. Inside the target space of the exome-plus, the exome-1.0 targets 17.57 Mb more than the exome-CDS and the exome-CDS targets 36.37 Mb more than the exome-1.0. Finally, 0.09 Mb, 0.12 Mb and 62.99 Mb are targeted uniquely by the exome-1.0, the exome-CDS and the exome-plus, respectively.

The difference between the exome-1.0 and exome-plus is attributable to a small number of genes not being shared by Ensembl Genes and the RefSeq Genes and/or mRNA database. The difference between the exome-CDS and the exome-plus is caused by a small number of additional exons and CDS identified by TransDecoder, as described in the design section. Inside the target space of the exome-plus (152 Mb), the exome-plus and the exome-CDS contain respectively 62.99 Mb and 36.37 Mb more compared with the exome-1.0. For the exome-plus, these differences are attributable to the inclusion of all the new protein-coding genes and the non-protein coding regions that were not available when the exome-1.0 was designed. For the exome-CDS, this difference is smaller due to the exclusion of UTRs from the newly discovered proteins. Finally, besides the 0.09 Mb already mentioned, the exome-1.0 contains an additional 17.57 Mb that is not shared with the exome-CDS. This difference is caused by the exclusion of the UTRs from the Ensembl Genes in the exome-CDS. These UTRs are incorporated in the exome-1.0.

Comparison with the exome-1.0: performance. An overall comparison of the average performance parameters of the exome-plus, the exome-CDS and the exome-1.0 can be found in table 5.5. The exome-plus has the lowest scores for the completely covered regions and the region reproducibility. This is attributable to the average size of each individual region. For the exome-plus, ≈ 152 Mb is divided over 242,914 regions, leading to an average size of each region of 624 base pairs. For the exome-1.0, similar calculations lead to an average region size of only 260 nucleotides. If for even one nucleotide in a region, a sequencing depth of 5x is not reached, this region is not covered completely. Theoretically, we can assume that the probability for this to happen is much more likely when the region size increases. This is in agreement with the experimental results: we divided the target regions in those with a length < 260 and ≥ 260 bp. Next, we compared in each subgroup the proportion of regions that were completely covered, with the total number of regions in this subgroup. On average, 25.0% more regions were completely sequenced if the region size was under 260 bp.

Table 5.5

Performance parameters of the exome-plus, the exome-CDS and the exome-1.0.

	exome-plus	exome-CDS	exome-1.0
fully covered regions (%)	79.7	85.4	84.9
base pairs covered (%)	95.1	93.4	90.2
% on target (%)	75.8	-	90.4
reproducibility regions (%)	63.5	71.4	79.9
reproducibility base pairs (%)	90.4	87.7	87.4
non-reference variants (n)	266,857	117,442	61,820

The exome-plus scores highest in terms of base pairs covered and base pair reproducibility. At the same time, the % reads on target is lower in the exome-plus compared with the exome-1.0. These results are explained by the settings applied when the baits were designed. For the exome-1.0, only unique baits were allowed, i.e. baits that only match one location. This is in contrast with the exome-plus that allowed up to 20 matches for each bait. This increases the number of target regions and target base pairs being sequenced at the expense of a lower % reads on target.

The contrasting results of regions and base pairs are due to the combination of the increased region size and the “more matches allowed” bait design settings. Overall, this leads to more regions (and base pairs) being covered for a relatively large proportion, but not completely (Figure 5.1). Compared with the exome-1.0, the exome-plus covers 0.9% regions more for 90%[1].

Intended use and user-specific customization options. With the development of the exome-plus and the exome-CDS, three WES enrichment assays are available for use. The exome-1.0 contains the core set of protein coding genes. As both the exome-plus and the exome-CDS contain many regulatory regions, they are especially valuable in complex disease studies where mutations influencing expression are more likely to be involved[2]. The exome-plus is the design of choice when one needs the most comprehensive capture based on the most recent annotation of the dog genome, including virtually all transcribed regions. The exome-CDS balances completeness and cost-efficiency.

An additional advantage of all three designs, is the ease of customization. Even in the exome-plus there is still room for ≈ 50 Mb of target regions to be added. For example, the few non-targeted RefSeq Genes and/or mRNA regions mentioned earlier or a new update in the non-coding RNA repertoire might be of interest. The regions uniquely identified by TransDecoder might also be added. BED files containing these regions are available on request.

5.4 Discussion

This study describes the development of two new target enrichment designs and the performance of the exome-plus. At a minimum sequencing depth of five, around 80% of the regions were covered completely and well over 90% of the base pairs were covered with a high reproducibility. In addition, a large number of variants were detected. Based on the results of the exome-plus, we estimate the performance of the exome-CDS. Together with the exome-1.0, these designs provide flexible solutions for a variety of research questions.

5.5 Methods

Sample collection. Sixteen canine Labrador Retriever blood samples were obtained from a canine blood bank available at Ghent University to study genetic disorders[10]. Approval was granted by the local ethical (Faculty of Veterinary Medicine, Ghent University, Belgium) and deontological (Federal Public Service Health, Food Chain Safety and Environment, Brussels, Belgium) committees (EC2013_193). All experiments were carried out in accordance with the approved guidelines.

Design. For the exome-plus, from the University Of California Santa Cruz (UCSC) (http://genome.ucsc.edu/) table browser (Dog, CanFam3.1), the Ensembl Genes were selected from the Genes and Gene prediction tracks[11,12]. The output format was a BED file with the setting "exons (plus 0 bases at each end)". MicroRNA sequence positions were downloaded from miRBase[3]. These files were combined with the protein coding genes, antisense transcripts and long non-coding transcripts[2,13]. Regions were merged using bedtools version v2.17.0. The total size of the design was 151,698,592 bp (≈ 6% of the genome) divided over 242,914 regions. For the exome-CDS, all files were identical except for the Ensembl Genes and the protein coding genes. From these 2 files, the CDS were predicted with TransDecoder[4] and selected. The total size of the exome-CDS is 71,254,801 bp (≈ 3% of the genome) divided over 244,543 regions. Both BED files are available on request.

Roche Nimblegen WES enrichment assay. Our design was processed by the Roche Nimblegen custom design group (Madison, USA). Using an

SSAHA algorithm, capturing baits were developed based on our design and the reference genome of the dog (*Canis familiaris* 3.1). Design settings for the baits allowed five or fewer single-base insertions, deletions or substitutions between the baits and the genome. Each bait was allowed to match at maximum up to 20 close matches in the genome. Regions under 100 bp were padded to 100 bp to increase capturing efficiency. After approval, the baits were generated and provided as SeqCap Developer Library.

DNA extraction. Genomic DNA was extracted with the DNeasy Blood & Tissue Kit (QIAGEN) with 100 µl of blood as input. The standard protocol was followed (including the RNAse A step) with the exception of the final elution step: instead of using 200 µl of Buffer AE, only 100 µl was used. The eluate was used again to elute a second and third time to increase the concentration. The DNA yield was measured with Quant-iT^{TM} Picogreen® dsDNA Assay (Life Technologies).

Sample preparation and sequencing. Extracted DNA was fragmented on a Covaris S2 System in a 130 µl volume (aim: 400 bp fragments, settings: duty cycle: 10%, intensity: 4, cycles per burst: 200, time: 55s). After shearing, another picogreen assay was performed. Depending on the yield after DNA-extraction, between 500 ng and 1 µg of the fragmented DNA was used as input for the library preparation. Samples were end repaired, A-tailed and ligated with TruSeq adapters using the reagents from the NEBNext Ultra DNA Library Prep Master mix set for Illumina (New England Biolabs) according to the manufacturer's protocol.

Size selection was performed on a 2% E-Gel (Invitrogen Life Technologies) (G4010-02), fragments were selected with an insert size around 300 bp. One µl of the ligated product was subsequently amplified in an enrichment PCR (10 cycles) for library quality assessment as recommended in the 'SeqCap EZ Library SR User's Guide' (Nimblegen, Roche). Thereafter, the pre-capture LM-PCR was performed on the samples for 8 cycles as prescribed in the SeqCap EZ library protocol. The concentration of each PCR product was determined using Quant-iT^{TM} Picogreen® dsDNA Assay (Life Technologies). Four times four samples were equimolarly pooled to obtain a total DNA input of 1250 ng. The pooled library was hybridized for 67-68 hours with the baits (SeqCap Developer Library). The hybridized library was washed and the captured and pooled DNA was recovered. After a final amplification (LM-PCR, 18 cycles), the quality of the library was checked using the High Sensitivity DNA chip (Agilent).

QPCR. To check the fold enrichment after capturing, a qPCR is performed as a quality control step before sequencing. Five primer pairs were used, as described previously[1]. An additional qPCR was performed to determine the quantity of the library to ensure optimal cluster densities.

Sequencing. Each pool was sequenced in a separate run on the NextSeq 500 PE 75 bp.

Data-analysis. Data-analysis was performed using the CLC Genomics Workbench (Version 7.5.1, CLC Bio, Aarhus, Denmark). Data were trimmed with the following settings: ambiguous trim = no, quality trim =

yes, quality limit = 0.05, use colourspace = no, create report = yes, also search on reversed sequence = yes, save discarded sequences = yes, remove 5' terminal nucleotides = no, discard short reads = no, discard long reads = no, remove 3' terminal nucleotides = no, trim adapter list = adapter list Illumina, save broken pairs = yes. The reference genome was downloaded from the UCSC genome browser[12]. For read mapping, the following parameters were used: mismatch cost = 2, insertion and deletion cost = 3, length fraction: 0.5, similarity fraction = 0.8, global alignment = no, auto-detect paired distances = yes, non-specific match handling = ignore, output mode = create reads track, create report = yes, collect un-mapped reads = yes, colour space alignment = no, masking mode = no masking. Duplicated reads were removed with the Duplicate Mapped Reads Removal (Version 1.0 beta 6) plugin (setting: maximum representation of minority sequence (percent) to 20.0), create a second output file to save the removed reads = yes. Reads were locally realigned with the following settings: realign unaligned ends = yes, multi-pass realignment = 3, guidance-variant track = not set, force realignment to guidance variants = no, output mode = create reads track, output track of realigned regions = yes. Variants were called using fixed ploidy variant detection with the following settings: ploidy = 2, required variant probability = 90.0, ignore positions with coverage above = 100000, minimum coverage = 5, minimum count = 2, minimum frequency = 20.0%, restrict calling to target regions = no, ignore broken pairs = yes, ignore non-specific matches = reads, minimum read length = 20, base quality filter = no, relative read direction filter = yes (significance 1.0%), remove pyro-error variants = no, create track = yes, create table = yes, variant report = yes.

5.6 References

1. Broeckx, B. J. G. *et al.* Development and performance of a targeted whole exome sequencing enrichment kit for the dog (Canis Familiaris Build 3.1). *Sci. Rep.* **4,** 5597 (2014).

2. Hoeppner, M. P. *et al.* An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. *PLoS One* **9,** e91172 (2014).

3. Kozomara, A. & Griffiths-Jones, S. MiRBase: Integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* **39,** 152–157 (2011).

4. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nat. Protoc.* **8,** 1494–1512 (2013).

5. Falk, M. J. *et al.* Mitochondrial disease genetic diagnostics: optimized whole-exome analysis for all MitoCarta nuclear genes and the mitochondrial genome. *Discov Med.* **14,** 389–399 (2012).

6. Clark, M. J. *et al.* Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.* **29,** 908–914 (2011).

7. Aird, D. *et al.* Analyzing and minimizing bias in Illumina sequencing libraries. *Genome Biol.* **12,** R18 (2011).

8. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36,** e105 (2008).

9. Willet, C. *et al.* A catalogue of common canine coding variants. Paper presented at the *7th international conference on advances in canine and feline genomics and inherited diseases*, Cambridge. Cambridge: Purina (2013, September 23-27).

10. Broeckx, B. J. G. *et al.* The Prevalence of Nine Genetic Disorders in a Dog Population from Belgium, the Netherlands and Germany. *PLoS One* **8,** e74811 (2013).

11. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438,** 803–819 (2005).

12. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32,** D493–D496 (2004).

13. Derrien, T. An extended repertoire of long non-coding RNAs in the domestic dog. Paper presented at the *8th international conference on advances in canine and feline genomics and inherited diseases* Cambridge. Cambridge: Purina (2015, June 22-26).

5.7 Supplementary material

Supplementary Table S5.1

Per sample per chromosome percentage reads on target (calculated as mapped reads on target/total number of mapped reads).

Chr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	81.0	80.6	80.8	81.0	77.6	77.4	76.9	77.4	77.5	77.7	77.2	77.2	77.1	77.1	78.0	77.7
2	80.5	80.1	80.2	80.1	76.7	76.6	76.0	76.8	77.1	77.0	76.8	76.7	76.4	76.6	77.6	77.2
3	78.2	77.9	78.2	78.1	74.4	74.2	73.5	74.1	74.3	74.4	73.9	74.0	73.8	73.5	74.6	74.4
4	80.5	80.2	80.5	80.3	77.2	76.9	76.3	77.0	77.1	77.3	76.8	76.7	76.6	76.5	77.7	77.3
5	82.0	82.0	82.3	82.1	79.2	78.8	78.1	79.1	79.2	79.2	79.1	79.0	78.7	78.6	79.8	79.3
6	81.1	80.8	81.2	80.9	77.3	77.1	76.4	77.0	77.2	77.1	76.8	77.1	76.8	76.7	77.8	77.4
7	80.7	80.9	82.0	79.6	76.5	76.0	74.9	76.0	75.8	76.2	76.0	76.2	75.2	74.8	76.6	75.7
8	79.5	79.3	79.6	79.3	76.2	75.8	74.9	75.8	75.9	76.0	75.7	75.9	75.6	75.4	76.5	76.3
9	84.5	84.2	84.6	84.4	81.4	81.0	80.6	81.5	81.5	81.5	81.3	81.4	81.2	81.5	82.4	82.0
10	81.9	81.6	81.9	81.7	78.7	78.4	77.7	78.5	78.7	78.8	78.4	78.4	78.2	78.3	79.3	78.9
11	79.4	79.2	79.4	79.5	76.2	75.9	75.2	75.9	76.2	76.2	75.8	75.9	75.7	75.6	76.7	76.3
12	80.2	79.9	80.2	80.0	76.7	76.5	75.6	76.4	76.6	76.8	76.2	76.3	76.1	76.2	77.1	76.8
13	75.3	75.0	74.9	77.8	74.3	74.1	73.2	74.1	74.2	74.2	73.9	73.9	73.7	73.6	74.8	74.4
14	78.3	78.0	78.5	78.1	74.6	74.3	73.5	74.2	74.4	74.5	73.9	74.0	73.7	73.7	74.8	74.6
15	80.6	80.3	80.6	80.4	76.9	76.6	75.9	76.8	77.0	77.1	76.6	76.6	76.3	76.3	77.6	77.2
16	74.9	74.6	74.6	74.7	71.1	70.8	70.0	71.0	71.3	70.9	70.5	70.7	70.9	70.6	71.8	70.6
17	80.9	80.9	80.6	80.4	77.4	77.0	76.5	77.2	77.7	77.7	77.2	77.2	77.0	77.1	78.2	77.5
18	80.2	80.0	80.3	79.8	77.1	76.6	75.8	76.6	76.8	76.9	76.5	76.6	76.5	76.6	77.3	77.0
19	72.0	71.3	72.4	71.7	68.4	68.3	66.6	67.7	67.5	67.8	67.5	67.4	67.2	66.5	68.4	67.9
20	82.7	82.4	82.7	82.4	79.6	79.2	78.4	79.4	79.5	79.7	79.3	79.5	79.2	79.5	80.1	79.7
21	78.0	77.6	77.6	77.3	74.4	74.0	73.4	74.2	74.4	74.5	74.2	74.3	73.8	73.7	74.8	73.5
22	73.8	73.4	73.8	73.6	70.0	70.0	69.1	69.4	69.6	69.9	69.1	69.4	69.2	68.9	69.9	69.7
23	80.2	80.0	80.3	80.0	76.6	76.2	75.6	76.3	76.5	76.7	76.2	76.3	76.0	75.9	77.0	76.6
24	81.1	81.0	81.4	81.2	77.7	77.5	77.0	77.7	77.9	78.2	77.7	77.8	77.7	77.8	78.8	78.3
25	79.1	78.9	79.3	79.1	75.7	75.4	74.7	75.5	75.7	75.8	75.5	75.4	74.9	74.9	76.1	75.7
26	81.3	80.9	81.3	81.1	77.7	77.5	77.1	77.8	77.8	78.0	77.8	77.8	77.7	77.5	78.6	78.2
27	81.4	81.1	81.3	81.0	77.5	77.1	76.5	77.3	77.5	77.5	77.2	77.2	77.1	77.2	78.2	77.8
28	81.4	81.0	81.4	81.1	77.7	77.3	76.8	77.7	77.7	77.8	77.6	77.5	77.0	77.1	78.6	78.0
29	76.5	76.1	76.4	76.3	72.3	72.0	71.3	72.1	72.1	72.3	71.6	71.7	71.5	71.2	72.5	72.5
30	82.4	82.1	82.6	82.1	78.6	78.1	77.6	78.4	78.7	78.7	78.3	78.3	78.1	78.3	79.5	79.0
31	70.3	69.8	69.6	69.4	65.3	65.1	64.4	64.8	64.9	64.9	64.2	63.5	64.5	63.8	64.8	64.4
32	70.7	69.8	70.3	69.6	62.6	62.7	63.4	63.4	63.2	62.1	62.4	60.5	60.9	59.4	62.9	62.9
33	80.6	80.4	80.7	80.4	76.9	76.5	75.8	76.7	76.8	76.8	76.5	76.5	76.1	76.2	77.6	77.1
34	77.7	77.4	77.6	77.4	73.4	73.2	72.7	73.2	73.2	73.2	72.9	72.8	72.5	72.0	73.6	73.2
35	80.5	80.2	80.4	80.4	77.1	77.0	76.3	76.9	77.1	77.1	76.8	76.9	76.4	76.4	77.4	77.1
36	80.6	80.3	80.7	80.3	76.7	76.2	75.5	76.7	76.7	76.7	76.4	76.2	75.8	76.2	77.7	77.1

37	80.3	80.0	80.2	79.9	75.8	75.5	74.9	75.5	75.3	75.5	75.1	75.3	75.2	75.0	76.2	75.9
38	78.3	77.9	78.2	78.0	74.9	74.5	73.3	74.1	74.5	74.5	74.0	74.2	73.8	74.2	74.6	74.4
X	66.5	66.1	67.8	67.7	63.0	62.9	60.8	61.7	61.9	62.0	62.5	61.6	61.1	61.2	62.6	62.7

6 An heuristic filtering tool to identify phenotype-associated genetic variants applied to human intellectual disability and canine coat colours

B.J.G. Broeckx[a], F. Coopman[b], G.E.C. Verhoeven[c], T. Bosmans[d], I. Gielen[c], W. Dingemanse[c], J.H. Saunders[c], D. Deforce[a*], F. Van Nieuwerburgh[a*]

a Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Ottergemsesteenweg 460, 9000 Ghent, Belgium

b Department of Applied Biosciences, Faculty of Bioscience Engineering, University College Ghent, Valentin Vaerwyckweg 1, 9000 Ghent, Belgium

c Department of Medical Imaging and Small Animal Orthopaedics, Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium

d Department of Medicine and Clinical Biology of Small Animals, Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium

* These authors contributed equally to this work.

BMC Bioinformatics 16 (2015), 391

6.1 Abstract

Background: Identification of one or several disease causing variant(s) from the large collection of variants present in an individual is often achieved by the sequential use of heuristic filters. The recent development of whole exome sequencing enrichment designs for several non-model species created the need for a species-independent, fast and versatile analysis tool, capable of tackling a wide variety of standard and more complex inheritance models. With this aim, we developed “Mendelian”, an R-package that can be used for heuristic variant filtering.

Results: the R-package Mendelian offers fast and convenient filters to analyze putative variants for both recessive and dominant models of inheritance, with variable degrees of penetrance and detectance. Analysis of trios is supported. Filtering against variant databases and annotation of variants is also included. This package is not species specific and supports parallel computation. We validated this package by reanalyzing data from a whole exome sequencing experiment on intellectual disability in humans. In a second example, we identified the mutations responsible for coat colour in the dog. This is the first example of whole exome sequencing without prior mapping in the dog.

Conclusion: We developed an R-package that enables the identification of disease-causing variants from the long list of variants called in sequencing experiments. The software and a detailed manual are available at https://github.com/BartBroeckx/Mendelian.

6.2 Background

The identification of genetic variation responsible for a phenotype, is one of the key aims in the field of genetics. This field has been revolutionized with the introduction of next generation sequencing technologies and is continuously evolving. Although several sequencing platforms exist, the analysis of sequencing data generated in disease-association studies is virtually identical: the platform-specific raw data is used for base-calling and subsequently for mapping and variant calling against a reference genome. These variants can subsequently be used to perform a disease-association analysis, where the typical aim is to identify one or several disease causing variant(s) from the large collection of variants present in an individual. This can be achieved by the sequential application of several heuristic filters[1].

As genetic diseases are heterogeneous, a wide range of filters is required. Compared to complex disorders, it is more straightforward to identify disease causing variants in Mendelian disorders. However, even in this subgroup of Mendelian disorders, a variety of factors might complicate the analysis: different inheritance models (dominant, recessive), de novo mutations, allelic or locus heterogeneity, reduced penetrance, phenocopies, etcetera[1].

Due to the recent development of whole exome sequencing (WES) enrichment designs for several non-model species, these species are likely to be sequenced more often[2-5]. To be of practical use, heuristic filtering software should thus be capable to deal with all the aforementioned situations for both model and non-model species. At this point however, most tools are specifically intended for human analyses (e.g. they only allow processing of files that can be linked to human-specific databases or annotation) and/or only allow the most basic filtering. This limits the broad application of sequencing based approaches as it requires access to bioinformaticians that have to write custom scripts for the analysis at hand. To avoid a constant reinvention of the wheel and to fulfil the need for a species-independent, fast and versatile analysis tool, capable of tackling a wide variety of inheritance models and complicating factors, we developed the R-package "Mendelian". It allows the analysis of several types of variants, including single nucleotide polymorphisms, insertion-deletions and structural variants.

We demonstrate its use with two practical examples. In the first example, we reanalyze the data of a human WES experiment that identified a *de novo* mutation responsible for intellectual disability[6]. The second example demonstrates the power of the combination of the exome-plus, a novel WES design in the dog, and Mendelian by revalidating the recessively inherited yellow and brown coat colour phenotypes in the Labrador Retriever[5,7–9]. This second analysis is also the first to use WES without prior mapping in the dog. The combination of WES and "Mendelian" will likely aid future disease-association studies.

6.3 Implementation

Flexibility of the applied software tool is an important aspect in disease-association studies as the species and phenotype studied might significantly alter the analysis process. For example, filtering steps might be omitted (e.g. when a variant database is not available for the studied species), the proposed inheritance model might be dominant or recessive and genetic heterogeneity might be present. An overview of the features of the tool is provided below. In addition, a detailed vignette is available together with the software package at https://github.com/BartBroeckx/Mendelian.

Input. "Mendelian" allows for the use of the standard variant call format (VCF). In addition, specific .txt output from the commercial platform CLC Genomics Workbench is also supported. If necessary, variant files can be annotated using .bed or .gtf files. The variants can be assigned to a variety of units from standard databases, e.g. an exon or a gene. User-specific custom annotations can also be used.

Filtering against variant databases. Often, the first step in filtering called variants consists of the removal of previously identified variants present in public databases such as dbSNP. This significantly reduces the number of putative variants. Depending on the disease studied, one can choose to use all the variants present in a database or to use only those variants that have a certain minor allele frequency (MAF). This step can be skipped if a dbSNP is not available for the species studied.

Filtering sequencing variants. There are four variant filters to support both dominant and recessive modes of inheritance, filtering at the nucleotide level or at a user-defined level (often an exon or a gene).

They can be applied on one or more affected individuals at once and allow for the inclusion of one or several unaffected control individuals.

The two (dominant and recessive) functions for filtering at the nucleotide level, consider individual variants at a single nucleotide position in the genome. Under a dominant mode of inheritance, no zygosity assumptions are made: every variant called in an affected individual is a putative disease causing variant. Every variant called in unaffected individuals can be used to filter the variants in affected individuals.

Under a recessive mode of inheritance, putative causal variants are assumed to be in a homozygous state. Only homozygous variants in unaffected individuals are used to filter variants in affected individuals.

The two functions for filtering at a user-specified level, consider the variants in a unit (e.g. an exon or a gene) together. This allows for allelic heterogeneity, which implies that different variants within one unit might be disease causing.

Under a recessive mode of inheritance, putative causal variants can both be homozygous and/or compound heterozygous. Compound heterozygosity means that an individual expresses a phenotype due to two different heterozygous alleles within a particular unit. Every unit with at least one homozygous variant or that is compound heterozygous, is retained. If several cases are available, the filter identifies shared units instead of shared nucleotides. Variants called in unaffected individuals are used to filter variants in cases in two consecutive steps. First, homozygous variants in controls are used for filtering. Next, all compound heterozygous variants within a unit are used for filtering.

Under a dominant mode of inheritance, no zygosity assumptions are made, resulting in every unit with at least one variant being retained in affected individuals. Every variant present in a control is used for filtering.

Detectance and penetrance. All four filters allow for a reduced penetrance and reduced detectance. Penetrance is defined as the probability of seeing a certain phenotype, given the genotype. Detectance is defined as the probability of identifying a certain genotype, given the phenotype. A 100% detectance and penetrance is often assumed. Under a reduced detectance, a causal variant can be identified, even under locus heterogeneity or when phenocopies are present. Under reduced penetrance, a causal variant can be present in an individual without the expression of the associated phenotype.

These theoretical definitions are translated into practice by Mendelian in two sequential steps. First, Mendelian calculates the possible detectance and penetrance levels using the following formulas:

$$penetrance = \frac{n_c}{n_c + c_g}$$

$$detectance = \frac{n_c}{n_c + n_d}$$

With for the phenotypically affected individuals:

n_s = {phenotypically “sick” animals (called “cases”)}; n_c ={phenotypically “sick” individuals with a shared (= “common”) genetic cause}; n_d ={ phenotypically “sick” individuals with a different genetic cause or phenocopies} and n_c + n_d = n_s.

and for the phenotypically unaffected individuals:

c = {phenotypically “healthy” animals (called “controls”)}, c_g = {phenotypically “healthy” animals with “sick” genotype}, c_c = {phenotypically “healthy” animals with “healthy” genotype} and c_g + c_c = c. The relation between these abbreviations is depicted in detail in Table 6.1. By varying c_g (restrictions: $0 \leq c_g \leq c$) for the penetrance and n_d (restrictions: $0 \leq n_d < n_s$) for the detectance over all the possible values, the different options are calculated and provided to the user to choose from.

Table 6.1

Relation between a genotype and a phenotype.

Genotype	**Phenotype**	
	Affected	Healthy
Affected	n_c	c_g
Healthy	n_d	c_c
Total	n_s	c

c_g reflects the number of animals that have a reduced penetrance. n_d is the number of animals that have a different genetic cause and/or that are phenocopies. n_c are the animals that share a genetic cause and are phenotypically affected. c_c are the animals that are both genetically and phenotypically healthy. A priori, only n_s and c are known.

After the user has chosen the appropriate levels of detectance and penetrance, c_g and n_c are calculated by rearranging both formulas:

$$c_g = \frac{n_c}{penetrance} - n_c$$

And

$$n_c = detectance\,.\,n_s$$

Practically, Mendelian assumes that under reduced penetrance a variant is allowed to be present in at most c_g phenotypical controls and that under reduced detectance the variant has to be present in at least n_c cases. The chosen penetrance and detectance levels are thus the lower limits, all variants with levels of penetrance and detectance at least as high will be returned by default. This can be adapted, if needed.

6.4 Results and discussion

The output of the heuristic filters is a data frame that for each variant contains the chromosome, the exact location, the allele and the number of samples with that allele. To show the possibilities of "Mendelian", we performed two separate analyses. All R commands used in this analysis are included (see supplementary file 1). All the data reanalyzed in this study was obtained from published studies that were approved by the institution's ethical committees.

Example 1: human intellectual disability. As a starting point, we reanalyzed WES data from a study on intellectual disability[6]. A trio of one affected child and two healthy parents was sequenced and a *de novo* mutation was expected. Trio sequencing has the benefit that the vast majority of variants in the child will be present in at least one of the parents and with a *de novo* mutation, one can additionally assume that the variant has to be heterozygous in the affected child. This allows for an enormous reduction of variants, even though only three samples are sequenced. Two sequential filters were used in our analysis: after preprocessing, the VCF file containing the variants of the patient (patient #3 in the original study) was filtered against a human variant database. In agreement with the original study, the dbSNP135 was used with a

MAF of 0% (i.e. every variant in the database can be used for filtering). This already reduced the number of variants with 72.1%. In the second filtering step the standard dominant filtering at the nucleotide level function was used, but with the "family" option specified. By specifying the "family" option, the parental variants were used to further reduce the number of variants, but with the additional assumption that the putative variant has to be heterozygous in the child. At this point, 99.99% of the variants were excluded and only 5 variants remained. The original de novo mutation on chromosome 17 (chr17:72341086G>A) was one of these 5. In the original paper, the number of variants was further reduced by filtering against a second control population and a Sanger sequencing step. An overview of the analysis is provided in Figure 6.1.

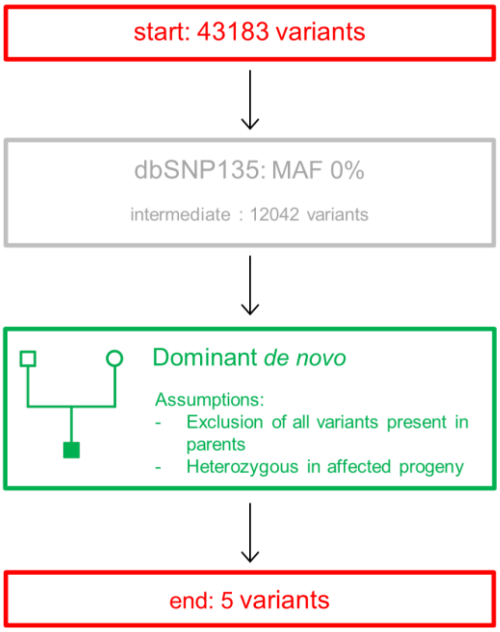

start: 43183 variants
dbSNP135: MAF 0%
intermediate : 12042 variants
Dominant *de novo*
Assumptions:
- Exclusion of all variants present in parents
- Heterozygous in affected progeny
end: 5 variants

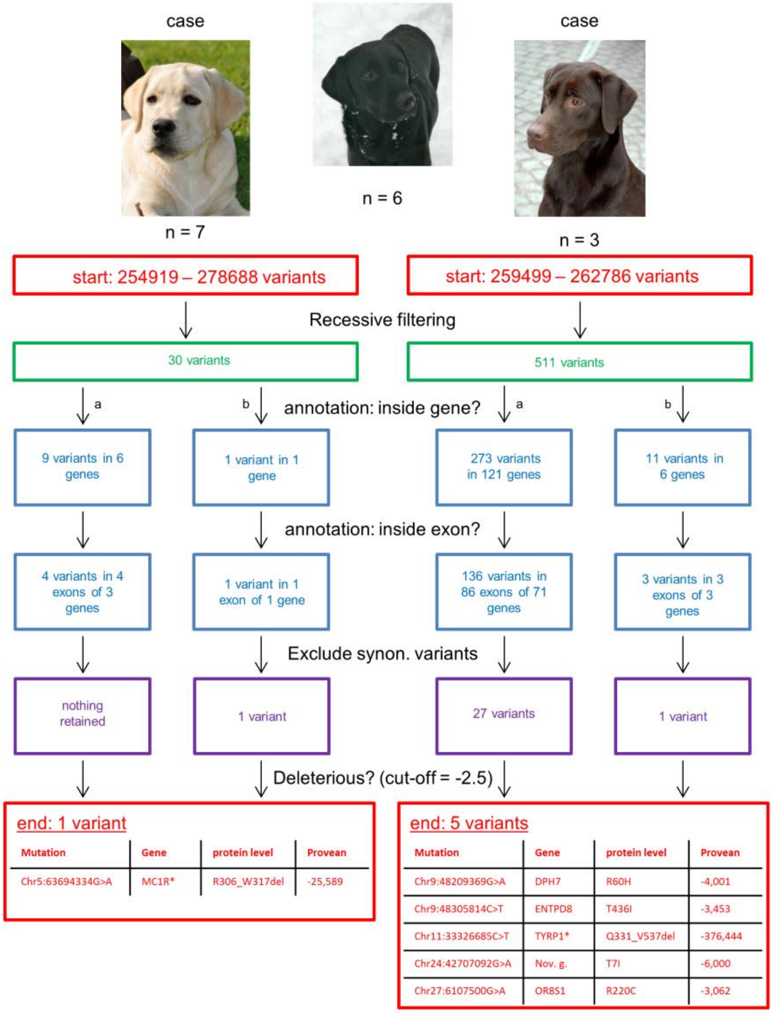
Figure 6.1. Consecutive filtering steps in the identification of putative causal variants for intellectual disability.

Two remarks have to be made when the “family” option is being used. First of all, each family should be analyzed separately. In addition, unrelated controls should not be included with the “family” option specified as the function would consider them to be parents. This would result in additional variants being filtered, based on assumptions that might not be valid.

Example 2: coat colour in the Labrador Retriever. In contrast with human studies, WES is not used frequently in domestic species. One of the reasons is likely the limited availability of WES capturing designs. For the dog, the first report on a WES design was published in 2014. The development of new WES designs, are likely to boost disease-association studies in these species. To demonstrate the power of WES studies combined with “Mendelian”, we revalidated the mutations responsible for the black, brown and yellow coat colour in the Labrador Retriever[7–9]. For this analysis, variant data of 16 dogs that were sequenced to validate the exome-plus design, were used[5]. The analysis is detailed in Figure 6.2. Based on previous reports and the available pedigree data (see supplementary file 2) of the sequenced dogs, it is known that both brown and yellow are inherited recessively as opposed to black[7–9]. For both yellow to black and brown to black, two separate analyses were conducted in parallel. The first step was simple recessive filtering, assuming 100% detectance and 100% penetrance. The analysis was continued by two filtering steps based on annotation: at first, only variants that were inside a gene were retained, followed by a second filtering to retain only those variants within known exons. In the final step, only non-synonymous variants were retained. At this point, only one putative variant remained in the comparison of yellow versus black. For the brown versus black analysis, 27 unique putative variants remained and one of them fell within the exon boundaries of both the Ensembl Genes and the RefSeq genes annotation. Further checking learned that both annotations actually referred to the same gene and that the effect on the protein sequence was identical. The two annotations for that specific variant were thus treated as one.

To further prioritize the putative variants, the analysis was followed by an assessment of the potential effect of the variant at the protein level with Provean[10]. Finally, the variant responsible for the yellow coat colour was identified to be a highly deleterious (Provean score of -25.589) mutation (chr5:63694334G>A) introducing a premature stop codon (R306_W317del in *MC1R*). For the brown coat colour, the variant which corresponds with the known mutation, was predicted to be the most deleterious (Provean score of -376.444). This mutation (chr11:33326685C>T) also results in the introduction of a premature stop codon and removes more than 200 amino-acids from the protein (Q331_V537del in *TYRP1*). None of the other mutations associated with yellow and brown colour in the *MC1R* and *TYRP1* genes were present in any of the dogs[11].

Even with a limited number of dogs, it was possible to identify the mutations responsible for the yellow coat colour and almost to identify the causal mutation for brown coat colour. Importantly, this analysis does not demonstrate the full power of WES for several reasons. First of all, this analysis was conducted without prior filtering to a variant database. For rare disease phenotypes, it is relatively safe to assume that the putative variant has a low MAF in such a database. For a common phenotype such as coat colours, this assumption is not valid and determining an appropriate MAF cut-off will be difficult. In addition, the sequenced dogs, were selected to study orthopedic disorders, not coat colour. Therefore, the case/control selection was not optimized for our analysis. For example, it is much more interesting to include two full siblings with opposite phenotypes than two siblings with the same phenotype (additional variant reduction of 27.6%, chapter 7). Finally, the yellow versus black analysis was somewhat overpowered. A simulation where we gradually included dogs, showed that with 5 yellow dogs and 4 black dogs, we still would have retained the same unique variant (see supplementary file 2).



Mutation	Gene	protein level	Provean
Chr5:63694334G>A	MC1R*	R306_W317del	-25,589

Mutation	Gene	protein level	Provean
Chr9:48209369G>A	DPH7	R60H	-4,001
Chr9:48305814C>T	ENTPD8	T436I	-3,453
Chr11:33326685C>T	TYRP1*	Q331_V537del	-376,444
Chr24:42707092G>A	Nov. g.	T7I	-6,000
Chr27:6107500G>A	OR8S1	R220C	-3,062

Figure 6.2. Sequence of heuristic filters to identify causal mutations for coat colours in the Labrador Retriever. The two analysis (yellow (n = 7) versus black (n = 6) and brown (n = 3) versus black (n = 6)) were performed separately. The annotation steps were split for the Ensembl Genes (a) and the RefSeq genes (b). The potential effect on the protein was predicted with Provean. The default threshold of -2.5 was used as the cut-off value. * = the causal mutations for brown and yellow coat colours, synon. = synonymous, Nov. g. = novel gene (ENSCAFG00000030103).

As the attention shifts towards complex disorders, the question is whether Mendelian can be used for those disorders as well. Complex disorders are in essence no more than a combination of genetic and environmental factors that lead to a reduced penetrance and detectance. As Mendelian allows both reduced penetrance and/or detectance, it should be possible technically. However, lowering the thresholds will also result in less variants being filtered. Overall, the power of Mendelian for complex disorders will probably be lower compared to simple disorders.

6.5 Comparison with existing software

A limited number of different software packages that deal with similar problems are available. Examples are VCFtools[12] and GEMINI[13]. Inside R Bioconductor, the packages VariantFiltering and VariantTools can be used. Compared with these tools, Mendelian has several advantages. GEMINI and VariantFiltering were developed specifically for humans only, which is a disadvantage since WES becomes increasingly popular in a variety of non-model species[2–5]. VariantFiltering does not support multi-allelic variants (variants with more than one alternate allele). Simple analysis tools such as VCFtools and VariantTools only allow for basic analysis (e.g. intersections or complements) and do not support various modes of inheritance[12]. Mendelian is the only package that allows the analysis of variants under reduced penetrance and detectance. To give an idea on the time required when analyzing variant data with Mendelian, some simulations on a standard desktop were added [Additional file 3].

6.6 Conclusions

The identification of one or several causal variant(s) from the vast amount of variant data generated in sequencing experiments, is often based on the sequential use of various filter steps. This software package was designed to provide a species-independent, fast and versatile analysis tool, capable of tackling a wide variety of inheritance models and complicating factors such as genetic heterogeneity and reduced penetrance. We demonstrated its possibilities by reanalyzing a dataset on human intellectual disability and were the first to use WES for the coat colour phenotype in the Labrador Retriever without prior mapping. Overall, this package is a valuable tool for causal variant identification in sequencing studies, especially in non-human species were the alternatives are very limited.

6.7 Availability and requirements

Project name: Mendelian

Project home page: https://github.com/BartBroeckx/Mendelian

Operating system(s): Platform independent

Programming language: R

Other requirements: R version 3.1.0 or higher

License: GPL-2

Any restrictions to use by non-academics: none

6.8 References

1. Stitziel, N. O., Kiezun, A. & Sunyaev, S. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol.* **12,** 227 (2011).

2. Fairfield, H. *et al.* Mutation discovery in mice by whole exome sequencing. *Genome Biol.* **12,** R86 (2011).

3. Robert, C. *et al.* Design and development of exome capture sequencing for the domestic pig (Sus scrofa). *BMC Genomics* **15,** 550 (2014).

4. Broeckx, B. J. G. *et al.* Development and performance of a targeted whole exome sequencing enrichment kit for the dog (Canis Familiaris Build 3.1). *Sci. Rep.* **4,** 5597 (2014).

5. Broeckx, B. J. G. *et al.* Improved canine exome designs, featuring ncRNAs and increased coverage of protein coding genes. *Sci. Rep.* **5,** 12810 (2015).

6. Helsmoortel, C. *et al.* Challenges and opportunities in the investigation of unexplained intellectual disability using family based whole exome sequencing. *Clin. Genet.* **88,** 140–148 (2015).

7. Everts, R. E., Rothuizen, J. & Van Oost, B. a. Identification of a premature stop codon in the melanocyte-stimulating hormone receptor gene (MC1R) in Labrador and Golden retrievers with yellow coat colour. *Anim. Genet.* **31,** 194–199 (2000).

8. Newton, J. M. *et al.* Melanocortin 1 receptor variation in the domestic dog. *Mamm. Genome* **11,** 24–30 (2000).

9. Schmutz, S. M., Berryere, T. G. & Goldfinch, A. D. TYRP1 and MC1R genotypes and their effects on coat color in dogs. *Mamm. Genome* **13,** 380–387 (2002).

10. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS One* **7,** e46688 (2012).

11. Schmutz, S. M. & Berryere, T. G. Genes affecting coat colour and pattern in domestic dogs: A review. *Anim. Genet.* **38,** 539–549 (2007).

12. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).

13. Paila, U., Chapman, B. a., Kirchner, R. & Quinlan, A. R. GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. *PLoS Comput. Biol.* **9,** e1003153 (2013).

6.9 Supplementary material

Supplementary file 1: command line used for example 1 and 2:

Installation of “Mendelian” in R:

```
devtools::install_github("BartBroeckx/Mendelian", build_vignettes=TRUE)
```

By setting build_vignettes to TRUE, the vignette (= manual) is downloaded as well. The vignette can be accessed by typing:

```
vignette("Mendelian-vignette")
```

Example 1:

Input: example of reading in the .vcf files in R:

```
patient1 <- read.table("a")

parent1 <- read.table("b")

parent2 <- read.table("c")
```

Preparing the .vcf files for further processing with a filtering to retain only those variants that passed the quality filters of the GATK pipeline.

```
patient1proc <- VCFfile(patient1, "V10", filter=TRUE, "PASS")

parent1proc <- VCFfile(parent1, "V10", filter=TRUE, "PASS")

parent2proc <- VCFfile(parent2, "V10", filter=TRUE, "PASS")
```

Reading in the dbSNP data:

```
dbSNP <- read.table("dbSNP135", header=TRUE, sep="\t")
```

Preparing the dbSNP for filtering:

a. Deciding how many processors are allowed to be used to prepare the dbSNP:

```
library(doParallel)

registerDoParallel()

nproc <- getDoParWorkers()
```

b. The actual filtering against a variant database with the MAF unspecified:

```
dbSNPfilter <- prepvarpar(dbSNP,,”refNCBI”, nproc)
```

Remark: an unspecified MAF gives the same result as setting MAF = 0, e.g.:

```
dbSNPfilter <- prepvarpar(dbSNP,0,”refNCBI”, nproc)
```

Removing all the variants present in the dbSNP from the variants in the patient:

```
filtered <- varfilter(patient1proc,dbSNPfilter)
```

Trio filtering:

```
nDom("filtered", c("parent1proc", "parent2proc"), "Ps-F")
```

Example 2:

Input: example of reading in of one .txt output file from CLC Genomcics Workbench in R:

```
a <- read.table("a", header=TRUE, sep="\t")
```

Standard recessive filtering (with three cases and three controls):

```
b <- nRec(c("a", "b", "c”), c("d", "e", "f"))
```

Annotation:

reading in the RefSeq Genes annotation as downloaded from the UCSC table browser:

```
RefBED <- read.table("bed", sep="\t", header=FALSE)
```

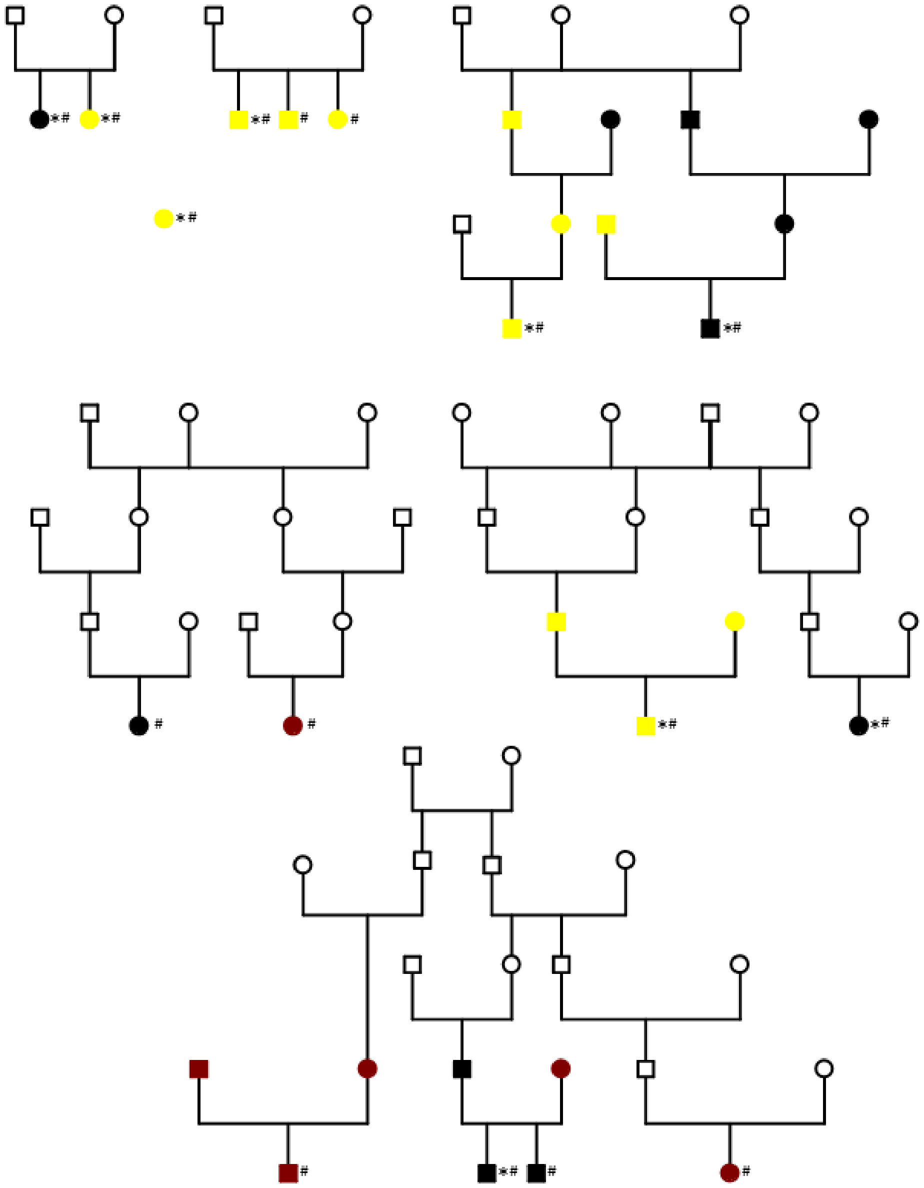
the actual annotation process with removal of all variants that do not fall inside the RefSeq regions

```
out <- annot(a,RefBED, type="BED", nomatch=FALSE, CLC=TRUE)
```

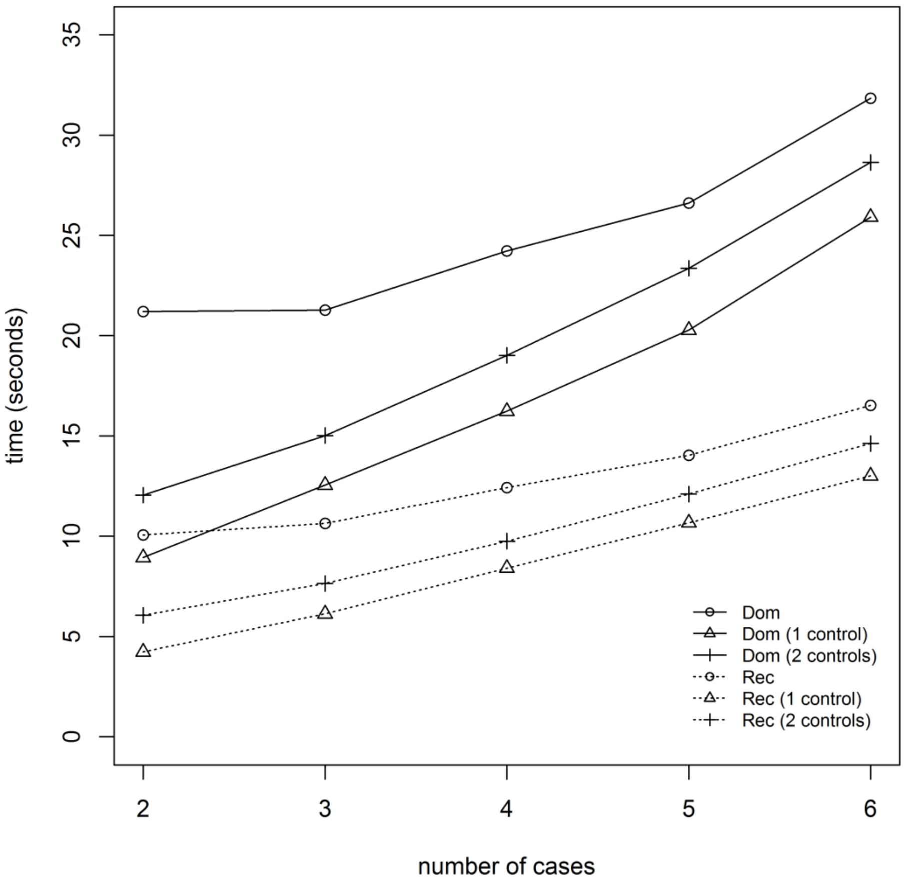
Additional examples for each function are provided together with the installation of the R-package. They can be accessed by combining “?” with the function you require information for:

Example:

```
?annot
```



Supplementary file 6.2. pedigree data of the dogs used in the coat colour analysis. In this figure, the familial relation between the dogs used in the analysis, is shown. The colour of the squares and circles corresponds with the coat colour of the dog (yellow, brown or black). If the coat colour is not known, an empty black circle or square was used. □ = male, ○ = female, # the dogs used in the general analyses, * the 5 yellow dogs and 4 black dogs needed to retain only one variant.



Supplementary file 3. Time duration required for processing a variable number of cases and controls with the dominant (Dom) and recessive (Rec) filter (at the nucleotide level) used in example 1 and 2. Even though each dog had well over 250000 variants, the analysis only took at most around 30 seconds on a standard desktop (Intel(R) Core(TM) i3-2100 CPU @ 3.10GHz, 4.00 GB RAM, 32-bit Windows 7). The inclusion of controls decreases the computing time through a reduction of the number of variants in the cases. The recessive filter outperforms the dominant filter here as the size of the data frames is reduced by the exclusion of heterozygous variants.

7 Toward the most ideal case-control design with related and unrelated dogs in whole exome sequencing studies

B.J.G. Broeckx[a], F. Coopman[b], G.E.C. Verhoeven[c], S. De Keulenaer[a], E. De Meester[a], V. Bavegems[d], P. Smets[d], B. Van Ryssen[c], F. Van Nieuwerburgh[a*], D. Deforce[a*]

a Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Ottergemsesteenweg 460, 9000 Ghent, Belgium

b Department of Applied Biosciences, Faculty of Bioscience Engineering, University College Ghent, Valentin Vaerwyckweg 1, 9000 Ghent, Belgium

c Department of Medical Imaging and Small Animal Orthopaedics, Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium

d Department of Medicine and Clinical Biology of Small Animals, Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium

* These authors contributed equally to this work.

Accepted: Animal Genetics (in press)

7.1 Abstract

With the recent development of whole exome sequencing enrichment designs for the dog, a novel tool for disease-association studies became available. The aim of disease-association studies is to identify one or a very limited number of putative causal variants or genes from the large pool of genetic variation. To maximize the efficiency of these studies and to provide some directions of what to expect, we evaluated the effect on variant reduction for various combinations of cases and controls for both dominant and recessive types of inheritance assuming variable degrees of penetrance and detectance. In this study, variant data of 14 dogs (13 Labrador Retrievers and 1 Dogue de Bordeaux) obtained by whole exome sequencing, were analyzed. In the filtering process, we found that unrelated dogs from the same breed share up to 70% of their variants, which is likely a consequence of the breeding history of the dog. For the designs tested with unrelated dogs, combining 2 cases and 2 controls gave the best result. These results were improved further by adding closely related dogs. Reduced penetrance and/or detectance has a drastic effect on the efficiency and is likely to have a profound effect on the sample size needed to elucidate the causal variant. Overall, we demonstrated that sequencing a small number of dogs, results in a marked reduction of variants that is likely sufficient to pinpoint causal variants or genes.

7.2 Introduction

The development of a technique that enables selective capturing of exons in 2007, announced the era of whole exome sequencing (WES)[1]. The practical use of this technique was soon demonstrated by the discovery of a mutation responsible for Miller syndrome[2]. Since then, WES has been widely used in disease-association studies in the human population. Gradually, WES designs also became available for several domestic species like the mouse and the pig[3,4]. In 2014, the first report on a WES design, called the exome-1.0, in the dog was published[5]. Typically, WES aims to selectively sequence all the regions that are transcribed to mRNA. However, many disease causing mutations are likely to fall outside these protein-coding regions. This was at least partially resolved by the recent development of two novel canine WES designs, the exome-plus and the exome-CDS. In addition to novel protein-coding regions, these designs target microRNAs, long non-coding RNAs and antisense transcripts from the recently published improved annotation[6,7]

Although WES is much cheaper compared to whole-genome sequencing, it remains relatively expensive. A priori selecting the optimal combination of samples is likely to improve variant reduction, increase cost-efficiency and thus the overall chance of successfully identifying causal mutation(s) in disease-association studies. This study evaluates various combinations of cases and controls, inheritance types and familial relatedness to determine the most efficient design for sample sizes up to 4 and can be used as a guideline for WES studies in the dog. As the goal in disease-association studies is filtering genetic variation until only one or a very limited number of putative causal variants or genes remains, we compared the efficiency of each design in terms of the proportion of variants from the case(s) that could be excluded. The most efficient designs are those where the smallest amount of variants were retained.

7.3 Materials and Methods

Animal selection. Dogs were selected from two prior whole-exome sequencing experiments[5,7]. Inclusion criteria were 1) the presence of a pedigree (at least up to four generations), 2) besides the familial degrees studied (parent – progeny, full sibs, half sibs) no additional ancestors were allowed to be shared to avoid bias due to relatedness. So called "unrelated dogs" did not share any ancestor in the four generation pedigree.

Exome-enrichment and data-analysis. Fourteen dogs were selected based on their known familial relationships (Figure 7.1). Details of the library preparation, sequencing and data-analysis are provided in the original papers[5,7]. Briefly, after shearing, samples were end repaired, A-tailed and ligated according to the manufacturer's protocol. Prior to each enrichment, four samples were equimolarly pooled. Each exome-1.0 pool (target size ≈ 53 Mb, CanFam 3.1) was sequenced in one lane on a HiSeq 2500 (PE 100 bp) whereas each exome-plus pool (target size ≈ 152 Mb, CanFam 3.1 and updated annotation) was sequenced in a separate run on the NextSeq 500 PE 75 bp. The data-analysis was performed with the CLC Genomics Workbench and consisted out of quality filtering, read mapping, duplicate read removal, local realignment and probabilistic variant calling (minimum coverage = 5). The average coverage for the eight samples sequenced with the exome-1.0 was 102x and 70x for the six samples sequenced with the exome-plus. Detailed sequencing statistics are available in Table S7.1.

Filtering. In concordance with standard practices, only the autosomal, non-synonymous variants were retained from each animal[8]. Mitochondrial variants and variants on the X chromosome were thus removed. The number of variants and functional units for each animal prior to filtering are shown in Figure 7.1. The filtering was performed with the R-package "Mendelian" (Chapter 6). The assumptions for each filter are detailed in Table 7.1. Four different filters have been used: two of them filter at the nucleotide level, the other two filter at the level of a unit. The two (dominant and recessive) functions for filtering at the nucleotide level, consider individual variants in the genome. Under a dominant mode of inheritance, every variant called in an affected individual is a putative disease causing variant whereas the recessive filter assumes that causal variants are homozygous in cases. All the variants called in controls or only those that are homozygous are used to remove variants in cases for the dominant and recessive filters, respectively.

Filtering units instead of individual variants allows the presence of allelic heterogeneity (different variants within one unit might be disease causing). As mentioned, non-synonymous variants were the starting point for the analysis. This means that, based on the known annotation, the effect of a variant at the protein level was determined. A logical choice for the functional unit is thus the protein: a functional unit is only retained if case(s) have variant(s) with an effect on the amino-acid sequence of the same protein. Under a dominant mode of inheritance, every unit with at least one variant is being retained in cases. Under a recessive mode of inheritance, every unit with at least one homozygous variant or that is compound heterozygous, is retained. If several cases are available, the filter identifies shared units instead of shared nucleotides. All variants called in controls are used to remove variants in cases when the dominant filter is used. When the recessive filter is used, the variants in controls are used in two consecutive steps. First, homozygous variants in controls are used for filtering. Next, all compound heterozygous variants within a unit are used for filtering.

Table 7.1

Assumptions during heuristic filtering under a dominant and recessive mode of inheritance for single variants (= no allelic heterogeneity) and for functional units (= allelic heterogeneity allowed). In this study, a protein is the functional unit of choice.

filter	retained when shared by cases	which variants called in control(s) can be used to remove variants from case(s)
Dominant		
variants	every variant	every variant
proteins	every protein ≥ 1 variant	every variant
Recessive		
variants	homozygous variants	homozygous variants
proteins	every protein ≥1 homozygous variant and/or ≥1 compound heterozygous	homozygous variants and/or pairwise combination of heterozygous variants in each protein

Instead of the actual number of variants that were retained, we calculated the proportionate reduction for each case. The rationale is that the number of variants retained will vary considerably depending on the number of variants initially obtained, whereas the proportionate reduction is expected to give more generally applicable results. This proportionate reduction was termed efficiency and is calculated as:

$$\frac{1}{n \times p} \sum_{j=1}^{p} \sum_{i=1}^{n} (1 - \frac{nvar_{ij1}}{nvar_{ij0}})$$

with:

- $nvar_{ij0/1}$ = the number of variants for case i in permutation j after filtering (1) and before filtering (0)

- n = the number of cases

- p = the number of possible permutations

A stepwise approach was applied. In the baseline analysis, only unrelated animals were selected. To avoid any bias due to relatedness, one individual from every family was selected randomly for all comparisons (Figure 7.1, individuals with *). All possible permutations were used to evaluate the efficiency. In the second analysis, various familial designs were evaluated. The third analysis combined the most efficient design for a sample size of three and four, as identified in the first analysis, with the most efficient familial combinations from the second analysis. Finally, one of the most efficient designs overall was used to assess the effect of reduced penetrance and detectance. Penetrance and detectance are defined respectively in literature, as[8]:

P (phenotype I genotype)

P (genotype I phenotype)

In theory, detectance and penetrance can vary between 0 and 100%. When having sequenced a number of individuals, this continuous variable turns into a discrete variable with a limited number of thresholds, based on the following formulas:

$$penetrance = \frac{n_c}{n_c + c_g}$$

$$detectance = \frac{n_c}{n_c + n_d}$$

With for the phenotypically affected individuals:

n_s = {phenotypically "sick" animals (called "cases")}; n_c ={phenotypically "sick" individuals with a shared (= "common") genetic cause}; n_d ={ phenotypically "sick" individuals with a different genetic cause or phenocopies} and n_c + n_d = $n_{s.}$

and for the phenotypically unaffected individuals:

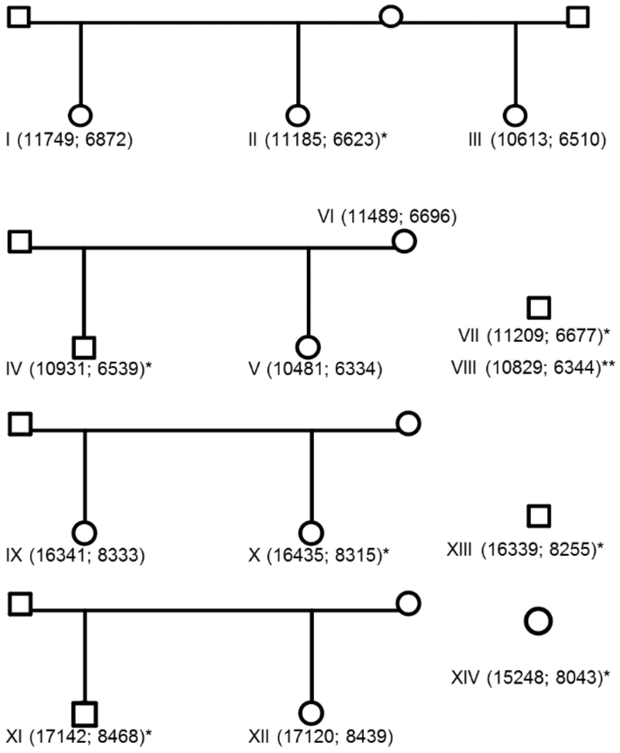
c = {phenotypically "healthy" animals (called "controls")}, c_g = {phenotypically "healthy" animals with "sick" genotype}, c_c = {phenotypically "healthy" animals with "healthy" genotype} and c_g + c_c = c. Only n_s and c are known a priori. Depending on the chosen detectance and penetrance levels, the other variables vary within the restrictions mentioned.

For analysis 2-4, the samples were permutated taking the familial relatedness into account. Throughout the manuscript, all comparisons are presented as x vs y with x the number of cases and y the number of controls.

7.4 Results

A total of 14 dogs were selected based on their known familial relationships (Figure 7.1). Thirteen of these fourteen dogs were Labrador Retrievers, one dog was a Dogue de Bordeaux.

Figure 7.1. Pedigree detailing the relationship for the 14 dogs (# of variants; # of proteins). Dog I to VIII were sequenced using the exome-1.0, dog IX to XIV with the exome-plus. ＊ denotes the dogs selected in the unrelated dogs analysis; ＊＊ is the only Dogue de Bordeaux, the other ones are Labrador Retrievers; □ = male, ○ = female.



The baseline: increasing the sample size by including unrelated individuals. The effect of the following parameters was assessed in unrelated individuals: the sample size, the proposed type of inheritance, the functional unit of choice (a single variant or a protein) and the choice to include additional case(s) or control(s) (Table 7.2A and B). Overall, an increase in sample size always results in less (albeit a single mutation or proteins) being retained. For all sample sizes, recessive inheritance is more efficient compared to dominant inheritance and the assumption of one identical variant being shared is more efficient than assuming a shared protein. A balanced design (with equal number(s) of case(s) and control(s)) outperforms all the other designs, but is followed closely by designs with several cases and one control.

For a sample size of two and three, we compared the exome-1.0 and the exome-plus enrichment designs directly. In general, the results are similar, with the maximum difference being 7.1% for the variants (2 vs 0, 62.3% for the exome-1.0 versus 69.4% for the exome-plus) and 8.5% for the proteins (1 vs 2, 73.7% for the exome-1.0 versus 65.2% for the exome-plus). For both the exome-1.0 and the exome-plus enrichment designs, the same combinations of cases and controls are designated as being the most efficient (for n = 2, 1 vs 1; for n = 3, 2 vs 1; for n = 4, 2 vs 2).

The effect of familial relatedness on variant reduction. The effect of the inclusion of individuals with a variable degree of familial relatedness was assessed next (Table 7.3). The more closely related the individual, the more pronounced the effect is in terms of variant reduction if the included individual is a control (1 vs 1). If the individual is a case as well, less related individuals have the benefit (2 vs 0). The most extreme example of unrelatedness in this study is the inclusion of a dog from a different breed (a Dogue de Bordeaux). Combining dogs from two different breeds results in the highest reduction if both dogs are cases (2 vs 0, e.g. 40.3% > 26.4% > 20.6%) and the lowest reduction if one of them is a control and one of them is a case (1 vs 1, e.g. 58.9% < 73.6% < 80.6% < 86.8%).

Table 7.2
The average efficiency for reduction of variants and proteins for various combinations of cases and controls (x vs y) from one breed. All animals were unrelated up to four generations. Two exome designs were used: the exome-1.0 (= 1.0) and the exome-plus (= plus).

A	n = 2				n = 3					
	2 vs 0		1 vs 1		3 vs 0		2 vs 1		1 vs 2	
Dominant	1.0	Plus	1.0	Plus	1.0	Plus	1.0	Plus	1.0	Plus
variants	31,3	35,8	68,7	64,2	43,5	47,7	87,8	88,2	80,9	76,0
proteins	21,2	22,4	59,7	51,9	30,8	31,9	79,8	76,0	73,7	65,2
Recessive										
variants	62,3	69,4	80,1	84,5	69,3	74,8	93,0	94,7	87,1	89,7
proteins	42,1	45,3	65,4	64,7	49,5	52,2	82,5	81,5	76,0	74,1

B	n = 4			
	4 vs 0	3 vs 1	2 vs 2	1 vs 3
Dominant	Plus	Plus	Plus	Plus
variants	54,1	93,5	94,6	81,4
proteins	37,5	84,2	86,4	71,8
Recessive				
variants	77,7	97,1	97,6	92,3
proteins	56,3	87,4	89,0	78,9

Combining related and unrelated individuals. Based on the first two analyses, the most efficient designs are likely to be those with a closely related family member as a control and an unrelated individual as an additional case. Based on these results, additional designs were evaluated further (Table 7.4).

For a sample size of three, the most efficient combination was a 2 vs 1 design. Three combinations were evaluated: a combination of two full sibs (two cases) with one parent (control), two full sibs (one case, one control) and an unrelated case, one descendant (case) with one parent (control) and an additional unrelated case. The results for all three designs were comparable with the maximum difference being only 2.4% (2 vs 1, Fs$_a$, U vs P$_a$ = 94.7% and Fs$_{1a}$, U vs Fs$_{2a}$ = 92.3%).

Table 7.3

The average efficiency for reduction of variants and proteins for various combinations of cases and controls (*x* vs *y*) for varying degrees of familial relatedness. Two exome designs were used: the exome-1.0 (= 1.0) and the exome-plus (= plus). Parents were included as a control only, not as a case.

	progeny - parent	full sibs				half sibs		unrelated dog	
	1 vs 1	2 vs 0		1 vs 1		2 vs 0	1 vs 1	2 vs 0	1 vs 1
Dominant	1.0	1.0	Plus	1.0	plus	1.0	1.0	1.0	1.0
variants	86,8	20,6	24,0	80,6	76,0	26,4	73,6	40,3	58,9
proteins	77,0	12,5	14,7	74,5	65,4	17,6	65,9	28,2	47,7
Recessive									
variants	91,6	55,8	63,0	87,3	90,6	61,6	83,3	66,5	75,5
proteins	76,8	35,1	38,9	77,2	74,4	41,1	69,6	48,1	57,3

For a sample size of four, the most efficient combination was a 2 vs 2 design. We evaluated two combinations: two pairs of full sibs, with one individual of each pair as a control and the other one as a case and one pair of full sibs with one parent and one descendant were evaluated. Overall, these designs were the most efficient in terms of variant reduction with almost no variants/proteins being retained.

Complicating factors: reduced penetrance and detectance. The 2 vs 2 design with a combination of 2 pairs of full sibs (one individual of each pair being a control, the other one being a case) was used to evaluate the effect of reduced penetrance and reduced detectance (Table 7.5).

For the detectance, two cut-offs can be evaluated in this design: 50 and 100%. A detectance of 50% means that a variant/protein present in at least one of the two cases is putatively causal. As the formula for penetrance takes n_c into account, the calculated percent penetrance depends on the choice of the detectance level. For a detectance of 100%, three cut-offs can be evaluated for penetrance: 100%, 67% and 50%. For a detectance of 50%, the three cut-offs are 100%, 50% and 33%. Practically however, the result is the same: reducing the penetrance one level, means that a putative causal variant/protein shared is also allowed to be present in up to one of the two controls. Dropping it two levels means the putative variant/protein is allowed in both controls. The baseline to compare these results is 100% detectance and 100% penetrance (Table 7.4). Incomplete penetrance or detectance has a large effect on the efficiency. Reducing either one of them with one step,

Table 7.4

The average efficiency for reduction of variants and proteins for a 2 cases vs 2 controls (2 vs 2) and a 2 cases vs 1 control (2 vs 1) design by combining individuals with varying degrees of familial relatedness. Two exome designs were used: the exome-1.0 (= 1.0) and the exome-plus (= plus). For each design, an example is given. Abbreviations: Fs1/2a/b = full sib pair with the respective individual 1 or 2 from family a or family b; Pa = parent of individual Fsa in family a; U = unrelated dog from the same breed.

	2 vs 2			2 vs 1			
	Fs_a,Fs_{1b} vs P_a,Fs_{2b}	Fs_{1a},Fs_{1b} vs Fs_{2a},Fs_{2b}		Fs_a,U vs P_a	Fs_{1a},Fs_{2a} vs P_a	Fs_{1a},U vs Fs_{2a}	
Dominant	1.0	1.0	plus	1.0	1.0	1.0	Plus
variants	98,9	98,1	97,8	94,7	93,7	92,3	93,0
proteins	96,2	96,0	92,0	88,2	87,1	86,5	83,0
Recessive							
variants	99,4	98,6	99,1	97,3	96,3	95,9	97,1
proteins	95,7	95,5	93,5	88,7	87,7	88,6	87,0
example	IV,I vs VI,II	IV,II vs V,I		IV,II vs VI	IV,V vs VI	IV,II vs V	

reduces the efficiency at its best to the same level of a 2 vs 1 design of unrelated individuals. The most extreme case is a penetrance of 33% and detectance of 50% under a dominant inheritance: with every variant/protein being putative disease causing for each case and no variant/protein in a control that can be used for filtering, we end up with more than we initially started.

Table 7.5

The effect of reduced penetrance (P) and/or detectance (D) on the average efficiency of reduction of variants and proteins for the 2 cases vs 2 controls design with two pairs of full sibs, where one individual of each pair is a control and the other one a case. Two exome designs were used: the exome-1.0 (= 1.0) and the exome-plus (= plus).

	100% D				100% P		Reduced D and P			
	67% P		50% P		50% D		50% P, 50% D		33% P, 50% D	
Dominant	1.0	plus	1.0	plus	1.0	plus	1.0	plus	1.0	plus
variants	86,8	88,1	33,3	38,5	78,3	68,2	31,1	20,7	-33,6	-38,6
proteins	75,4	71,3	22,8	23,6	71,3	55,5	23,0	13,1	-23,0	-23,6
Recessive										
variants	93,2	95,1	63,9	70,1	84,0	87,8	56,9	65,8	22,2	36,8
proteins	75,6	75,3	43,4	46,0	72,4	67,5	34,1	36,0	5,3	11,3

7.5 Discussion

This study evaluates the effect of various combinations of individuals for different sample sizes, types of inheritance, functional units, familial relatedness and penetrance and detectance levels in the dog. A baseline to compare the other designs with was established by comparing the effect of the inclusion of several unrelated individuals from one dog breed, the Labrador Retriever. As the general dog population went through several bottlenecks (domestication and breed creation), a high degree of shared variants was expected[9,10]. For the Labrador Retriever, this resulted in close to 70% of the variants being shared between two unrelated dogs (Table 7.2). These results are likely to vary between breeds as they reflect among other the population history. Even within one breed, our results will vary somewhat. These differences are likely due to differences between the dogs. The analysis pipeline, the choice for whole exome or whole genome sequencing and the exome capturing design used, are not likely to affect the proportionate reduction. However, they do influence the exact number of variants being retained eventually. To provide some guidance to what variation one might expect, we analyzed two different sets of Labrador Retrievers with two different exome designs (exome-1.0 and exome-plus) with two different analysis pipelines. As the results for both analyses are rather similar, these estimates can be considered relatively reliable. A remark is that the difference between the two analyses is sometimes larger than the difference between designs (e.g. 2 vs 1 and 1 vs 2 in Table 7.2). However, inside each analysis, the same combination of cases and controls was the most efficient. If one has a choice, choosing the most optimal design for a certain sample size will always pay off. A joint analysis of exome-1.0 and exome-plus sequenced dogs was not conducted here, even though it would have given us the chance to analyze designs with n > 4. The reason is that, as the exome-plus and exome-1.0 WES enrichment designs are not the same, neither in size, nor in targets, mingling them would affect the variant count and the proportionate reduction: e.g. filtering a case sequenced with the exome-plus, using a control sequenced with the exome-1.0, will always result in at least 5377 variants and 1734 proteins being retained as this is the average difference in variants/proteins at the start. In general, combining different enrichment designs should always be done with caution and it is best to decide beforehand what is likely to be the most appropriate enrichment design.

The majority of the presented analyses were conducted in one breed, the Labrador Retriever. The question is whether the presented results can be extrapolated directly to these other breeds. As each breed has its own population history, the baseline and thus the other efficiency numbers will vary between breeds. However, the trends that were observed when the different combinations of cases and controls were compared, will remain. The reason is that for all the analyses conducted in one breed, the breed-specific genetic background remained the same. The observed differences between the various case-control designs originated thus from the parameters that were varied and not from the breed. A design that is more efficient in the Labrador Retriever will thus also be more efficient in other breeds, but the exact proportionate reduction will differ. In general, we expect the results presented here to be rather conservative (i.e. underestimated) as the population size of the Labrador Retriever is relatively big compared to other breeds[9,10].

It is generally accepted that disease-association studies for recessive diseases are more likely to be successful compared with the dominant ones[8,11]. The advantage of recessive inheritance lies in the fact that homozygous variants are easier to detect compared to heterozygous variants and the a priori exclusion of variants or functional units that are not homozygous and/or compound heterozygous in cases. In general, this trend is confirmed here, although with increasing sample sizes and more efficient combinations of cases and controls, the difference tends to diminish. We did identify three exceptions in this study where dominant was more efficient than recessive: the 1 vs 1 design (parent – progeny, Table 7.3), and the 2 vs 2 design for parent-progeny and full-sibs (Table 7.4). These differences are very small with the maximum difference being only 0.5%. It can be expected however that at some point, with increasing numbers of cases and controls, the dominant model of inheritance will reduce variants more efficiently than the recessive model. This will not be the case if only affected individuals are included, but it will occur when controls are included. The reason is that when assuming recessive inheritance, only homozygous or compound heterozygous variants present in controls can be used to filter variants in cases. For a dominant type of inheritance, every variant present in controls can be used for filtering. The balance favoring recessive inheritance, is thus likely to tilt towards dominant inheritance when several cases and controls are combined. Overall, the difference became negligible in the most efficient 2 vs 2 designs (Table 7.4).

Due to the high relatedness of “unrelated” dogs from one breed, designs including one control are always preferred over designs with the same sample size but with cases only. Even the most extreme case of unrelatedness presented in this study (= a dog from a different and relatively distinct breed) favors inclusion as a control instead of an additional case. Considering that even unrelated dogs are highly similar and favor inclusion as a control, the inclusion of more closely related individuals should result in a more efficient variant reduction. This trend is visible in Table 7.3. Although full sibs and one parent vs a descendant are mathematically expected to be equally related, the inclusion of at least one parent has some additional benefits. Under a recessive mode of inheritance with no genetic heterogeneity and with healthy parents, you can assume additionally that the parent has to be heterozygous for the causal mutation. Under the same conditions, but for a dominant mode of inheritance, it can be assumed that the causal mutation is heterozygous in the affected descendant. Under these assumptions, including a parent is preferred over a full sib. If allelic heterogeneity is present, the advantage of parents over full sibs is non-existent. This is confirmed by our experiments (Table 7.3).

One downside of this study is that we did not have access to two parents, so no results for "trio sequencing" (a combination of a healthy mother, healthy father and an affected child) could be provided. Trio sequencing is a very efficient design due to the additional assumptions of a mutation to be de novo under a dominant mode of inheritance or the requirement of heterozygosity in both parents under a recessive mode of inheritance. Based on previous studies, we estimate the number of retained variants to be between ten and three hundred depending on the amount of prior filtering and additional assumptions made[12,13]. If these numbers are directly transferred to our data, this would mean a reduction between 98.2% and 99.9%. This would make trio sequencing the most efficient design for a sample size of 3 and at least as efficient as the other designs tested here with a sample size of 4.

The 2 vs 2 design is the most efficient one, especially if closely related individuals are combined (Table 7.2 and 7.4). In our study, up to 99% of the variants and 96% of the proteins were removed with only 4 samples, which is the usual number of samples pooled prior to capturing[5,7]. An important remark is that these results were obtained without prior filtering to a variant database and without additional filtering based on predicted deleteriousness by tools as PolyPhen and Provean[14,15]. These steps were omitted here to avoid retaining no single variant or protein as this would make comparisons on the efficiency impossible. However, these steps are standard in most disease-association studies[8]. Therefore, it is likely that a sample size of 4 is sufficient in standard studies, if detectance and penetrance are 100%.

The effect of reduced penetrance and detectance was evaluated for the 2 vs 2 design of 2 pairs of full sibs as this was one of the most efficient designs until now (Table 7.5). Reducing either of them immediately decreases efficiency. The most extreme design (33% penetrance and 50% detectance under a dominant mode of inheritance) results in even more variants and proteins being retained, since all the variants from all the affected individuals are just added together and no filtering occurs. Under these assumptions, proceeding to sequencing would be futile. For a recessive mode of inheritance, this extreme design still results in a limited reduction. This is a consequence of the assumption of compound heterozygosity and/or homozygosity in the individual affected cases. These results demonstrate that a reduced penetrance and/or detectance requires a dramatic increase of the sample size. Unfortunately, it is not unlikely for this situation to occur. Reduced penetrance is reported quite often. Some common examples in the dog are the environmentally influenced exercise-induced collapse (affected by physical activity) and the age-dependent reduced penetrance of degenerative myelopathy[16,17]. Reduced detectance has been reported less often, but was necessary to identify the disease-causing mutations for Kabuki syndrome in humans[18]. Reduced detectance enables the presence of phenocopies and genetic heterogeneity (albeit locus or allelic heterogeneity). This might be important if dogs from several breeds are combined. For some diseases, the disease-causing mutation is the same in a wide range of breeds (e.g. the c.118G>A mutation in the SOD1 gene for degenerative myelopathy)[16,19]. However, for other diseases, several genes and mutations are known to be disease-causing with some being unique for individual breeds. One example in the dog is the group of progressive retinal atrophies (PRA)[20]. In this group of conditions, a mutation in the PRCD gene is shared by several breeds, but in the Golden Retriever at least two other genes are also linked to PRA and even phenocopies have been reported[20-25].

Overall, these results have various consequences. As dogs from two distinct different breeds share up to 60% of their variants and unrelated dogs within a breed up to 70%, it is likely that sequencing experiments with a limited sample size will be successful, even without access to a (large) variant database. This efficiency for low sample sizes has already been demonstrated for genome-wide association studies in the dog[10]. Due to the relatively large litter size and the frequent use of the same breeding dogs, it should be fairly easy to obtain related animals, as opposed to human studies. Although genetic heterogeneity has of course been reported in the dog, it is still less extensive compared with the human population[23]. Due to the population bottlenecks and the high degree of relatedness, it is not unlikely that complex diseases are relatively easy to study in the dog. Therefore, the dog is likely to be a good animal model to study disorders in the human population, as demonstrated for Retinitis Pigmentosa, the human variant of PRA, and narcolepsy[21,26].

This is the first study evaluating the effect of various parameters in disease-association studies in the dog. We demonstrated that, with a small sample size, a large reduction can be achieved, with recessive inheritance in general being more efficient than dominant inheritance. A balanced combination of cases and controls should be preferred and, if available, closely related family members should be selected. Reduced detectance and/or penetrance are likely to have a profound effect on the sample size needed to elucidate the variant responsible for a genetic disease.

7.6 References

1. Hodges, E. *et al.* Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* **39,** 1522–1527 (2007).
2. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42,** 30–35 (2010).
3. Fairfield, H. *et al.* Mutation discovery in mice by whole exome sequencing. *Genome Biol.* **12,** R86 (2011).
4. Robert, C. *et al.* Design and development of exome capture sequencing for the domestic pig (Sus scrofa). *BMC Genomics* **15,** 550 (2014).
5. Broeckx, B. J. G. *et al.* Development and performance of a targeted whole exome sequencing enrichment kit for the dog (Canis Familiaris Build 3.1). *Sci. Rep.* **4,** 5597 (2014).
6. Hoeppner, M. P. *et al.* An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. *PLoS One* **9,** e91172 (2014).
7. Broeckx, B. J. G. *et al.* Improved canine exome designs, featuring ncRNAs and increased coverage of protein coding genes. *Sci. Rep.* **5,** 12810 (2015).
8. Stitziel, N. O., Kiezun, A. & Sunyaev, S. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol.* **12,** 227 (2011).
9. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438,** 803–819 (2005).
10. Karlsson, E. K. *et al.* Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat. Genet.* **39,** 1321–1328 (2007).
11. Goh, G. & Choi, M. Application of whole exome sequencing

to identify disease-causing variants in inherited human diseases. *Genomics Inf.* **10,** 214–219 (2012).

12. Ahonen, S. J., Arumilli, M. & Lohi, H. A CNGB1 Frameshift Mutation in Papillon and Phalène Dogs with Progressive Retinal Atrophy. *PLoS One* **8,** e72122 (2013).

13. Helsmoortel, C. *et al.* Challenges and opportunities in the investigation of unexplained intellectual disability using family based whole exome sequencing. *Clin. Genet.* **88,** 140–148 (2015).

14. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS One* **7,** e46688 (2012).

15. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* 1–41 (2013). doi:10.1002/0471142905.hg0720s76

16. Coates, J. R. & Wininger, F. A. Canine degenerative myelopathy. *Vet. Clin. North Am. - Small Anim. Pract.* **40,** 929–950 (2010).

17. Minor, K. M. *et al.* Presence and impact of the exercise-induced collapse associated DNM1 mutation in Labrador retrievers and other breeds. *Vet. J.* **189,** 214–219 (2011).

18. Ng, S. B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Am. J. Med. Genet.* **42,** 790–793 (2011).

19. Awano, T. *et al.* Genome-wide association analysis reveals a SOD1 mutation in canine degenerative myelopathy that resembles amyotrophic lateral sclerosis. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 2794–2799 (2009).

20. Petersen-Jones, S. Advances in the molecular understanding of canine retinal diseases. *J. Small Anim. Pract.* **46,** 371–80

(2005).

21. Zangerl, B. *et al.* Identical Mutation in a Novel Retinal Gene Causes Progressive Rod-Cone Degeneration (prcd) in Dogs and Retinitis Pigmentosa in Man. *Genomics* **88,** 551–563 (2006).

22. Downs, L. M. *et al.* A frameshift mutation in Golden Retriever dogs with progressive retinal atrophy endorses SLC4A3 as a candidate gene for human retinal degenerations. *PLoS One* **6,** e21452 (2011).

23. Miyadera, K., Acland, G. M. & Aguirre, G. D. Genetic and phenotypic variations of inherited retinal diseases in dogs: The power of within- and across-breed studies. *Mamm. Genome* **23,** 40–61 (2012).

24. Narfström, K., Jeong, M., Hyman, J., Madsen, R. W. & Bergström, T. F. Assessment of hereditary retinal degeneration in the English Springer Spaniel dog and disease relationship to an RPGRIP1 mutation. *Stem Cells Int.* **2012,** 685901 (2012).

25. Downs, L. M. & Mellersh, C. S. An Intronic SINE Insertion in FAM161A that Causes Exon-Skipping Is Associated with Progressive Retinal Atrophy in Tibetan Spaniels and Tibetan Terriers. *PLoS One* **9,** e93990 (2014).

26. Chen, L., Brown, R. E., Mckenna, J. T. & Mccarley, R. W. Animal Models of Narcolepsy. *CNS Neurol Disord Drug Targets* **8,** 296–308 (2009).

7.7 Supplementary material

Supplementary Table S7.1

General sequencing statistics.

Sample	Total Reads	Mapped Reads	Duplicate Reads	Remaining Reads (%)	Sequencing depth (x)
Exome-1.0 (HiSeq 2500 PE 100 bp)					
I	96,041,166	93,261,066	8,278,081	84,982,985 (88.5)	106.9
II	90,534,096	83,841,822	4,680,806	79,161,016 (87.4)	102.0
III	77,786,110	72,147,586	4,457,341	67,690,245 (87.0)	87.1
IV	82,574,410	77,392,469	4,820,648	72,571,821 (87.9)	93.0
V	74,657,388	69,542,653	4,518,820	65,023,833 (87.1)	82.6
VI	111,624,766	108,781,536	9,882,797	98,898,739 (88.6)	125.1
VII	86,094,438	83,226,207	5,926,249	77,299,958 (89.8)	99.3
VIII	103,290,412	100,440,603	8,653,736	91,786,867 (88.9)	116.7
Exome-plus (NextSeq 500 PE 75 bp)					
IX	266,996,086	251,028,902	17,002,907	234,025,995 (87.7)	75.3
X	187,857,302	176,207,940	12,226,544	163,981,396 (87.3)	53.9
XI	284,357,886	264,735,195	9,414,179	255,321,016 (89.8)	85.8
XII	281,522,490	261,170,320	9,366,318	251,804,002 (89.4)	84.3
XIII	233,403,500	216,361,182	13,330,685	203,030,497 (87.0)	65.0
XIV	181,728,820	168,679,105	4,382,284	164,296,821 (90.4)	55.5

8 The Prevalence of Nine Genetic Disorders in a Dog Population from Belgium, the Netherlands and Germany

B.J.G. Broeckx[a], F. Coopman[b], G.E.C. Verhoeven[c], W. Van Haeringen[d], L. van de Goor[d], T. Bosmans[e], I. Gielen[c], J.H. Saunders[c], S.S.A. Soetaert[a], H. Van Bree[c], C. Van Neste[a], F. Van Nieuwerburgh[a], B. van Ryssen[c], E. Verelst[b], K. Van Steendam[a], D. Deforce[a]

a Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Ottergemsesteenweg 460, 9000 Ghent, Belgium

b Department of Applied Biosciences, Faculty of Bioscience Engineering, University College Ghent, Valentin Vaerwyckweg 1, 9000 Ghent, Belgium

c Department of Medical Imaging and Small Animal Orthopaedics, Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium

d Dr. Van Haeringen Laboratorium b.v., AgroBusinessPark 100, 6708 PW Wageningen, The Netherlands

e Department of Medicine and Clinical Biology of Small Animals, Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium

PLoS ONE 8 (2014), e74811

8.1 Abstract

The objective of this study was to screen a dog population from Belgium, the Netherlands and Germany for the presence of mutant alleles associated with hip dysplasia (HD), degenerative myelopathy (DM), exercise-induced collapse (EIC), neuronal ceroid lipofuscinosis 4A (NCL), centronuclear myopathy (HMLR), mucopolysaccharidosis VII (MPS VII), myotonia congenita (MG), gangliosidosis (GM1) and muscular dystrophy (Duchenne type) (GRMD). Blood samples (K3EDTA) were collected for genotyping with Kompetitive Allele Specific PCR (n = 476). Allele and genotype frequencies were calculated in those breeds with at least 12 samples (n = 8). Hardy-Weinberg equilibrium was tested. Genetic variation was identified for 4 out of 9 disorders: mutant alleles were found in 49, 15, 3 and 2 breeds for HD, DM, EIC and NCL respectively. Additionally, mutant alleles were identified in crossbreeds for both HD and EIC. For HD, DM, EIC and NCL mutant alleles were newly discovered in 43, 13, 2 and 1 breed(s), respectively. In 9, 2 and 1 breed(s) for DM, EIC and NCL respectively, the mutant allele was detected, but the respective disorder has not been reported in those breeds. For 5 disorders (HMLR, MPS VII, MG, GM1, GRMD), the mutant allele could not be identified in our population. For the other 4 disorders (HD, DM, EIC, NCL), prevalence of associated mutant alleles seems strongly breed dependent. Surprisingly, mutant alleles were found in many breeds where the disorder has not been reported to date.

8.2 Background

The reduction of genetic disorders remains an important goal for both veterinarians and breeders[1]. It is the responsibility of the scientific community to provide detailed information regarding the existence, application and importance of diagnostic genetic tests that have been developed[1–3]. An important step in this process is to evaluate the prevalence of disorders and mutant alleles in the population. This information is needed to provide proper breeding advice. So far, only a few studies have been conducted to identify allele frequencies in a canine population[3–6]. The presence of mutant alleles in different breeds with each having their specific genetic background, might provide interesting (clinical) information for both the animal and human population as many canine disorders are animal models for human genetic disorders.

This study reports on the prevalence of mutant alleles associated with 9 canine genetic disorders (Table 8.1) that influence the neuronal and/or musculoskeletal system: hip dysplasia (HD), degenerative myelopathy (DM), exercise-induced collapse (EIC), neuronal ceroid lipofuscinosis 4A (NCL), centronuclear myopathy (HMLR), mucopolysaccharidosis VII (MPS VII), myotonia congenita (MG), gangliosidosis (GM1) and muscular dystrophy (Duchenne type) (GRMD). The tests were performed in a dog population from Belgium, the Netherlands and Germany. Eight out of these 9 disorders are animal models for similar conditions in humans. The mutations predicted to cause HD, DM, EIC and NCL, were detected in a wide variety of breeds.

Table 8.1
Overview of 9 disorders tested with their corresponding chromosome number, the mutation, the effect, the inheritance, the breeds where the mutation has been reported before and the animal model. CFA = chromosome number; AR = autosomal recessive, XR = X-linked recessive, MP = multifactorial; MS = missense, SP = splice variant, FS = frame shift, ES = exon skipping, SINE = short interspersed element, SNP = single nucleotide polymorphism; [1] leading to premature termination, [2] reduced penetrance.

Disorder	Gene	CFA	Mutation		Effect	inheritance	Breeds with mutation reported so far	Similar human disease	Ref
Hip dysplasia (HD)	*Fibrillin 2 (FBN2)* (Gene ID: 481491)	11	intronic (3 SNPs + deletion)	GAT > AGC	?	MP	Labrador Retriever, Border Collie, German Shepherd Dog, Golden Retriever, Newfoundland, Rottweiler, Great Dane	Congenital Contractural Arachnodactyly	7
Degenerative myelopathy (DM)	*Superoxide dismutase 1 (SOD1)* (Gene ID: 403559)	31	exon (1 SNP)	G > A	MS	AR[2]	German Shepherd Dog, Boxer, Rhodesian Ridgeback, Chesapeake Bay Retriever, Pembroke Welsh Corgi	Amyotrophic lateral sclerosis	8
Exercise-induced collapse (EIC)	*Dynamin 1 (DNM1)* (Gene ID: 491319)	9	exon (1 SNP)	G > T	MS	AR[2]	Labrador Retriever, Chesapeake Bay Retriever, Curly-coated Retriever, Boykin Spaniel, Pembroke Welsh Corgi	-	9
Neuronal ceroid lipofuscinosis 4A (NCL)	*Arylsulfatase G (ARSG)* (Gene ID: 480460)	9	exon (1 SNP)	G > A	MS	AR[2]	American Staffordshire Terrier	Kufs disease	10
Centronuclear myopathy (HMLR)	*Protein tyrosine phosphatase-like (PTPLA)* (Gene ID: 574011)	2	exon (SINE)	SINE	MS	AR	Labrador Retriever	Human centronuclear myopathy	11
Mucopolysaccharidosis VII (MPS VII)	*Beta-glucuronidase (GUSB)* (Gene ID: 403831)	6	exon (1 SNP)	G>A	MS	AR	German Shepherd Dog	Mucopolysaccharidosis VII	12
	Beta-glucuronidase (GUSB) (Gene ID: 403831)	6	exon (1 SNP)	C>T	MS	AR	Brazilian Terrier	Mucopolysaccharidosis VII	13
Myotonia congenita (MG)	*Chloride channel, voltage sensitive 1 (CLCN1)* (Gene ID: 403723)	16	exon (1 SNP)	C>T	MS	AR	Miniature Schnauzer	Generalized myotonia (Beckers disease)	14
Gangliosidosis (GM1)	*Beta-galactosidase (GLB1)* (Gene ID: 403873)	23	exon (deletion)	C	FS[1]	AR	Shiba	Gangliosidosis	15
Muscular dystrophy (Duchenne) (GRMD)	*Dystrophin (DMD)* (Gene ID: 606758)	X	intron (1 SNP)	A>G	ES	XR	Golden Retriever	Duchenne Muscular Dystrophy	16

8.3 Materials and methods

Ethics Statement. Approval from the local ethical (Faculty of Veterinary Medicine, Ghent University, Belgium) and deontological (Federal Public Service Health, Food Chain Safety and Environment, Brussels, Belgium) committees was granted (EC2010_171 and EC2011_193). All efforts were made to minimize suffering. Informed consent was obtained from owners of dogs before enrollment in the study.

Sample Collection. Blood samples (K3EDTA) were collected for a genetic database to study HD. Veterinarians from Belgium, the Netherlands and Germany were asked to take a blood sample from every dog that had a hip radiograph taken. Reasons for performing the procedure varied from screening purposes (breeding and assistance dogs) to dogs with clinical complaints (with HD in the differential diagnosis). No prerequisites were made regarding breed, sex and age.

Irrespective of breed, samples (n = 476) were tested for the presence of mutations associated with 9 disorders (HD, DM, EIC, NCL, HMLR, MPS VII, MG, GM1 and GRMD). Breeds where the mutant allele has already been reported, can be found in Table 8.1. A summary of breeds and samples per breed can be found in Table 8.2. Additionally, a mixed breed group of 28 dogs was tested.

Genotyping. Genomic DNA was isolated using routine procedures. For blood samples, 10 µl of blood was washed 3 times with 150 µl of a Tris-HCL based buffer. The procedure was performed with the use of

Table 8.2
Breed and samples per breed tested. a = not specified whether English or Welsh Springer Spaniel.

Breed	n	%	Breed	n	%
Airedale Terrier	2	0.4	Gordon Setter	1	0.2
Akita	1	0.2	Great Dane	1	0.2
American Bulldog	1	0.2	Hovawart	1	0.2
American Cocker Spaniel	1	0.2	Hungarian Vizsla	7	1.5
American Staffordshire Terrier	18	3.8	Jack Russell Terrier	2	0.4
Anatolian Shepherd Dog	2	0.4	Labrador Retriever	75	15.8
Appenzeller Sennenhund	1	0.2	Laekenois	1	0.2
Australian Kelpie	1	0.2	Large Munsterlander	1	0.2
Australian Shepherd	6	1.3	Leonberger	4	0.8
Basset Hound	1	0.2	Malinois	7	1.5
Berger de Picardie	2	0.4	Maltese	1	0.2
Bernese Mountain Dog	20	4.2	Mastino Napoletano	1	0.2
Blue Picardy Spaniel	1	0.2	Miniature Pinscher	1	0.2
Boerboel	1	0.2	Munsterlander	1	0.2
Border Collie	29	6.1	Nederlandse Schapendoes	1	0.2
Bouvier des Flandres	3	0.6	Newfoundland	4	0.8
Boxer	15	3.2	Nova Scotia Duck Tolling Retriever	3	0.6
Briard	1	0.2	Rhodesian Ridgeback	1	0.2
Bull Terrier	1	0.2	Rottweiler	7	1.5
Cavalier King Charles Spaniel	4	0.8	Saarlooswolfhond	2	0.4
Collie Rough	2	0.4	Saint Bernard Dog	3	0.6
Dalmatian	2	0.4	Samoyed	1	0.2
Dobermann	2	0.4	Shar Pei	4	0.8
Dogo Argentino	1	0.2	Shetland Sheepdog	1	0.2
Dogue de Bordeaux	5	1.1	Shiba	3	0.6
Dwergschnauzer	1	0.2	Siberian Husky	2	0.4
English Bulldog	4	0.8	Spanish Water Dog	12	2.5
English Cocker Spaniel	2	0.4	Springer Spaniel[a]	1	0.2
English Setter	2	0.4	Stabyhoun	3	0.6
English Springer Spaniel	1	0.2	Standard Poodle	2	0.4
Epagneul Breton	2	0.4	Tervueren	2	0.4
Flat Coated Retriever	8	1.7	Tibetan Mastiff	1	0.2
French Bulldog	3	0.6	Weimaraner	4	0.8
German Shepherd Dog	73	15.3	White Swiss Shepherd Dog	5	1.1
Golden Retriever	62	13.0	Wire-Haired Pointing Griffon Korthals	1	0.2
			Total	476	

robotic equipment. The cell pellet was lysed with Proteinase K (0.5 units for 45 minutes at 56°C followed by heat inactivation at 95°C for 5 minutes).

Genotyping was conducted using KASP (http://www.kbioscience.co.uk), a competitive allele specific polymerase chain reaction system, according to the manufacturer's instructions. Primer sequences were based on literature (see Table 8.1) and transformed to be compatible with KASP. All tests are routinely run at the dr. Van Haeringen Laboratorium (Wageningen, the Netherlands).

Statistical Analysis. Breed specific prevalence was analyzed in 8 breeds for which at least 12 samples were available (German Shepherd Dog, Labrador Retriever, Golden Retriever, Border Collie, Bernese Mountain Dog, American Staffordshire Terrier, Boxer, Spanish Water Dog). Hardy – Weinberg equilibrium (HWE) was tested with an online calculator (www.tufts.edu/~mcourt01/Documents). Data are available on request.

8.4 Results

No variation was found in any of the breeds for the mutations putatively responsible for 5 of the 9 disorders (HMLR, MPS VII, MG, GM1, GRMD). The mutations predicted to cause HD, DM, EIC and NCL, were observed in a wide variety of breeds (49, 15, 3 and 2 breeds, respectively). Breeds where mutant alleles were found are listed in Tables 8.3, 8.4, 8.5 and 8.6 for HD, DM, EIC and NCL, respectively. For HD, DM, EIC and NCL mutant alleles were newly discovered in 43, 13, 2 and 1 breed(s), respectively. The mutant alleles for HD and EIC were also identified in mixed breed dogs. For 25 of the 28 crossbreds, the parental breeds were known and we have shown these breeds to possess the respective mutant allele.

For 8 breeds, breed specific prevalence of alleles were reported (Table 8.3–8.6). Comparisons with previous reports could only be made for EIC in the Labrador Retriever: our population contained a very high number of genetically affected dogs (Table 8.7)[6]. HWE could be tested for HD in the German Shepherd Dog and the Golden Retriever (no significant deviation). For the Labrador Retriever, genotype frequencies for HD and EIC both deviated significantly from HWE ($p \leq 0.001$). The other breeds and disorders were not tested due to low number of samples and/or absence of variation.

8.5 Discussion

Since its first report in 1935, intensive research has focused on hip dysplasia (HD), one of the most frequent orthopedic disorder in dogs[17]. HD can be found in a wide variety of breeds[18,19]. Recently, a haplotype in the Fibrillin 2 (FBN2) gene has been reported to be associated with this highly prevalent, multifactorial disorder[7]. Fibrillins are components of extracellular microfibrils and have both a structural and a regulatory function[20]. The mutant AGC haplotype was identified in 49 different breeds in the population under study (Table 8.3). In 44 breeds, this mutant allele was reported for the first time. We identified the AGC haplotype as the only allele in 10 breeds (Cavalier King Charles Spaniel (n = 4), English Setter (n = 2), English Springer Spaniel (n = 1), Gordon Setter (n = 1), Laekenois (n = 1), Mastino Napoletano (n = 1), Rhodesian Ridgeback (n = 1), Saarlooswolfhond (n = 2), Siberian Husky (n = 2), Standard Poodle (n = 2)), however few conclusions can be made as the sample count is low (n = 17 overall). One additional dog was homozygous for the mutant allele, but as the breed was not correctly specified (Springer Spaniel without mentioning English or Welsh Springer Spaniel), this sample was excluded.

Table 8.3

Breeds where mutant alleles for hip dysplasia were found and breed specific prevalence for breeds with at least 12 samples.

	NN	NA	AA	Total	HWE	q (%)
	n (%)	n (%)	n (%)	n (%)	P-value	
American Staffordshire Terrier	4 (24)	9 (53)	4 (24)	17 (100)	-	50
Bernese Mountain Dog	7 (35)	12 (60)	1 (5)	20 (100)	-	35
Border Collie	25 (89)	3 (11)	0 (0)	28 (100)	-	5
Boxer	15 (100)	0 (0)	0 (0)	15 (100)	-	0
German Shepherd Dog	34 (47)	34 (47)	4 (6)	72 (100)	0.225	29
Golden Retriever	10 (16)	31 (51)	20 (33)	61 (100)	0.728	58
Labrador Retriever	32 (45)	21 (30)	18 (25)	71 (100)	0.001	40
Spanish Water Dog	4 (33)	8 (67)	0 (0)	12 (100)	-	33
Others:	Airedale Terrier, Appenzeller Sennenhund, Australian Shepherd, Blue Picardy Spaniel, Boerboel, Bouvier des Flanders, Briard, Bull Terrier, Cavalier King Charles Spaniel, Collie Rough, Dalmatian, Dobermann, Dogo Argentino, Dogue de Bordeaux, English Bulldog, English Cocker Spaniel, English Setter, English Springer Spaniel, Epagneul Breton, Flat Coated Retriever, Gordon Setter, Hovawart, Hungarian Vizsla, Laekenois, Large Munsterlander, Leonberger, Malinois, Maltese, Mastino Napoletano, Miniature Pinscher, Newfoundland, Nova Scotia Duck Tolling Retriever, Rhodesian Ridgeback, Rottweiler, Saarlooswolfhond, Saint Bernard Dog, Shetland Sheepdog, Siberian Husky, Springer Spaniel[a], Stabyhoun, Standard Poodle, Weimaraner, White Swiss Shepherd Dog					

NN = 2 normal alleles, NA = heterozygous, AA = 2 mutant alleles, HWE = Hardy-Weinberg equilibrium, q = mutant allele frequency, % = percent of dogs belonging to specific category, - = not applicable, a = not specified whether English or Welsh Springer Spaniel.

Degenerative myelopathy (DM) is characterized by progressive ataxia and upper motor neuron spastic paresis. The majority of dogs with DM start to develop symptoms from 5 years of age[21]. Diagnosis is not straightforward when patients are alive. Based on clinical symptoms and exclusion of other disorders (for example intervertebral disc disease, spinal cord neoplasia), DM can be the most likely etiology, but formal diagnosis can only be achieved post-mortem on histopathology[21]. Only for a subset of those breeds in which DM has been reported clinically, this disorder has been confirmed on histopathology. In 2009, a causal mutation was discovered in the superoxide dismutase 1 (SOD1) gene for 5 breeds (Table 8.1) with an age-dependent incomplete penetrance[8]. This gene encodes a free radical scavenger[21]. We identified the causal mutation in 15 breeds. In 2 breeds (Standard Poodle and Bernese Mountain Dog) the disorder was confirmed both clinically and on histopathology while in another 2 breeds (Border Collie and Collie Rough) the diagnosis of DM was made only on clinical examination[21]. We report the presence of the mutant allele in these 4 breeds and in an additional 9 breeds, in which the disorder has not yet been reported. We also confirmed the presence of the mutant allele in the German Shepherd Dog and Boxer (Table 8.4). DM has been clinically reported in the Labrador Retriever and confirmed by histopathology in the Golden Retriever, but the mutation has not yet been reported in these breeds[21]. In our population of respectively 74 and 62 individuals, we also did not identify the SOD1 mutation. The mutant allele might be infrequently present in the population or a different mutation might be responsible for the same disorder, as recently reported[22].

Table 8.4

Breeds where mutant alleles for degenerative myelopathy were found and breed specific prevalence for breeds with at least 12 samples.

	NN	NA	AA	Total	HWE	q (%)
	n (%)	n (%)	n (%)	n (%)	P-value	
American Staffordshire Terrier	18 (100)	0 (0)	0 (0)	18 (100)	-	0
Bernese Mountain Dog	12 (60)	7 (35)	1 (5)	20 (100)	-	23
Border Collie	27 (96)	1 (4)	0 (0)	28 (100)	-	2
Boxer	13 (87)	2 (13)	0 (0)	15 (100)	-	7
German Shepherd Dog	53 (73)	18 (25)	2 (3)	73 (100)	-	15
Golden Retriever	62 (100)	0 (0)	0 (0)	62 (100)	-	0
Labrador Retriever	74 (100)	0 (0)	0 (0)	74 (100)	-	0
Spanish Water Dog	12 (100)	0 (0)	0 (0)	12 (100)	-	0
Others:	Airedale Terrier, Australian Shepherd, Cavalier King Charles Spaniel, Collie Rough, Dobermann, Dogo Argentino, Saarlooswolfhond, Shetland Sheepdog, Stabyhoun, Standard Poodle, White Swiss Shepherd Dog					

NN = 2 normal alleles, NA = heterozygous, AA = 2 mutant alleles, HWE = Hardy-Weinberg equilibrium, q = mutant allele frequency, % = percent of dogs belonging to specific category, - = not applicable.

Dogs with exercise-induced collapse (EIC) develop incoordination of the hind limbs, paraparesis and/or tetraparesis and have an increased body temperature after strenuous exercise. A mutation in the dynamin 1 (DNM1) gene was found to be responsible for this disorder and an incomplete penetrance (influenced by the level of physical activity) was suggested[9]. DNM1 is important in neuronal synaptic vesicle recycling, especially during high levels of activity[23]. This disorder is reported mainly in Labrador Retrievers, but presence of the mutant allele has also been reported in other retriever breeds[6,9]. We report on the presence of the mutant allele for the first time in one English Cocker Spaniel and one Hungarian Vizsla (Table 8.5). The prevalence of EIC in our Labrador population was higher than in a previous report (Table 8.7)[6]. The sample source seems to affect the percentage of affected dogs: when samples were collected from Labrador Retrievers in dog shows, field trials and from local pet owners in Canada and the United States, the prevalence of affected dogs was relatively low and in agreement with HWE expectations[6]. The opposite was true for dogs that were specifically tested for EIC and those results were in agreement with the results from our study, although we did not specifically select for those dogs. In our population of Labrador Retrievers, genotype frequencies for both HD and EIC deviate significantly from HWE. This might be a reflection of non-random sampling or selection[4]. For HD, non-random sampling can be expected since our dogs were initially collected to study this disorder. For EIC, the reason for rejection of the HWE is not clear. For both HD and EIC, our results tend to overestimate the number of affected dogs based on HWE. A similar result was found in the pet population in a previous study (26.9 versus 29.2%)[6].

Table 8.5

Breeds where mutant alleles for exercise-induced collapse were found and breed specific prevalence for breeds with at least 12 samples.

	NN	NA	AA	Total	HWE	q (%)
	n (%)	n (%)	n (%)	n (%)	P-value	
American Staffordshire Terrier	18 (100)	0 (0)	0 (0)	18 (100)	-	0
Bernese Mountain Dog	18 (100)	0 (0)	0 (0)	18 (100)	-	0
Border Collie	25 (100)	0 (0)	0 (0)	25 (100)	-	0
Boxer	12 (100)	0 (0)	0 (0)	12 (100)	-	0
German Shepherd Dog	69 (100)	0 (0)	0 (0)	69 (100)	-	0
Golden Retriever	51 (100)	0 (0)	0 (0)	51 (100)	-	0
Labrador Retriever	30 (46)	16 (25)	19 (29)	65 (100)	< 0.001	42
Spanish Water Dog	12 (100)	0 (0)	0 (0)	12 (100)	-	0
Others:	Hungarian Vizsla, English Cocker Spaniel					

NN = 2 normal alleles, NA = heterozygous, AA = 2 mutant alleles, HWE = Hardy-Weinberg equilibrium, q = mutant allele frequency, % = percent of dogs belonging to specific category, - = not applicable.

Neuronal ceroid lipofuscinosis 4A (NCL) has been reported in American Staffordshire Terriers with progressive ataxia[24]. The causal mutation was discovered in 2010 in the Arylsulfatase G (ARSG) gene and an incomplete penetrance was suggested[10]. ARSG encodes for a lysosomal enzyme[25]. The mutant allele has a prevalence of approximately 9% in our 17 American Staffordshire Terriers which is less than the frequency expected based on a previous study[24]. We report the presence of the same mutation in the Bull Terrier, a breed where NCL has not been reported (Table 8.6). No dogs were homozygous for the mutant allele in the population studied.

Table 8.6

Breeds where mutant alleles for neuronal ceroid lipofuscinosis 4A were found and breed specific prevalence for breeds with at least 12 samples.

	NN n (%)	NA n (%)	AA n (%)	Total n (%)	HWE P-value	q (%)
American Staffordshire Terrier	14 (82)	3 (18)	0 (0)	17 (100)	-	9
Bernese Mountain Dog	19 (100)	0 (0)	0 (0)	19 (100)	-	0
Border Collie	26 (100)	0 (0)	0 (0)	26 (100)	-	0
Boxer	13 (100)	0 (0)	0 (0)	13 (100)	-	0
German Shepherd Dog	71 (100)	0 (0)	0 (0)	71 (100)	-	0
Golden Retriever	57 (100)	0 (0)	0 (0)	57 (100)	-	0
Labrador Retriever	70 (100)	0 (0)	0 (0)	70 (100)	-	0
Spanish Water Dog	12 (100)	0 (0)	0 (0)	12 (100)	-	0
Others:			Bull Terrier			

NN = 2 normal alleles, NA = heterozygous, AA = 2 mutant alleles, HWE = Hardy-Weinberg equilibrium, q = mutant allele frequency, % = percent of dogs belonging to specific category, - = not applicable.

For the other 5 disorders (HMLR, MPS VII, MG, GM1, GRMD) the mutant alleles were not detected in our population. Non-detection of the mutant allele might indicate that the allele is absent, that it is present but at a low frequency and that the number of samples tested was too low. For HMLR (Labrador Retriever, n = 66) and MPS VII (German Shepherd Dog, n = 67), we had, with the numbers of animals tested, a 99% chance of detecting every allele with a minor allele frequency (MAF) of at least 3.5%[26]. For GRMD (Golden Retriever, n = 19), we had a 99% chance of detecting alleles with a frequency of at least 11.5%[26]. For MG, GM1 and MPS VII (in the Brazilian Terrier), the sample size was too small to make any conclusions regarding the absence of the mutant allele[26]. The allele frequency for HMLR has previously been investigated[4]. In that study, the mutant allele frequency was very low (1.8% or 0.47%). To be able to detect all alleles with a MAF of 1% with a 99.9% probability, 344 samples would be needed. Since our sample count was much lower, it cannot be concluded that any of these 5 mutant alleles are completely absent.

Table 8.7

Comparisons of genotype frequencies for exercise-induced collapse in the Labrador Retriever.

			Reference populations[6]	
Predicted Phenotype	Our population	Est. Freq (HWE)	Source: Public	Source: researchers
	% (n)	% (n)	% (n)	% (n)
Healthy (homo)	46.2 (30)	34.2 (22.2)	52.9 (4826)	59.2 (509)
Healthy (hetero)	24.6 (16)	48.6 (31.6)	37.2 (3392)	34.5 (297)
Affected	29.2 (19)	17.2 (11.2)	9.9 (907)	6.3 (54)
Total	100 (65)	100 (65)	100 (9125)	100 (860)

The reference population consists of 2 different subsets based on collection method. Source: Public = based on request by the owner to perform genetic testing for EIC, Soure: researchers = researchers went to several competitions and took samples from every dog. Est. Freq (HWE) = estimated frequencies under Hardy-Weinberg equilibrium.

For DM, EIC and NCL breeding advice can be given based on our population study. For some disorders, the mutant allele frequency is quite high in certain breeds (> 40% in Labrador Retriever for EIC). However, because of their recessive nature, a relatively fast reduction of both the mutant allele frequency and affected dogs can be achieved based on genetic tests. We propose to exclude certain genotypic combinations of dogs from mating, rather than excluding individuals, especially if high mutant allele frequencies are found. For autosomal recessive diseases, dogs homozygous for the mutant allele should not be combined with each other or with heterozygous dogs. However, they can still be used for breeding, but require a mating combination involving only homozygous wild type animals to reduce the number of affected dogs. As heterozygous dogs can be used the same way, no dogs need to be excluded. As essentially every dog can still be used for breeding, the clinical outcome of disorders can be prevented without excessive exclusion of carriers or genetically affected dogs from the breeding population. Reduction of the prevalence of HD based only on FBN2 will be more difficult since it is a multifactorial polygenetic disorder.

Surprisingly, mutant alleles for 3 autosomal recessive disorders (DM, EIC and NCL) were found in 9, 2 and 1 breed(s) respectively where the disorder has not been clinically reported. The most plausible explanation might be that the disorder just has not been recognized in those breeds. A second explanation is the influence of the breed specific genetic background: the effect of mutations might be different in different breeds. This has been reported in Drosophila and the mouse[27,28]. In a meta-analysis in humans where studies on a wide variety of diseases and genes were compared, opposite effects between races were found, but none of them significant[29]. To the authors' knowledge, this phenomenon has not yet been reported in dogs for the diseases studied here.

This study reports on the presence of mutant alleles for 9 disorders in a wide variety of dog breeds. Veterinarians and dog breeders should be aware that mutations are present in breeds even where the disorder has not been reported. Dogs from non-suspected breeds that show comparable symptoms to the disorders reported in this and other studies should be genotyped and results should be reported in order to create a reliable database. Ideally, phenotypical information for every disorder should be included in this database. As this is not available for all diseases in our database, this is a major limitation to this study.

8.6 References

1. Mellersh, C. DNA testing and domestic dogs. *Mamm. Genome* **23,** 109–123 (2012).

2. Nicholas, F. W. Online Mendelian Inheritance in Animals (OMIA): A comparative knowledgebase of genetic disorders and other familial traits in non-laboratory animals. *Nucleic Acids Res.* **31,** 275–277 (2003).

3. Karmi, N. *et al.* Estimated Frequency of the Canine Hyperuricosuria Mutation in Different Dog Breeds. *J. Vet. Intern. Med.* **24,** 1337–1342 (2010).

4. Gentilini, F. *et al.* Frequency of the allelic variant of the PTPLA gene responsible for centronuclear myopathy in Labrador Retriever dogs as assessed in Italy. *J. Vet. Diagn. Invest.* **23,** 124–126 (2011).

5. Gould, D. *et al.* ADAMTS17 mutation associated with primary lens luxation is widespread among breeds. *Vet. Ophthalmol.* **14,** 378–384 (2011).

6. Minor, K. M. *et al.* Presence and impact of the exercise-induced collapse associated DNM1 mutation in Labrador retrievers and other breeds. *Vet. J.* **189,** 214–219 (2011).

7. Friedenberg, S. G. *et al.* Evaluation of a fbrillin 2 gene haplotype associated with hip dysplasia and incipient osteoarthritis in dogs. *Am. J. Vet. Res.* **72,** 530–540 (2011).

8. Awano, T. *et al.* Genome-wide association analysis reveals a SOD1 mutation in canine degenerative myelopathy that resembles amyotrophic lateral sclerosis. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 2794–2799 (2009).

9. Patterson, E. E. *et al.* A canine DNM1 mutation is highly associated with the syndrome of exercise-induced collapse. *Nat. Genet.* **40,** 1235–1239 (2008).

10. Abitbol, M. *et al.* A canine Arylsulfatase G (ARSG) mutation leading to a sulfatase deficiency is associated with neuronal ceroid lipofuscinosis. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 14775–14780 (2010).

11. Pelé, M., Tiret, L., Kessler, J. L., Blot, S. & Panthier, J. J. SINE exonic insertion in the PTPLA gene leads to multiple splicing defects and segregates with the autosomal recessive centronuclear myopathy in dogs. *Hum. Mol. Genet.* **14,** 1417–1427 (2005).

12. Ray, J. *et al.* Cloning of the canine beta-glucuronidase cDNA, mutation identification in canine MPS VII, and retroviral vector-mediated correction of MPS VII cells. *Genomics* **48,** 248–253 (1998).

13. Hytönen, M. K. *et al.* A novel GUSB mutation in Brazilian terriers with severe skeletal abnormalities defines the disease as mucopolysaccharidosis VII. *PLoS One* **7,** 1–11 (2012).

14. Rhodes, T. H. *et al.* A missense mutation in canine ClC-1 causes recessive myotonia congenita in the dog. *FEBS Lett.* **456,** 54–58 (1999).

15. Yamato, O. *et al.* A novel mutation in the gene for canine acid B-galactosidase that causes GM1-gangliosidosis in Shiba dogs. *J. Inherit. Metab. Dis.* **25,** 525–526 (2002).

16. Sharp, N. J. *et al.* An error in dystrophin mRNA processing in golden retriever muscular dystrophy, an animal homologue of Duchenne muscular dystrophy. *Genomics* **13,** 115–121 (1992).

17. Schnelle, G. B. Some new diseases in the dog. *Am. Kennel Gaz.* **52,** 25–26 (1935).

18. Coopman, F., Verhoeven, G., Saunders, J., Duchateau, L. & van Bree, H. Prevalence of hip dysplasia, elbow dysplasia and humeral head osteochondrosis in dog breeds in Belgium. *Vet. Rec.* **163,** 654–658 (2008).

19. Rettenmaier, J. L., Keller, G. G., Lattimer, J. C., Corley, E. a & Ellersieck, M. R. Prevalence of canine hip dysplasia in a veterinary teaching hospital population. *Vet. Radiol. ultrasound* **43,** 313–318 (2002).

20. Ramirez, F. & Dietz, H. C. Fibrillin-Rich Microfibrils: structural determinants of morphogenetic and homeostatic events. *J. Cell. Physiol.* **213,** 326–330 (2007).

21. Coates, J. R. & Wininger, F. A. Canine degenerative myelopathy. *Vet. Clin. North Am. - Small Anim. Pract.* **40,** 929–950 (2010).

22. Wininger, F. a. *et al.* Degenerative Myelopathy in a Bernese Mountain Dog with a Novel SOD1 Missense Mutation. *J. Vet. Intern. Med.* **25,** 1166–1170 (2011).

23. Ferguson, S. M. *et al.* A selective activity-dependent requirement for dynamin 1 in synaptic vesicle endocytosis. *Science* **316,** 570–574 (2007).

24. Olby, N. *et al.* Cerebellar cortical degeneration in adult American Staffordshire Terriers. *J. Vet. Intern. Med.* **18,** 201–8 (2004).

25. Frese, M. -a., Schulz, S. & Dierks, T. Arylsulfatase G, a Novel Lysosomal Sulfatase. *J. Biol. Chem.* **283,** 11388–11395 (2008).

26. Gregorius, H. R. The probability of losing an allele when diploid genotypes are sampled. *Biometrics* **36,** 643–652 (1980).

27. Coleman, D. L. & Hummel, K. P. The influence of genetic background on the expression of the obese (Ob) gene in the mouse. *Diabetologia* **9,** 287–293 (1973).

28. Huang, W. *et al.* Epistasis dominates the genetic architecture of Drosophila quantitative traits. *Proc. Natl. Acad. Sci.* **109,** 15553–15559 (2012).

29. Ioannidis, J. P. a, Ntzani, E. E. & Trikalinos, T. a. ‘Racial’ differences in genetic effects for complex diseases. *Nat. Genet.* **36,** 1312–1318 (2004).

9 Discussion: the promises of whole exome sequencing in canine genetics

This chapter can be divided into two parts. The first, more general part (9.1 to 9.4) starts with a discussion of the phenotypical issues encountered when studying HD and briefly states the consequences for screening. This is followed by a demonstration of the practical use of DNA testing, using DM as an example. Difficulties in estimating disease prevalence with either DNA tests or phenotypically are discussed as is the importance of the dog as an animal model. Although some of these topics might be considered evidentiary or outside the scope of this dissertation, these topics were included to situate our research within the larger picture of canine genetics. This dissertation is a result from a fruitful collaboration with a wide range of professionals: veterinarians, dog breeders, assistance and rescue dogs federations and individual dog owners. We are indebted to all participants, for which this research is of utmost importance. The second part (9.5 to 9.7) focuses on WES and "Mendelian" as tools for mutation identification in genetic research and discusses future expectations.

9.1 Phenotypical problems: the issue with HD

The importance of correct phenotyping cannot be stressed enough. HD is a perfect disorder to demonstrate potential sources for phenotypical misclassification. Worldwide, the standard VD is the technique used in screening programs. However, it has been demonstrated that this technique lacks sensitivity to diagnose laxity in the hip joint[1,2]. Besides at the level of the technique that is used, several factors have been reported to influence the diagnosis. In chapter 3, we demonstrated a clear effect of the evaluating assessor and the positioning of the dog on the radiographic diagnosis of HD. Additional factors that interfere with the diagnosis are age of the dog and whether the dog is anesthetized or not; even the choice of sedatives has been reported to influence the diagnosis[3-7].

All these factors increase the environmental noise relative to the genetic contribution and it is the combination of both that results in a complex phenotype such as HD. If not handled correctly, these factors will decrease the effectiveness of screening programs and of genetic studies that aim to discover disease-contributing variants. Based on literature and our results, we propose the following improvements for radiographic HD screening:

- Define clear measurable quality criteria for radiographs (e.g. maximum amount of rotation of the pelvis around the longitudinal axis)
- Add a laxity based radiographical technique and, if this is not possible, increase the age of screening
- Standardize the anesthesia protocol

The assessment by multiple observers, familiar with the HD enigma, should be continued.

9.2 Underneath the (phenotypical) surface: the genotype

It is clear that, because of the complexity of complex disorders, reduction of disease prevalence purely based on the phenotype will be difficult. However, also for simple disorders, DNA testing offers a tremendous advantage. We demonstrate the practical use of DNA testing based on the results of chapter 8. DM is a recessive disorder with incomplete penetrance. Assuming that penetrance is complete (which is a simplification of reality), DM would be seen in 3% of the studied German Shepherd population. By excluding all dogs that are phenotypically affected (genotype = aa), only dogs with genotype Aa and AA will remain for breeding. As long as at least one AA dog is used in every mating, no phenotypically affected dogs will be born. However, if Aa dogs are mated, on average 25% will be phenotypically affected. Without knowing the genotype, these matings cannot be avoided and affected individuals will be born. DNA tests can completely remove phenotypical DM from the population in one generation. This is achieved by only allowing aa x AA and Aa x AA matings. In addition, no animals have to be excluded from breeding when a DNA test is used, as even aa animals (that are phenotypically affected) combined with AA animals will not result in affected progeny. Thus, aside from the more efficient elimination of disease, DNA tests increase the population size of breeding dogs which is beneficiary for the overall population health.

An additional point of consideration is that in reality, phenotypical selection is even less efficient: due to the reduced penetrance, even genetically affected animals that did not (yet?) express DM, would be used for breeding from time to time. Based on these results, it is clear that DNA tests can speed up health improvement.

9.3 Disease prioritization

In reality, health improvement is not an issue of one disease only. As the canine population suffers from several diseases, it is likely that some form of disease prioritization will be necessary in order to maintain a sufficiently large population for breeding. Disease prioritization requires an assessment of the impact of the disease on the health of the dog, aside from a determination of the prevalence of the disorder[8–10]. Knowledge on the prevalence of diseases is also important to define the optimal strategy for health improvement. The results from chapter 3 and chapter 8 are two examples of studying disease prevalence based on either the phenotype or a genetic test. As detailed in the previous section, DNA tests are preferred, but unfortunately, are not available for every disorder.

Based on the results from chapter 3, reliable disease estimates are difficult to obtain. Aside from the previously discussed factors that influence prevalence indirectly by adding noise to the phenotype, two factors influence the prevalence directly. Firstly, the selection bias in radiograph submission decreases the disease prevalence overall and also tends to reduce the differences in disease prevalence between breeds. The second point is that the sampling population should reliably represent the true population. It is clear that the orthopedic clinical complaint dog population is probably (hopefully) not a reliable representation of the general dog population with over 70% of the dogs having HD. However, the breeding population is also not representative due to the selection bias.

The results of the DNA tests from chapter 8 demonstrate that the disease prevalence is highly variable between breeds (from absent to 30% genetically affected individuals) and that the carrier frequency can be high (up to 35%). As shown by the rejection of HWE for two tests in the Labrador Retriever, an accurate selection of the test population is also important for DNA tests.

Overall, obtaining reliable disease prevalence estimates is important, but not easy. Estimating disease prevalence requires the collaboration of all parties involved in animal breeding. All factors that contribute to underidentification of a disease should be addressed because decision makers base their plans on these results.

9.4 Why do we study the dog?

Based on the previous sections, genetic research on canine diseases will likely improve their health. In addition to being an enjoyable pet, the dog is a good animal model and thus beneficial for biomedical research. Compared to established laboratory animals, such as mice, dogs have several advantages. Here is an overview of why the dog has so much potential as an animal model:

Phenotype. The dog population has a remarkable phenotypical diversity and a high frequency of spontaneously occurring genetic diseases. Frequently occurring spontaneous diseases reduce the need to establish and maintain research colonies as the pet population itself can be used directly. In addition, the willingness of individual owners to participate in research studies is often considerable. With litters of sometimes 10 puppies or more, the collection of closely related cases and controls is often relatively easy. With the health care standard close to that of human medicine, the majority of specialized diagnostical and therapeutical techniques are available for the dog, resulting in improved phenotypical characterization[11]. A large number of diseases share clinical and laboratory abnormalities in humans and dog and at this moment, well over 50% of the genetic diseases in the dog are considered to be potential models for human diseases[11,12]. Especially for complex diseases, spontaneous models have the benefit over induced models as the latter often are simplified models that knock out only one gene[13]. Both in size and structure, the anatomy of several organs of the dog resembles the human anatomy more than routinely used laboratory animals[14,15]. As companions, dogs share the world of their owners, resulting also in a shared environment that may influence the pathogenesis of disease. Extrapolation of results and studying of the bigger picture of interactions between genes and environment will certainly be more realistic than in the typical laboratory animal environment[14].

Genetic constitution. The creation of the dog as we know it, has clearly left its marks on the dog genome. The large haplotype blocks and low genetic heterogeneity have already been reported[13,16]. The practical advantage this has for WES has only recently been demonstrated (chapter 7). The same processes that caused the high prevalence of spontaneous diseases can be put to good use now because they will likely limit the number of animals necessary to identify causal mutations[13,17] (chapter 7). Additional advantages are that the dog genome is less diverged from the human than the mouse genome[18]. With the recent release of WES designs, several modern tools to study diseases are now available. Especially for complex diseases, the combination of high disease prevalence, genetic population isolation due to the pedigree barrier and sequencing based approaches is promising[11,19].

Translation to human diseases. Knowledge of disease-causing mutations in the dog has already led to the identification of several disease-causing mutations in humans. The identification of a mutation in the PRCB gene, responsible for progressive rod-cone degeneration in the dog, led to the subsequent discovery of that identical mutation in a woman with autosomal recessive Retinitis Pigmentosa[20]. Two other examples are the identification of mutations in the *hypocretin receptor 2* gene and the *preprohypocretin* gene, responsible for canine narcolepsy and an autosomal dominant early onset narcolepsy in a human patient, respectively and the association between *PNPLA1* mutations and canine and human ichthyosis[21–23].

Promises for treatment. Aside from the importance for breeding to improve health in future generations, knowledge of the causal mutation also holds the promise of improvement of the health of affected individuals[14]. Especially for eye diseases, the dog has been a valuable partner. Congenital stationary night blindness or retinal dystrophy is the spontaneously occurring canine homologue of Leber congenital amaurosis[24].

Upon identification of the causal mutation, gene therapy with a recombinant adeno-associated virus successfully restored vision in the dog[24]. It was only after this success that gene therapy studies were conducted in the mouse, which subsequently led to clinical trials with promising results in humans[14,15,25]. For other diseases such as achromatopsia and hereditary nephropathies, results obtained by gene therapy in the dog are also promising[14,15].

All these examples illustrate that the relationship between dogs and humans is a relationship that benefits both species and proves again that the dog really is "man's best friend".

9.5 Additional resources for the toolbox: WES

Currently, the majority of genetic studies conducted in the dog rely on GWAS. However, the transition from the indirect GWAS to direct sequencing is being made, as demonstrated on the latest Canine and Feline genomics conference (Cambridge, UK, 2015). In human genetics however, sequencing-based approaches such as WES are already being used for several years[26]. Possible reasons are the cost (GWAS is less expensive than WES), the success of GWAS ("never change a winning team") and the limited availability of WES designs for the dog[27]. As sequencing costs are plummeting, the difference in costs between GWAS and WES will decrease[26,28]. With the recent release of several WES designs (chapter 4 and 5), availability will be less of an issue from now on.

As already mentioned in chapter 5, the three available designs differ mainly in terms of: whether or not to target (3' and 5') UTRs and whether the focus should be almost exclusively on protein-coding regions or whether a large non-protein coding catalogue is to be targeted as well. Especially in cancer research and complex diseases, extending the focus beyond the protein-coding regions is advisable[29-31]. However for Mendelian disorders, the majority of causal variants are located in protein-coding regions[32,33]. Choosing the smallest design, the exome-1.0, for these disorders is the most cost-efficient option[32,33].

In comparison with human WES designs, two aspects are to be highlighted. First of all, with the availability of three canine WES designs, the same flexibility as commercial WES designs for humans is guaranteed. The flexibility is increased further as up to 200 Mb can be targeted nowadays, so additional regions of interest can be easily added. A second remark is that none of the human designs offer the combination of all the non-protein coding regions, all the UTRs and all the protein-coding regions in one design. Thus, the exome-plus is superior in terms of completeness.

9.5.1 Potential limitations

Just like GWAS, it is important to realize that WES is a tool and as with every other tool, it has its limitations.

One of the most important points of consideration, is that WES is only as strong as the quality of the annotation it is based on. If the annotation misses some genes, WES designs will not target them, unless there is overlap with other targets by coincidence. Actually, one of the reasons for the extended WES designs (exome-CDS and exome-plus), was the release of an updated annotation[30].

A second point is that, even if regions are targeted, some of them will not be sequenced. The most important reason for low coverage was found to be related to a high GC-content, followed by low-complexity/highly repetitive regions, as discussed in chapter 5. This is in agreement with previous reports and also for other sequencing technologies, these regions can be difficult to cope with[34,35].

Consequently, when the results of several samples are compared, there will always be some variability in the sequencing depth of certain regions between samples. Some regions might be covered at a sufficient depth in one sample, while this might not be the case in others.

A final issue relates to the detection of structural variants. A structural variant (SV) is defined as a genomic rearrangement of more than 50 bp and has already been associated with canine diseases (e.g. a 133 kb duplication is associated with dermoid sinus in the Rhodesian Ridgeback)[36,37]. Although WES has been used successfully to detect SVs, it remains difficult[37–40].

The problem of incomplete annotation is difficult to solve: clairvoyance is needed to know what has been missed. Due to the recent high-quality update, the annotation has certainly improved significantly. Details on the number of un(der)covered regions or base pairs of interest are presented in chapter 4 and 5. Overall, with on average 90% of the base pairs consistently covered at a sequencing depth of at least 5x, the majority of the targets will be sequenced. Whether this will be sufficient or not, depends on the study. However, it is important to realize that a variable coverage is not uniquely related to WES. Even when WGS is performed, some regions will be un(der)represented and some of them might have been sequenced when WES was performed[28].

9.5.2 Future perspective of WES

Realistically, the overall decrease of sequencing costs will pave the way for WGS as the additional preparatory steps for capturing are relatively expensive. We agree that the transition is likely to be made at some point, probably within the next five or ten years, but we believe that the current development of WES designs will serve their purpose for some time.

Although the field of genomics is evolving at an incredible pace, “older” techniques are still used widely side-by-side to new ones. Although Sanger sequencing was invented almost 40 years ago, it is still used in routine clinical settings as it can be used to fill gaps missed by NGS or when the size of the target regions is small[41,42]. Even though the advantage of sequencing-based approaches compared to GWAS is clear, a PubMed search (("genome-wide association study") AND (human), limiting results to those published between 1/1/2015 to 16/7/2015) returned 484 hits, illustrating that the GWAS approach is still applicable. In addition, although the $1000 genome has been announced by Illumina in 2014, it requires considerable investments initially and a large volume of samples to reach that price.

Where other WES designs often only focus on the protein-coding regions, the large non-coding catalogue of the exome-plus provides an intermediate option between the pure "exome" and complete genome.

Although sequencing itself is becoming less of a limitation, the computational burden remains important[26,29]. The entire genome is considerably larger (the exome-plus targets only 6% of the genome) and the analysis requires increased processing time and data storage. It will not be easy to solve this problem.

9.5.3 Conclusion

Currently, the developed WES enrichment designs provide a cost-efficient alternative between GWAS and WGS. Although WGS will probably replace WES at some point, we believe WES is here to stay (for a while).

9.6 Prior to and after the wet lab: optimal case-control selection and variant analysis with "Mendelian"

Where the shelf life of the WES designs might be limited, the R-package "Mendelian" and the results from the analysis of the various case-control designs are not. Whether the data are obtained from WES or WGS, from Illumina or newly developed sequencers, the end result will likely remain the same: a list of sequencing variants that need to be filtered. In chapter 6, “Mendelian” was used successfully to revalidate mutations for two autosomal recessively inherited phenotypes (brown and yellow coat colour in the Labrador Retriever) and a *de novo* mutation associated with mental retardation in humans[43–46].

The power of heuristic filtering of sequencing variants for autosomal Mendelian disorders was demonstrated further in chapter 7. By comparing several case-control designs of individuals with variable degrees of relatedness for both dominant and recessive types of inheritance, practical guidelines for optimal sample selection are provided. Although the effect of the population bottlenecks has been thoroughly investigated, it was still surprising that dogs from two different breeds share up to 60% of their variants[13,17,47]. Within a breed, this increases to 70%. This has several implications. First of all, the number of animals necessary to retain only the disease-causing variant will be limited. In addition, due to the close relatedness between dogs, extensive genetic heterogeneity is rather unlikely, as reported previously[16,31]. Due to the high number of shared variants, it is more efficient to include one control than to include a second case, especially if this control is a close relative (full sibs or parents) of the case. This is an advantage: as long as the prevalence of a disease does not exceed 50%, it is much easier to collect additional controls than additional cases.

Although “Mendelian” was not specifically designed for identifying X-linked disorders, it can be used for this purpose, as discussed in the online manual. This might require some tweaking of the VCF-files, depending on the program and the exact combination of male and female individuals.

As the attention shifts towards complex disorders, the question is whether "Mendelian" will be of any help here. Basically, complex disorders are no more than a combination of genetic and environmental factors that lead to a reduced penetrance and detectance. As "Mendelian" allows both reduced penetrance and/or detectance, it should be technically possible. Realistically however, lowering the thresholds also results in less variants being filtered. Especially for rare variants, the thresholds will have to be set so low that almost no variant is being filtered. In chapter 7, the effect of lowering the thresholds for penetrance and/or detectance confirmed that heuristic filtering will be inefficient for complex disorders. To identify those variants, more statistical approaches need to be used, but even then the identification remains challenging[19]. Overall however, the high degree of shared variants makes one wonder what the allele frequency of (rare and common) variants is in canine complex disorders. Based on our results, we expect the allele frequencies of both to be higher, but this remains to be determined.

9.7 Short-term evolution

As demonstrated in previous sections, the combination of WES and "Mendelian" is promising. WES has already been used in the dog 1) after mapping with a GWAS to identify the causal mutation[48] and 2) without mapping to revalidate a causal mutation (the coat colour loci in chapter 7). The next step is the identification of new causal genetic variation for unstudied phenotypes. Due to the added complexity of common disorders, it might be safer to study Mendelian disorders first, but the transition to complex disorders will follow soon.

Future studies will likely benefit from an extension of the dbSNP variant database. This database was very important in the development of the GWAS kits, but it was built with a limited number of samples (one dog from every breed, for a total of 11 breeds) which makes it less useful for filtering sequencing variants[13,17]. Due to the high prevalence of disorders in specific breeds, it is very likely that carriers of these breed-specific diseases were unintentionally selected, resulting in contamination (i.e. the database contains mutant alleles associated with disease) of the database for that disease. For example, for a recessive disease with a prevalence of 5%, under the assumptions of HWE and 100% detectance and penetrance, 35% of the population is a carrier of the mutant allele for that disease (see Appendix A for calculations). To avoid exclusion of the causal variant when filtering sequencing data, it is important to use only those variants in the variant database that have allele frequencies higher or equal to a certain MAF. That MAF can be estimated based on the disease prevalence in a population. However, this requires a variant database with reliable allele frequencies which can only be obtained by sequencing several dogs from one breed. A joint effort will be required to sequence a sufficient amount of dogs from the majority of the dog breeds, but initiatives in these direction are being proposed[49].

9.8 Final conclusion

In 2000, Elaine Ostrander stated: “canine genetics comes of age”. Since then, the dog genome and several improved annotations have become publicly available and linkage studies have been (largely) replaced by GWAS. With the release of several WES designs, an R-package for heuristic filtering and guidelines for case and control selection in sequencing studies, we hope to have contributed to the maturation of canine genetics and the healthy aging of the dog.

9.9 References

1. Culp, W. T. N. *et al.* Evaluation of the Norberg angle threshold: A comparison of Norberg angle and distraction index as measures of coxofemoral degenerative joint disease susceptibility in seven breeds of dogs. *Vet. Surg.* **35,** 453–459 (2006).
2. Powers, M. Y. *et al.* Evaluation of the relationship between Orthopedic Foundation for Animals' hip joint scores and PennHIP distraction index values in dogs. *J. Am. Vet. Med. Assoc.* **237,** 532–541 (2010).
3. Broeckx, B. J. G. *et al.* The effects of positioning, reason for screening and the referring veterinarian on prevalence estimates of canine hip dysplasia. *Vet. J.* **201,** 378–384 (2014).
4. Genevois, J.-P. *et al.* Influence of anaesthesia on canine hip dysplasia score. *J. Vet. Med. A. Physiol. Pathol. Clin. Med.* **53,** 415–417 (2006).
5. Malm, S. *et al.* Impact of sedation method on the diagnosis of hip and elbow dysplasia in Swedish dogs. *Prev. Vet. Med.* **78,** 196–209 (2007).
6. Verhoeven, G. *et al.* Interobserver agreement in the diagnosis of canine hip dysplasia using the standard ventrodorsal hip-extended radiographic method: Paper. *J. Small Anim. Pract.* **48,** 387–393 (2007).
7. Smith, G. K. *et al.* Lifelong diet restriction and radiographic evidence of osteoarthritis of the hip joint in dogs. *J. Am. Vet. Med. Assoc.* **229,** 690–693 (2006).
8. Asher, L., Diesel, G., Summers, J. F., McGreevy, P. D. & Collins, L. M. Inherited defects in pedigree dogs. Part 1: Disorders related to breed standards. *Vet. J.* **182,** 402–411 (2009).
9. Summers, J. F., Diesel, G., Asher, L., McGreevy, P. D. & Collins, L. M. Inherited defects in pedigree dogs. Part 2: Disorders that are not related to breed standards. *Vet. J.* **183,** 39–45 (2010).

10. Bell, J. S. Researcher responsibilities and genetic counseling for pure-bred dog populations. *Vet. J.* **189,** 234–235 (2011).

11. Ostrander, E. a., Galibert, F. & Patterson, D. F. Canine genetics comes of age. *Trends Genet.* **16,** 117–124 (2000).

12. Nicholas, F. W. Online Mendelian Inheritance in Animals (OMIA): A comparative knowledgebase of genetic disorders and other familial traits in non-laboratory animals. *Nucleic Acids Res.* **31,** 275–277 (2003).

13. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438,** 803–819 (2005).

14. Tsai, K. L., Clark, L. A. & Murphy, K. E. Understanding hereditary diseases using the dog and human as companion model systems. *Mamm. Genome* **18,** 444–451 (2007).

15. Petersen-Jones, S. M. & Komáromy, A. M. Dog Models for Blinding Inherited Retinal Dystrophies. *Hum. Gene Ther. Clin. Dev.* **26,** 150211074229002 (2015).

16. Miyadera, K., Acland, G. M. & Aguirre, G. D. Genetic and phenotypic variations of inherited retinal diseases in dogs: The power of within- and across-breed studies. *Mamm. Genome* **23,** 40–61 (2012).

17. Karlsson, E. K. *et al.* Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat. Genet.* **39,** 1321–1328 (2007).

18. Karlsson, E. K. & Lindblad-Toh, K. Leader of the pack: gene mapping in dogs and other model organisms. *Nat. Rev. Genet.* **9,** 713–725 (2008).

19. Auer, P. L. & Lettre, G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med.* **7,** 1–11 (2015).

20. Zanger, B., Goldstein, O., Alisdair R. Philp, Sarah J.P. Lindauer, S. E., Pearce-Kelling, Robert F. Mullins, Alexander S.

Graphodatsky, Daniel Ripoll, J. & S. Felix, Edwin M. Stone, g, Gregory M. Acland, and G. D. A. Identical Mutation in a Novel Retinal Gene Causes Progressive Rod-Cone Degeneration (prcd) in Dogs and Retinitis Pigmentosa in Man. *Genomics* **88,** 551–563 (2006).

21. Lin, L. *et al.* The sleep disorder canine narcolepsy is caused by a mutation in the hypocretin (orexin) receptor 2 gene. *Cell* **98,** 365–376 (1999).

22. Peyron, C. *et al.* A mutation in a case of early onset narcolepsy and a generalized absence of hypocretin peptides in human narcoleptic brains. *Nat. Med.* **6,** 991–997 (2000).

23. Grall, A. *et al.* PNPLA1 mutations cause autosomal recessive congenital ichthyosis in golden retriever dogs and humans. *Nat. Genet.* **44,** 140–147 (2012).

24. Acland, G. M. *et al.* Gene therapy restores vision in a canine model of childhood blindness. *Nat. Genet.* **28,** 92–95 (2001).

25. Trapani, I., Banfi, S., Simonelli, F., Surace, E. M. & Auricchio, A. Gene Therapy of Inherited Retinal Degenerations: Prospects and Challenges. *Hum. Gene Ther.* **26,** 193–200 (2015).

26. Goh, G. & Choi, M. Application of whole exome sequencing to identify disease-causing variants in inherited human diseases. *Genomics Inf.* **10,** 214–219 (2012).

27. Bushell, K. R. *et al.* Genetic inactivation of TRAF3 in canine and human B-cell lymphoma. *Blood* **125,** 999–1006 (2015).

28. Clark, M. J. *et al.* Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.* **29,** 908–914 (2011).

29. Biesecker, L. G., Shianna, K. V & Mullikin, J. C. Exome sequencing: the expert view. *Genome Biol.* **12,** 128 (2011).

30. Hoeppner, M. P. *et al.* An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. *PLoS One* **9,** e91172 (2014).

31. Davis, B. W. & Ostrander, E. a. Domestic Dogs and Cancer Research: A Breed-Based Genomics Approach. *ILAR J.* **55,** 59–68

(2014).

32. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12,** 745–755 (2011).

33. Stitziel, N. O., Kiezun, A. & Sunyaev, S. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol.* **12,** 227 (2011).

34. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36,** e105 (2008).

35. Aird, D. *et al.* Analyzing and minimizing bias in Illumina sequencing libraries. *Genome Biol.* **12,** R18 (2011).

36. Salmon Hillbertz, N. H. C. *et al.* Duplication of FGF3, FGF4, FGF19 and ORAOV1 causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. *Nat. Genet.* **39,** 1318–1320 (2007).

37. Tattini, L., D'Aurizio, R. & Magi, A. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Front. Bioeng. Biotechnol.* **3,** 1–8 (2015).

38. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12,** 363–376 (2011).

39. Tan, R. *et al.* An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data. *Hum. Mutat.* **35,** 899–907 (2014).

40. Han, S. M. *et al.* Genetic Testing of Korean Familial Hypercholesterolemia Using Whole-Exome Sequencing. *PLoS One* **10,** e0126706 (2015).

41. Katsanis, S. H. & Katsanis, N. Molecular genetic testing and the future of clinical genomics. *Nat. Rev. Genet.* **14,** 415–26 (2013).

42. Rehm, H. L. Disease-targeted sequencing: a cornerstone in the clinic. *Nat. Rev. Genet.* **14,** 295–300 (2013).

43. Everts, R. E., Rothuizen, J. & Van Oost, B. a. Identification of a

premature stop codon in the melanocyte-stimulating hormone receptor gene (MC1R) in Labrador and Golden retrievers with yellow coat colour. *Anim. Genet.* **31,** 194–199 (2000).

44. Newton, J. M. *et al.* Melanocortin 1 receptor variation in the domestic dog. *Mamm. Genome* **11,** 24–30 (2000).

45. Schmutz, S. M., Berryere, T. G. & Goldfinch, A. D. TYRP1 and MC1R genotypes and their effects on coat color in dogs. *Mamm. Genome* **13,** 380–387 (2002).

46. Helsmoortel, C. *et al.* Challenges and opportunities in the investigation of unexplained intellectual disability using family based whole exome sequencing. *Clin. Genet.* **88,** 140–148 (2015).

47. Quignon, P. *et al.* Canine population structure: Assessment and impact of intra-breed stratification on SNP-based association studies. *PLoS One* **2,** e1324 (2007).

48. Ahonen, S. J., Arumilli, M. & Lohi, H. A CNGB1 Frameshift Mutation in Papillon and Phalène Dogs with Progressive Retinal Atrophy. *PLoS One* **8,** e72122 (2013).

49. Bai, B. *et al.* DoGSD: the dog and wolf genome SNP database. *Nucleic Acids Res.* **43,** D777–D783 (2014).

Summary

Dogs（Canis familiaris）have been part of human society for thousands of years. Over this large time period, specific characteristics have developed in terms of hair colour, body size, skull shape, etc. Unfortunately, the processes that created the dog as we know it, inadvertently resulted in genetic diseases in the dog being far from rare. This dissertation studies several aspects of canine genetics and begins with a general overview on the methodologies that can be used to study genetic diseases（chapter 1）.

Genetic analyses start with defining the phenotype. Although this appears easy, the classification of an individual as healthy or sick can be difficult. A striking example of a complex phenotype is canine hip dysplasia（chapter 3）. In the first study, the aim was to identify factors that influence the radiographic diagnosis and prevalence estimates of canine hip dysplasia. A total of 583 radiographs were assessed independently by two different observers for hip conformation and for positioning. Overall, the agreement between observers for positioning and hip conformation was limited and can affect estimates of the prevalence of hip dysplasia. Aside from these factors, the prevalence estimates were further influenced by selection bias and the population that was sampled from. This study stresses the importance of correct phenotypical classification for screening programs as well as genetic studies.

A large section of this dissertation focuses on the development and performance of whole exome sequencing designs（chapter 4 and 5）.

Whole exome sequencing is a targeted sequencing method that aims to selectively sequence all the regions from the genome that are transcribed to mRNA. Three different designs were made: the exome-1.0 (53 Mb), the exome-CDS (71 Mb) and the exome-plus (152 Mb). They differ mainly in terms of the choice to target (3' and 5') UTRs and whether the focus should be almost exclusively on protein-coding regions or whether a large non-protein coding catalogue is to be targeted as well. Both the performance of the exome-1.0 and the exome-plus were evaluated for several samples, while the performance of the exome-CDS was estimated based on the results of the exome-plus. Overall, the exome designs all performed well, with the differences between them being mostly related to region size and bait design.

When Mendelian disorders are studied, several assumptions to filter the sequencing variants are typically relied on, until one or a very limited number of putative causal variants remain. The release of several whole exome sequencing designs for the dog, created the need for a heuristic filtering tool capable of analyzing variant data under the assumptions of recessive or dominant modes of inheritance, with variable degrees of penetrance and detectance. With this aim, the R-package "Mendelian" was developed. We demonstrated its performance by revalidating a *de novo* mutation responsible for human intellectual disability and two recessively inherited mutations responsible for the yellow and brown coat colours in the Labrador retriever (chapter 6).

As whole exome sequencing is relatively new, no guidelines on sample selection have been published. Which combinations result in the most efficient variant reduction is not yet known. In order to provide guidelines and some directions on what to expect, several combinations of cases and controls with variable degrees of familial relatedness and assuming dominant or recessive modes of inheritance with variable degrees of penetrance and detectance, were evaluated (chapter 7). Remarkably, up to 60% of the variants were shared between dogs from two distinct breeds. Within a breed, unrelated dogs shared up to 70% of their variants. Although the same processes that caused this phenomenon resulted in the high prevalence of genetic diseases, these same characteristics can be used when studying genetic diseases. Lower sample sizes are needed, genetic heterogeneity is reduced and, as the inclusion of one control is preferred over a second case, sample collection is made easier.

One of the outcomes of disease-association studies is typically a DNA test. An important step in defining the optimal strategy for health improvement and to decide whether the disease should be treated with priority or not, is an assessment of the prevalence of the disorder. In a population of 476 dogs, we determined the allelic frequencies of mutant alleles associated with nine genetic disorders (chapter 8). For 5 disorders, the mutant allele could not be identified in our population. For the other 4 disorders (degenerative myelopathy, exercise-induced collapse, neuronal ceroid lipofuscinosis 4A and a mutation found to be associated with hip dysplasia in a population in the United States), the prevalence of the mutant alleles was strongly breed dependent. The high carrier frequencies in specific breeds reduce the efficacy of phenotypical selection programs. In addition, mutant alleles were found in many breeds where the disorder has not been reported yet. Whether the disease has not been recognized in these breeds or whether a difference in genetic background influences the effect of the mutant alleles, still needs to be determined.

The final chapter of this dissertation, chapter 9, discusses the results of the previous chapters. In addition, it situates our research in the world of canine genetics and stresses the potential of the dog as an animal model for human diseases.

Samenvatting

De hond ofte *Canis familiaris*, is reeds ettelijke duizenden jaren onze metgezel en heeft een opmerkelijke fenotypische diversiteit. De processen die tot deze diversiteit hebben geleid, hebben er echter onbedoeld ook voor gezorgd dat genetische aandoeningen verre van uitzonderlijk zijn. In deze thesis worden meerdere aspecten van deze genetische aandoeningen onderzocht. Alvorens over te gaan tot het eigenlijke onderzoek, wordt in hoofdstuk 1 een algemene inleiding gegeven over welke moderne methodologische middelen er bestaan om genetische aandoeningen te bestuderen.

Het ontrafelen van genetische aandoeningen start met het bestuderen van het fenotype van die aandoening. Beslissen of iemand gezond of ziek is, is niet altijd gemakkelijk. Het ideale voorbeeld van een aandoening die moeilijk te fenotyperen is, is heupdysplasie (hoofdstuk 3). In de eerste studie werd het effect van factoren die de radiografische diagnose van heupdysplasie en schattingen van de prevalentie beïnvloeden, bestudeerd. Hiervoor werden er 583 radiografische opnames van het heupgewricht geblindeerd beoordeeld door 2 personen waarbij hen gevraagd werd om enerzijds de heupen te scoren en anderzijds te beoordelen of de positionering van de hond op de radiografie goed was. Algemeen kon er besloten worden dat de beoordelingen van zowel positionering als heupscore slechts in beperkte mate overeenkwamen. Dit beïnvloedt eveneens schattingen van de prevalentie. Andere factoren die het moeilijk maken om exact in te schatten hoe vaak heupdysplasie voorkomt, zijn selectiebias en de populatie waarvan men steekproeven neemt. Selectiebias betekende in dit geval dat röntgenopnames van honden met duidelijke HD minder vaak ingezonden werden voor officiële beoordeling. Algemeen benadrukt deze studie het belang van correcte fenotypische classificatie voor zowel programma's die screenen op genetische aandoeningen als in genetische studies.

Een groot deel van deze thesis focust zich op de ontwikkeling en de prestaties van zogenaamde "whole exome sequencing" designs (hoofdstuk 4 en 5). Whole exome sequencing is een techniek die gericht bepaalde delen van het genoom van de hond analyseert die worden vertaald naar het boodschapper RNA (= mRNA). In totaal werden er drie van deze designs gemaakt: de exome-1.0 (53 Mb), de exome-CDS (71 Mb) en de exome-plus (152 Mb). De verschillen tussen deze designs zijn voornamelijk te wijten aan de keuze om al dan niet 3' en 5' UTRs mee te sequeneren en of men zich voornamelijk op de regio's richt die coderen voor eiwitten of eveneens een groot deel niet-eiwit coderende regio's mee wilt nemen. Zowel de exome-1.0 als de exome-plus werden uitvoerig getest. De prestaties van de exome-CDS werden geschat op basis van de resultaten van de exome-plus. De voornaamste oorzaken van prestatieverschillen zijn gerelateerd aan de grootte van de te sequeneren regio's en andere instellingen die werden gebruikt tijdens het ontwikkelen van de eigenlijke kit. Evenwel presteerden alle designs goed.

Bij het bestuderen van Mendeliaanse aandoeningen, gaat men vaak uit van bepaalde assumpties. Deze assumpties worden dan gebruikt om van alle varianten die geïdentificeerd worden het overgrote deel weg te filteren totdat uiteindelijk slechts één of een zeer beperkt aantal varianten overblijft. Door het uitbrengen van de whole exome designs ontdekten we dat er nood was aan een software pakket dat kon om gaan met deze verschillende assumpties en ook gebruikt kon worden bij de hond. Met dat doel voor ogen werd "Mendelian" ontwikkeld, software gebaseerd op de programmeertaal R. Om te demonstreren dat "Mendelian" geschikt is voor deze taak, werden een *de novo* mutatie, verantwoordelijk voor mentale retardatie bij de mens, en twee recessief overervende mutaties voor de bruine en gele vachtkleur bij de Labrador Retriever opnieuw aangetoond (hoofdstuk 6).

Aangezien whole exome sequencing bij de hond nieuw is, bestonden er nog geen richtlijnen over welke combinatie of aantallen van gezonde en aangetaste dieren het meest efficiënt zijn om het aantal geïdentificeerde varianten te reduceren. Om enerzijds richtlijnen te kunnen aanbieden en anderzijds een idee te hebben over welke reductie men kan verwachten, werden er verscheidene combinaties van gezonde en aangetaste dieren, met variabele verwantschapsgraden en voor zowel dominante als recessieve kenmerken met variabele penetrantie en detectiegrenzen, getest (hoofdstuk 7). Een opmerkelijk resultaat was dat honden uit verschillende rassen tot 60% van hun varianten delen. Binnen een ras, loopt dit zelfs op tot 70% voor honden die niet verwant zijn. De processen die tot deze hoge percentages hebben geleid, zijn in dit geval positief voor het genetisch onderzoek, maar hebben terzelfdertijd net tot die hoge prevalentie aan genetische aandoeningen geleid. De voordelen van deze hoge percentages zijn de kleinere aantallen honden die nodig zullen zijn om genetisch onderzoek uit te voeren en de beperktere genetische heterogeniteit. Uit deze resultaten bleek ook dat het voordeliger is om als tweede hond een gezonde hond te sequeneren in plaats van een tweede zieke hond. Dit vergemakkelijkt de staalcollectie aangezien gezonde honden (in dit geval, honden zonder die specifieke aandoening) meer zullen voorkomen dan een tweede aangetaste hond, zeker indien het gaat over zeldzame aandoeningen.

Eén van de uiteindelijke doelen van genetisch onderzoek is de ontwikkeling van een DNA test. Om de optimale strategie te bepalen om een ziekte te bestrijden en ook om te kunnen uitmaken of een ziekte absolute prioriteit verdient of niet, is het belangrijk om de prevalentie van de aandoening te kennen. In hoofdstuk 8 werden de allelfrequenties bepaald van 9 verschillende genetische aandoeningen in een populatie van 476 honden. De mutaties verantwoordelijk voor 5 van deze 9 aandoeningen werden niet ontdekt. Van 4 andere aandoeningen (degeneratieve myelopathie, collaps geïnduceerd door inspanning, neuronale ceroïd lipofuscinosis 4A en een mutatie die geassocieerd zou zijn met heupdysplasie in een Amerikaanse populatie honden) werden de causale varianten teruggevonden. Afhankelijk van het ras verschilden de allelfrequenties sterk. Uit de grote aantallen dragers blijkt dat puur fenotypische selectie een stuk minder efficiënt zal verlopen dan selectie met behulp van DNA testen. Opmerkelijk is ook dat de mutaties werden teruggevonden in rassen waar de ziekte nog niet gerapporteerd werd. Het kan zijn dat de ziekte in deze rassen nog niet herkend werd. Een andere mogelijkheid is dat het effect van de causale varianten beïnvloed wordt door de rasgebonden genetische achtergrond.

In het laatste hoofdstuk, hoofdstuk 9, worden de resultaten uit de eerdere hoofstukken bediscussieerd. Het kadert ook dit doctoraat binnen het domein van het genetisch onderzoek bij de hond. Tot slot wordt er verwezen naar de belangrijke rol die de hond kan spelen als diermodel voor de mens.

Appendix

For a recessive disease, the carrier frequency and the frequencies of the mutant and wild type alleles, are calculated starting from the disease prevalence.

The following assumptions are made:

1) The population is in HWE for the alleles studied

2) 100% penetrance

3) 100% detectance

Due to the first assumption, the following relationship exists between alleles and genotypes:

	Genotype		
	AA	Aa	aa
Frequency	p^2	2pq	q^2

With p and q the allele frequencies for A (= the wild type allele) and a (= the mutant allele), respectively.

Due to assumption two and three:

P(disease phenotype) = P(aa) = q^2

To calculate q, the square root of the prevalence of disease is taken. As p+q=1, 2pq = 2(1-q)q = the carrier frequency

The relation between disease prevalence, q and the carrier frequency is depicted graphically in Figure A.1:

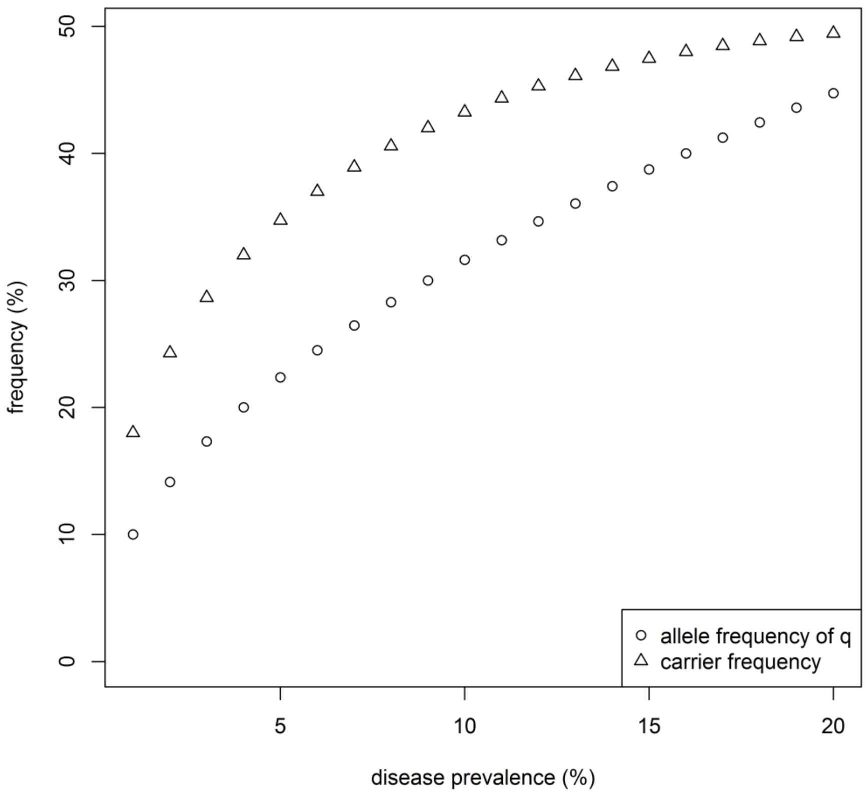


Figure A.1. Relation between disease prevalence (%), carrier frequency and the allele frequency of the mutant allele q.

It is clear that even for rare diseases, carriers occur at high frequencies in the population.

Curriculum vitae

A. General information

Name:	Bart Broeckx
Date of birth:	24/03/1987
Place of birth:	Antwerp
Nationality:	Belgian

B. Education and working experience

2011 - 2015	PhD Student (IWT), Laboratory of Pharmaceutical Biotechnology
2008 – 2011	Master in Veterinary Medicine, Ghent University; summa cum laude
2005 – 2008	Bachelor in Veterinary Medicine, University of Antwerp; magna cum laude
1999 - 2005	Secondary school: Science – Mathematics, Xaveriuscollege

C. Additional courses and certificates

General

- Leadership Foundation course, Transferable skills Doctoral Schools, Ghent University (6, 13, 20 October 2014)

- PennHIP Certified Member (2012)

- Introduction to linux for bioinformatics, VIB (3 and 7 October 2011)

- Expert Laboratory Animal Leader, FELASA category C, Ghent University (2011)

Statistical training

- **Computational Biology** (part of Master of Statistical Data-analysis, organized in 2nd semester of academic year 2014 - 2015). Ghent, Ghent University, Belgium, 5 credits, exam score: 16/20

- **Analysis of Continuous Data** (part of Master of Statistical Data-analysis, organized in 1st semester of academic year 2014 - 2015). Ghent, Ghent University, Belgium, 5 credits, exam score: 16/20

- Flames Summer School in Methodology and Statistics, **Concepts of Multilevel, Longitudinal and Mixed Models**, Flanders Training Network For Methodology and Statistics, Ghent, Belgium (15-19 September 2014)

- Flames Summer School in Methodology and Statistics, **Advanced Regression Topics**, Flanders Training Network For Methodology and Statistics, Ghent, Belgium (15-19 September 2014)

- Flames Summer School in Methodology and Statistics, **Basic Regression Analysis**, Flanders Training Network For Methodology and Statistics, Ghent, Belgium （8-12 September 2014）

- Advanced statistical methods, module 4: **statistical genome analysis**, Ghent University （12, 19 and 26 March and 2, 23, 30 April and 7, 14 May 2014）

- Advanced statistical methods, module 3: **Multivariate Data Analysis**, Ghent University （15, 22, 29 January and 5, 12, 19 February 2014）

- Advanced statistical methods, module 2: **non-parametrical techniques**, Ghent University （6, 13, 20, 27 November and 11, 18 December 2013）

- Advanced statistical methods, module 1: **R**, Ghent University （11, 18, 25 September and 2, 9, 16 October 2013）

- **Multivariate Data Analysis**, Ghent University （19 – 21 October and 4 November 2011）

- **Design of experiments**, Ghent University, （6-7, 12 and 18 October 2011）

D. Scientific curriculum

A1 Publications （Accepted）

1. **B.J.G. Broeckx** et al. （2013）. The Prevalence of Nine Genetic Disorders in a Dog Population from Belgium, the Netherlands and Germany. PLoS ONE 8(9): e74811. doi: 10.1371/journal.pone.0074811.

2. **B.J.G. Broeckx** et al. (2014). The effects of positioning, reason for screening and the referring veterinarian on prevalence estimates of canine hip dysplasia. Vet. J. 201, 378–384. doi: 10.1016/j.tvjl.2014.05.023.

3. **B.J.G. Broeckx** et al. (2014). Development and performance of a targeted whole exome sequencing enrichment kit for the dog (Canis Familiaris Build 3.1). Sci. Rep. 4, 5597. doi: 10.1038/srep05597.

4. F. Coopman, **B.J.G. Broeckx** et al. (2014). Combined prevalence of inherited skeletal disorders in dog breeds in Belgium. Vet Comp Orthop Traumatol 27. doi: org/10.3415/VCOT-13-11-0140.

5. K. Kromhout, H. van Bree, **B.J.G. Broeckx** et al. (2014). Low-Field MRI and Multislice CT for the Detection of Cerebellar (Foramen Magnum) Herniation in Cavalier King Charles Spaniels. J. Vet. Intern. Med. 29, 238-242. doi: 10.1111/jvim.12498.

6. Fortrie R.R., Verhoeven G., **B.J.G. Broeckx** et al. (2015). Intra- and Interobserver Agreement on Radiographic Phenotype in the Diagnosis of Canine Hip Dysplasia. Vet Surg. 44, 467-473. doi: 10.1111/j.1532-950X.2014.12309.x.

7. K. Kromhout, H. van Bree, **B.J.G. Broeckx** et al. (2015). Low-field MRI and multislice CT for the detection of cervical syringomyelia in dogs. J. Vet. Intern. Med. 29, 1354-1359. doi: 10.1111/jvim.13579.

8. V. Gobin, M. De Bock, **B.J.G. Broeckx** et al. (2015). Fluoxetine suppresses calcium signaling in human T lymphocytes through depletion of intracellular calcium stores. Cell Calcium 06/2015, 25. doi: 10.1016/j.ceca.2015.06.003.

9. **B.J.G. Broeckx** et al. (2015). Improved canine exome designs, featuring ncRNAs and increased coverage of protein coding genes. Sci. Rep. 5, 12810. doi: 10.1038/srep12810.

10. A. Villamonte Ch, H. van Bree, **B.J.G. Broeckx** et al (2015). Assessment of Medial Coronoid Disease in 180 Canine Lame Elbow Joints: A Sensitivity and Specificity Comparison of Radiographic, Computed Tomographic and Arthroscopic Findings. BMC Veterinary Research 11, 243. doi: 10.1186/s12917-015-0556-9.

11. **B.J.G. Broeckx** et al. Towards the most ideal case-control design with related and unrelated dogs in whole exome sequencing studies. Accepted in Animal Genetics (in press).

12. **B.J.G. Broeckx** et al (2015). An heuristic filtering tool to identify phenotype-associated genetic variants applied to human intellectual disability and canine coat colours. BMC Bioinformatics 16, 391. doi: 10.1186/s12859-015-0822-7.

A1 Publications (Other)

13. A. Villamonte Ch, W. Dingemanse, **B.J.G. Broeckx** et al. Bone Density of the canine elbow joints in Labrador retriever and Golden retriever dogs: A comparison of healthy and diseased joints. Revised manuscript (The Veterinary Journal)

14. E. Coppieters, I. Gielen, G. Verhoeven, E. de Bakker, Y. Samoy, **B.J.G. Broeckx** et al. The effect of a bi-oblique dynamic proximal ulnar osteotomy as a single treatment in immature dogs with a fissure

of the medial coronoid process: short- and long-term results. Submitted to VCOT.

15. E. Coppieters, G. Verhoeven, **B.J.G. Broeckx** et al. Spectrum of arthroscopic findings in 92 canine elbow joints diagnosed with medial compartment erosion. Submitted to VCOT.
16. C.P. Crijns, Y. Baeumlin, L. De Rycke, **B.J.G. Broeckx** et al. Intra-arterial versus intra venous contrast-enhanced computed tomography of the equine head. Revised manuscript (BMC Veterinary Research).
17. E. Coppieters, B. Van Ryssen, H. van Bree, G. Verhoeven, **B.J.G. Broeckx** et al. Tomographic Findings In Canine Elbows 1 With Erosion Of The Medial Compartment. Submitted to Veterinary Radiology and Ultrasound
18. C. Casper, H. van Bree, **B.J.G. Broeckx** et al. Computed tomographic findings on the physiologic parameters influencing the size of the pituitary gland in horses without PPID. Submitted to Equine Veterinary Journal.
19. E. Royaux, V. Martlé, **B.J.G. Broeckx** et al. Multislice Computed Tomography and Low-Field Magnetic Resonance Imaging for the Detection of Hydrated Nucleus Pulposus Extrusion in dogs. Submitted to the Journal of Veterinary Internal Medicine.

Active participation in international conferences

1. **B.J.G. Broeckx** et al. (2012). Relation between the FBN2 haplotype and phenotypical hip dysplasia. Advances in Canine and Feline Genomics and Inherited Diseases, 6th International conference, Poster.

2. **B.J.G. Broeckx** et al. (2015). Whole exome sequencing the dog. Advances in Canine and Feline Genomics and Inherited Diseases, 8th International conference, Poster.
3. **B.J.G. Broeckx** et al. (2015). Heuristic filtering of sequencing variants with the R-package "Mendelian". Advances in Canine and Feline Genomics and Inherited Diseases, 8th International conference, Poster.
4. H. Versnaeyen, Pieter Defauw, **B.J.G. Broeckx** et al. (2015). Progressive juvenile nephropathy in an 11 week old Bloodhound. Joint European congres of the ECVP and ESVP, Abstract.

Active participation in national conferences and workshops

1. **B.J.G. Broeckx** et al. (2013). Prevalence of genetic disorders in dogs from Belgium, the Netherlands and Germany. Knowledge for Growth, Ghent, Belgium, Poster.
2. Oral presentation at the Savab-Flanders najaarsmeting: work in progress entitled "De genetica van heupdysplasie ontrafeld: an ongoing research project". Drongen, Belgium, 2013.
3. Keynote speaker at the Assistance Dogs Europe/Assistance Dogs International conference entitled "Medical assistance for assistance dogs: some hot topics". Spa, Belgium, 2013.
4. Workshop at the Assistance Dogs Europe/Assistance Dogs International conference entitled "The genetics of hip dysplasia unravelled: hip, hip, hurray?". Spa, Belgium, 2013.

5. Lecture for the regional veterinary association “Rupelstreek-Klein Brabant” entitled: “Heupdysplasie, what the FOK, een praktische kijk op HD en andere genetische aandoeningen”. Recognized by the NGROD, 2 erkende bijscholingspunten, Rumst, Belgium, 2014.

6. **B.J.G Broeckx** et al (2015). Who lives in the human gut? Unipept Shotgun Metagenomics Analysis Pipeline. Annual Symposium Bioinformatics Institute Ghent Nucleotides to Networks, Ghent, Belgium, Poster.

7. Oral presentation at the Biotech seminar of the Ghent University College entitled “The genetics of hip dysplasia: development of the whole exome sequencing enrichment kit”, Ghent, Belgium, 2015.

8. Oral presentation at the 7th Flemish Breeding Day of the KU Leuven entitled “Genetisch (h)onderzoek: more than meets the eye”, Leuven, Belgium, 2015.

<u>Attendance at national conferences and workshops without active participation</u>

1. NCSA: workshop on hip, elbow and shoulder dypsplasia, Belgium, Ghent, 4 Sep 2010.

2. Small Animal Veterinary Association Belgium Flanders Conference 2011: From DNA to AND, Belgium, Heusden Zolder, 18 Mar 2011 – 19 Mar 2011.

3. Symposium on inbreeding and inherited diseases, Belgium, Ghent, 10 Sep 2011.

4. Seminar Nucleotides2Networks (Prof. Mar Alba): Long non-coding RNAs as a source of new peptides, Belgium, Zwijnaarde, 9 April 2014.

E. Educational experience

- Supervisor of practical courses in “Biotechnology”

- Supervisor of practical courses in “Pharmacognosy and Phytochemistry”

- Guest lecture on “canine hip dysplasia” in 2nd Master of Veterinary Medicine, Orthopedics (Professor B. Van Ryssen)(2013, 2014)

- Guest lecture on “canine hip dysplasia” in 1st Bachelor of Veterinary Medicine, Statistics (Professor L. Duchateau) (2015)

- Lecture on “Introduction to sequencing (overview of sequencing techniques)”, 14 augustus 2014, Faculty of Pharmaceutical Sciences, Ghent University (Organizer/lecturer)

- Lecture on “Introduction to sequencing (part 2: data analysis)”, 14 augustus 2014, Faculty of Pharmaceutical Sciences, Ghent University (Organizer/lecturer)

- Member of PhD supervisory committee of the dissertation presented in the fulfillment of the requirements for the degree of Doctor in Veterinary Medicine by Kaatje Kromhout (2015)

- Co-supervisor of master theses:

  * 2013: Elien Verelst, Master of Science in Bioscience Engineering Technology. “De associatie van genetische variatie in SOD1 en FBN2 met

objectieve heupdysplasieparameters bij de hond". Promotor: Frank Coopman. Co-promotor: Bart Broeckx.

⁎ 2014: Sylvie D'Hooghe, Master of Veterinary Medicine. "Tricuspidalisklepdysplasie bij een familie jacht Labrador Retrievers". Promotor: Valérie Bavegems. Co-promotor: Bart Broeckx.

⁎ 2014: Sara Henckens, Master of Veterinary Medicine. "Preliminaire studie ter ontwikkeling van een genetische test voor elleboogdysplasie" (2014). Promotor: Bernadette Van Ryssen. Co-promotor: Bart Broeckx.

⁎ 2014: Sara Henckens, Master of Veterinary Medicine. "Het effect van rotatie rond de longitudinale as op de beoordeling van heupdysplasie" (2014). Promotor: Jimmy Saunders. Co-promotor: Bart Broeckx.

- Reading committee of master theses:

⁎ 2014: Saskia Grammens, Master of Science in Bioscience Engineering Technology. "Genetische achtergrond cryptorchidie bij hond en paard".

⁎ 2014: Vicky Tas, Master of Science in Bioscience Engineering Technology. "Validatie Combibreed genetische test bij de Boerboel".

Dankwoord

Vier jaar geleden leek dit nog zo ver weg en met momenten zelfs onmogelijk tijdens de onzekere periode van de beursaanvragen. Ondertussen is echter reeds het einde van de predoctorale periode aangebroken. Het is onvoorstelbaar hoe snel deze vier jaren zijn voorbij gevlogen. Een doctoraat wordt gekenmerkt door hoogtes en laagtes, periodes waarin alles lukt en periodes waarin het lijkt dat je niet vooruit komt. Daar nu op terugkijkend, kan ik alleen maar zeggen: het komt (altijd) goed! Altijd blijven gaan!

Het is een speciaal moment als je dat boekje met jouw naam op daar ziet liggen: een combinatie van blijdschap, fier dat je het hebt voltooid, maar vooral ook dankbaar want dit werk zou nooit tot stand zijn gekomen zonder de nodige hulp en ondersteuning. Eerst en vooral, zou ik het Agentschap voor Innovatie door Wetenschap en Technologie (IWT) willen bedanken: mede dankzij hun geloof in het project en hun financiële steun ligt dit boekje hier.

Naast de financiële steun is dit vooral ook een project dat tot stand is gekomen dankzij vele gemotiveerde personen die elk op hun manier hebben bijgedragen. Als eerste in de lijn zou ik graag mijn promotor en copromotoren willen bedanken:

Prof. Dieter Deforce, u gaf mij de kans om als dierenarts in uw labo te doctoreren, de vrijheid om het project mee uit te stippelen en te groeien in wetenschappelijk onderzoek. Uw brede wetenschappelijk kennis en inzicht, gecombineerd met de nodige “drive” hebben dit project ongetwijfeld gestimuleerd. Een oprechte dank u!

Prof. Frank Coopman, ontsproten aan jouw gedachtegang, werd het de bron van de gemeenschappelijke passie: de zoektocht naar de genetische oorzaak voor heupdysplasie. Van samen grasduinen in een stapel rx’en tot het leren nemen van PennHIP, HD diagnostiek op zijn best! Op dit moment hebben we het nog niet gevonden, maar ik ben er van overtuigd dat er een hele stap in de goede richting is gezet.

Prof. Geert Verhoeven, waar is de tijd dat ik voor het eerst binnenstapte in Dierenkliniek Randstad. Vanaf het eerste moment was ik onder de indruk van je theoretische kennis gecombineerd met je chirurgische vaardigheden. Als student diergeneeskunde leek “de praktijk” de logische optie, tot jij de vraag stelde om te doctoreren. Niet iedereen heeft het geluk een persoonlijke mentor te hebben die je helpt te ontwikkelen, bij deze is het dan ook mijn beurt om je te bedanken!

Prof. Filip Van Nieuwerburgh, waarom doen we geen “whole exome sequencing”? Een goede vraag die de basis vormde voor het overgrote deel van het doctoraat. De nuchtere en kritische kijk op onderzoek, de sterke verhalen after hours: de ideale mix!

Tevens zou ik de andere leden van de lees- en examencommissie willen bedanken voor de tijd die ze genomen hebben om dit werk te lezen en te beoordelen. Speciale dank aan Dr. Wim Van Haeringen voor de goede samenwerking!

Ik heb het geluk mijn doctoraat te hebben mogen doorbrengen omringd door geweldige collega's, zowel op mijn "thuisbasis", de Faculteit Farmaceutische Wetenschappen, als op de Faculteit Diergeneeskunde. Samen hebben we gelachen, gevloekt en samen raken we vooruit. Logischerwijs komen we dan onmiddellijk terecht bij mijn "medebokaalbewoners" op het tweede. Liesbeth en Katleen, elks op jullie eigen manier onmisbaar voor het labo, maar bij beide de altijd "openstaande deur" voor vragen en om te lachen. Katleen, de "mama" van het labo. De manier waarop jij de begeleiding combineert van externen en interne projecten en de rust die je uitstraalt. Liesbeth, gedreven en doorzettingsvermogen. Zoveel praktische zaken regelen en ook nog tijd vinden om te combineren met je eigen research, desnoods in de late/vroege uurtjes: we gaan je missen tijdens je verblijf in de VS! Paulien, samen met mij begonnen en ons door de IWT procedure geworsteld en ondertussen ook aan het afronden: ik ben blij dat ik met jou het hele traject heb kunnen afleggen. Ellen, altijd enthousiast en vrolijk. Elisabeth, de goedheid zelve, harde werker en perfectionist! You can do it! Maarten, mede-koffieverslaafd, altijd enthousiast en brenger van de Japanse origami's en "chihuahuas"! Trees, altijd vrolijk en een geweldige lach. David, de wekelijkse leverancier van de groene cadeautjes, maar bovenal iemand waar je mee kan lachen en op kan rekenen. Christophe, het statistiek-clubje, loop-clubje en het R-clubje (ja, ja, het blijft R, zelfs al is het via Python ;-)): veel gemeenschappelijke interesses en altijd een plezier om "even" mee te babbelen. Ondertussen zit je "aan den overkant", maar laat dat je niet tegen houden om regelmatig binnen te springen! Laura, staat haar mannetje, weet wat ze wil! Yannick, ik kan mij zelfs niet meer voorstellen hoe we ooit computer problemen konden oplossen zonder jou. “When all else fails, read the manual” is bij ons ongetwijfeld, “just go to Yannick”! Dieter, mede R-adept, allround bio-informaticus en gewoon “ne sympathieke mens”. Ellen en Sarah, queens of the wet lab! Lieselot, altijd lachen! Senne, king of the road, of het nu op de fiets is of te voet! Sander, internationalisering van het labo en samen met Heleen, de toekomst! Evelien, nog maar net gestart, maar ongelooflijk gedreven en ik kijk vol vertrouwen uit naar de verdere uitwerking van je project en samenwerking. Saskia, Sylvie, Sabine, Petra, Leen, Eveline, Evelien ofte de collega’s van “den DNA” en Sofie: we hadden misschien iets minder contact, maar het was daarom niet minder aangenaam! Bedankt voor de leuke babbels en fijne momenten! Nadine en Astrid, altijd bereid om te helpen, van groene enveloppen, onkostennota’s tot VAT en overhead: zonder jullie was ik nooit door de administratieve soep geraakt. Inge, altijd lachen en vrolijk en bereid om alles te regelen.

Onze oud-collega’s, Mado, Marlies, Bert, Shahid, Pieter, Veerle, Yens, Nicky, we zijn jullie zeker niet vergeten, bedankt voor al de fijne momenten! Een speciaal woordje van dank voor Sandra: een ongelooflijke hulp tijdens het schrijven van mijn masterproef en nog meer tijdens de voorbereiding van mijn IWT. Zonder jou was dit project misschien nooit echt uit de startblokken geraakt.

Het woord dat dit doctoraat het beste kenmerkt is naar mijn mening niet DNA en zelfs niet heup, maar wel samenwerking. Aan de basis van dit doctoraat lag een brede samenwerking en de volgende personen hebben daar allemaal aan bijgedragen:

Onder leiding van Prof. Bernadette Van Ryssen, het team van orthopedie. Speciale dank aan mijn (oud-)bureaugenoten: Stijn, Lynn en Astrid. Op een teambuilding of gewoon op het werk, de sfeer zat er altijd in!

Onder leiding van Dr. Ingrid Gielen, het CT/MR team. Speciale dank aan Kaatje: ist oem da wij alletwiee van 't stad zen, k wieet et ni, mor gij zijt de max! Aqui, you had to cross an ocean and go through Spain to end up in Ghent, but I'm glad you did!

Onder leiding van Prof. Ingeborgh Polis, het team van anesthesisten. Zo slaapverwekkend jullie zijn voor de hond, zo fijn om mee samen te werken als mens!

Het team van de dienst cardiologie. Samen naar die ene klep staren, lekt hij of lekt hij niet, samen die ene tabel van meer dan 200 kolommen verwerken. Ik kijk uit naar het volgende onderzoek, the best is yet to come!

Onder leiding van Prof. Jimmy Saunders, het voltallige team van medische beeldvorming. Speciale dank aan Elke, samen trekken en sleuren op zoek naar de perfecte heupfoto van de onmogelijk te positioneren hond: het was niet altijd evident om het doel te bereiken, maar we kwamen er steeds!

Speciale dank ook aan Prof. Em. Van Bree voor het vertrouwen en de aangename samenwerking!

Naast collega's op de Faculteit Diergeneeskunde, waren een heleboel praktijkdierenartsen uit het binnen- en buitenland onmisbaar voor dit onderzoek. Wat ben je met sequencing als je niets hebt om te sequeneren? Hun bijdrage is van onschatbare waarde. Graag zou ik dan ook de volgende collega's en hun teams bedanken:

In België:

Dierenkliniek Sint-Jan – DAP aan de Heikant – Dierenkliniek Randstad – DAP Kleidal – Dierenkliniek Sint-Jan – Dierenkliniek Het Binnenhof – Huisdierchirurgie-Verdonck – Dierenkliniek De Vliet – Dierenkliniek Avanti/Vandecan – DAP Sonuwe – DAP De Bruycker-Criel – DAP Vetuatuca – DierenArtsenCentrum Assist – AC-DAP – Dierenkliniek Sanimalia – DAP Hulsterheide – Dierenkliniek De Bosdreef – DAP De Botermarkt – DAP Meirsschaut – Dierenartsencentrum Malpertuus – Dierenartsencentrum De Vledermuis – Dierenartsencentrum Herckenrode – Cabinet Buchet-Mathieu – Dierenkliniek Kerberos – DAP De Roeck Landen – DAP Hoogland – DAP Merckx – DAP Het Neerhof – DAP De Bosberg – Dierenkliniek Drogenboom – DAP Nachtegaele – Dierenkliniek Causus – DAP De Lovaart – Kasteel Vetcare – Centre vétérinaire Animalliance – Di Duca

In Nederland:

Sterkliniek Dierenartsen Ermelo – Dierenkliniek Brandersstad – Sterkliniek Dierenartsen Leeuwarden – DAP Horst e.o. – DAP de

Roosberg – Veterinair Centrum Holland Noord – KVGD Eersel – DAP De Meemortel – de Tweede Lijn – Dierenkliniek De Rashof – Dierenziekenhuis Drechtstreek – Sterkliniek Oss

In Duitsland:

Rosin Tiergesundheid

Speciale dank aan:

- Hans Nieuwendijk (de Tweede lijn): na een lange tocht door België en Nederland deed het bijzonder veel deugd om zo gastvrij ontvangen te worden in het verre Wilhelminaoord en mij te laten overnachten om uitgeslapen terug huiswaarts te kunnen keren.

- het voltallige team van Dierenkliniek Randstad: om mij gedurende zoveel jaren te laten proeven van het mooiste beroep ter wereld.

I would also like to thank Kerstin Lindblad-Toh, Jessica Alfoldi, Christophe Hitte and Thomas Derrien for our fruitful "exome" collaboration.

Tijdens dit doctoraat kwam ik dankzij Daphné in contact met de wereld van de assistentiehonden. Keer op keer was ik onder de indruk van de drive en kennis van de medewerkers, de trainers, de gastgezinnen en overige vrijwilligers en de wil om samen te werken en vooruit te gaan. Mijn uitdrukkelijke dank aan de teams van Hachiko – Dyadis – Scale Dogs – Vrienden der Blinden Koksijde – het Belgisch centrum voor Geleidehonden - Blindengeleidehondenschool Genk – Stichting Hulphond en Martin Gaus Geleide- en Hulphondenschool. Daarnaast bedank ik eveneens graag de geleiders van "Rescue Dog Belgium" en

Diensthondenvereniging Zennevallei. Daphné kan er zelf jammer genoeg niet meer bij zijn, maar ze mag trots zijn op wat dankzij haar gestart is.

Fysiek overal langsgaan zou meer dan een fulltime bezigheid zijn geweest. Dankzij de steun van de diergeneeskundige laboratoria, Medvet, Mediclab en Velab en in Nederland, ophaaldienst Miedema, werd de staalverzameling praktisch mogelijk gemaakt. De rol van het dr. Van Haeringen Laboratorium kan zeker niet geminimaliseerd worden: het praktische, overzichtelijke beheer kon niet beter gedaan worden! Ook mijn thesisstudenten Evelien en Sara zou ik graag willen bedanken voor hun hulp en enthousiasme!

Jullie bijdrage was misschien onrechtstreeks, maar daarom niet minder belangrijk: mijn mede-jaargenoten van het jaar 2011: Ruth, Sofie D., Sofie M. en Olivier: van samen in de les tot samen afstuderen en uiteindelijk 3 verschillende richtingen uitgegaan. Ruth en Sofie D.: bedankt voor de ontspannende wandelingen met de hondjes en nog zoveel meer. Ondertussen zitten jullie niet meer aan de UGent, maar aan de andere kant van België en zelfs aan de andere kant van de wereld, even afspreken om te wandelen wordt dus wat moeilijker, maar dat komt ongetwijfeld nog goed! En aangezien we met slechts een half uur verschil synchroon aan het afleggen zijn: succes, you can do it, Sofie! Sofie M., bedankt voor de goede zorgen voor onze hondjes en de toffe babbels. Olivier, maat, ex-kotgenoot, merci!

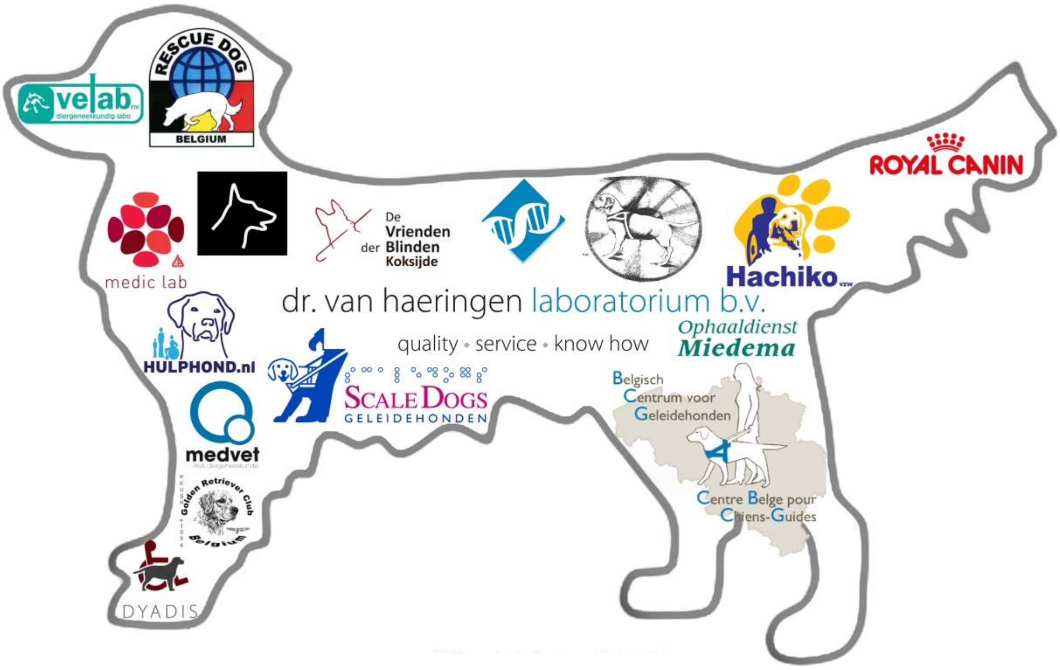
Papa, mama, alles beschrijven wat jullie hebben gedaan is niet mogelijk. Van de steun tijdens het studeren, tot het vervoer naar alle mogelijke hobby's, jullie hebben mij altijd alle kansen gegeven en mede dankzij jullie sta ik waar ik nu sta. Bedankt! Ben (en Charis), Michiel (en Nicky), Hanneleen: bedankt voor de broodnodige ontspanning, de geweldige Broeckx-humor en het samen knutselen van het "kanteldinges"!

Bompa en Bomma, hoe ik er op gekomen ben om dierenarts te worden, weet ik niet meer, maar jullie hebben daar ongetwijfeld een belangrijke rol in gespeeld: van het slapen bij Douchka tot het opnemen van alle afleveringen van "All creatures great and small", de liefde voor dieren werd er met de paplepel ingegoten. Bedankt!

Zoals vaak is de laatste persoon zeker niet diegene die de kleinste bijdrage heeft geleverd. Astrid, research verloopt in ups en downs, maar op elk moment kon ik op je onvoorwaardelijke steun rekenen, was je daar om me te helpen relativeren, hielp je mij vooruit. Je bijdrage is onschatbaar, niet alleen voor het doctoraat, maar gewoon in mijn leven. Bedankt voor alles!

Snoopy, Ruby, Bathida: woef!

Met dank aan:



velab
diergeneeskundig labo
RESCUE DOG
BELGIUM
ROYAL CANIN
medic lab
De Vrienden der Blinden Koksijde
Hachiko vzw
dr. van haeringen laboratorium b.v.
quality • service • know how
Ophaaldienst Miedema
HULPHOND.nl
SCALE DOGS
GELEIDEHONDEN
Belgisch Centrum voor Geleidehonden
Centre Belge pour Chiens-Guides
medvet
Golden Retriever Club Belgium
DYADIS