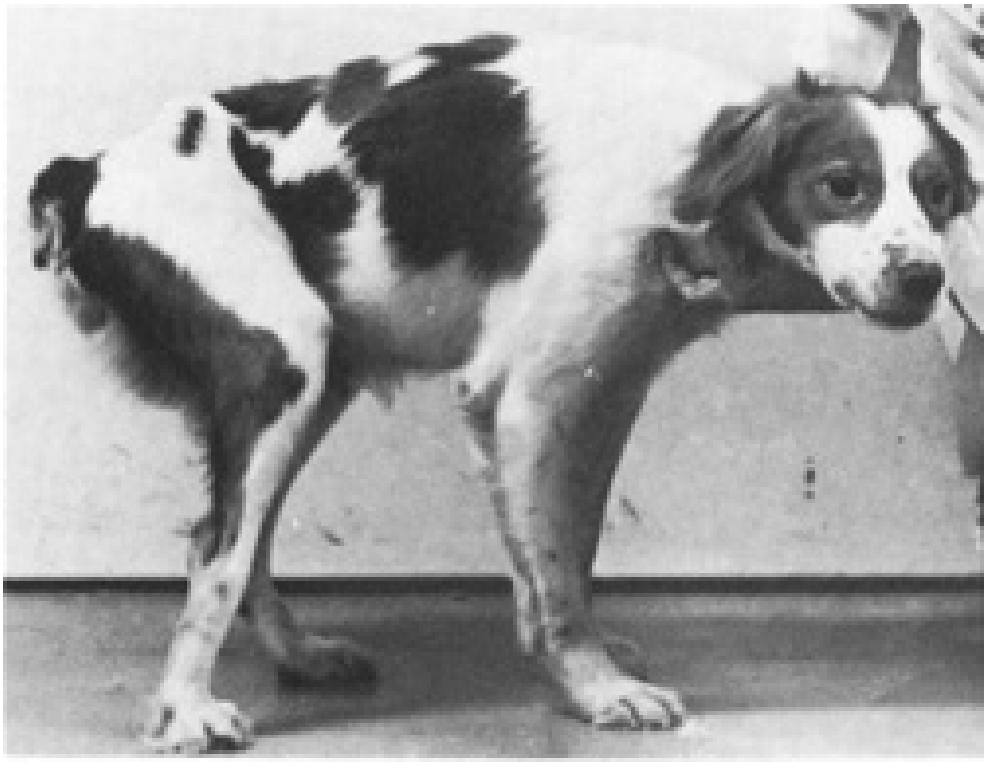
UPPSALA
UNIVERSITET

Genetic mapping of Hereditary Canine Spinal Muscular Atrophy



Axel Ericsson

Degree project in biology, Master of science (2 years), 2012
Examensarbete i biologi 45 hp till masterexamen, 2012
Biology Education Centre, Uppsala University, and Broad Institute of Harvard and MIT
Supervisors: Kerstin Lindblad-Toh and Noriko Tonomura
External opponent: Carl-Johan Rubin

Table of Contents

Summary

Motor neurons disease (MND) is classified as any disease that involves degeneration of motor neurons. Motor neurons can be classified into the lower motor neurons, which are present within the spinal cord and upper motor neurons located in the motor cortex and cerebellum. The most common motor neuron diseases are Spinal Muscular Atrophy (SMA) and Amyotrophic Lateral Sclerosis (ALS). The SMA only targets the lower motor neurons, and mutations found in the Survival Motor Neuron (*SMN-1*) gene is the major cause of SMA, while ALS targets both upper motor neurons and lower motor neurons. Approximately, 20% of the familiar ALS cases are caused by mutations in the *SOD1* gene. The clinical manifestations include muscle weakness and paralysis and subsequently death due to respiratory failure.

Hereditary Caine Spinal Muscular Atrophy (HCSMA) is a motor neuron disease targeting the lower motor neurons, demonstrating similar pathophysiology as human SMA and ALS. HCSMA originates from a spontaneous mutation in a Brittney spaniel population. The single proband carrying the mutation was experimentally bred in California many years ago, generating the HCSMA disease pedigree. HCSMA is a co-dominant trait, meaning a varied phenotype is observed dependent on the number of disease alleles carried. Two phenotypes have been characterized; an accelerated form of the disease defined as “Homozygous mutant” and a late onset phenotype defined as “heterozygous mutant”.

Here we present the genetic mapping of HCSMA, starting with a Genome Wide Association Study (GWAS). The trait was mapped as a recessive trait, designating the accelerated dogs as affected and the late onset animals as carriers, identifying a 10Mb region on chromosome 13. In addition, a homozygosity analysis was executed and identified a segment that overlapped with the candidate region on chromosome 13. To further narrow down the region, the recombination breakpoints were investigated in the GWAS cohort, which recognized a ~4Mb associated region. A targeted sequencing capture was performed on 4 dogs with the accelerated disease phenotype and 4 dogs with later onset disease, as well as a fine-mapping study including genotyping of 475 Single Nucleotide Polymorphisms (SNPs) across the ~4Mb region. The sequencing variants were filtered by several parameters to narrow down the number of candidates, the variants passing the filter were selected for extended pedigree analysis, where the SNPs were genotyped using Sequenom technology and indels by Sanger sequencing.

No causative mutation was found, however the region where the mutation must reside, was narrowed down by haplotype analysis to a ~2Mb region. Both the accelerated disease cohort and four dogs with late onset disease, are homozygous across this ~2Mb segment in regards of ~300SNPs, indicating the mutation to be present on the common haplotype. No variants were discovered by the targeted sequencing within the ~2Mb region concordant with the mode of inheritance.

The ~2Mb segment contains 21 genes, of which four are known to be expressed within the Central nervous system; *REST, TDGF1, MSL2* and *HopX*. Further analysis should include investigation of these candidate genes.

Introduction

Discovering the genetic cause of human disease is a major component in the area of personalized medicine, and is important for both diagnosis and for development of treatment strategies. The DNA molecule was discovered by James Watson and Francis Crick almost 60 years ago, and DNA, which provides the blueprint of life has ever since been carefully studied and characterized[1].

Genetic mapping

One strategy to identify genomic regions associated with disease utilizes genetic markers, distributed across the genome. These regions can be located by investigating co-segregation of markers with disease in a pedigree[2], also referred to as linkage studies. Another approach is association studies, where genetic markers in the genome of cases and controls are compared to identify the markers that are associated with the disease status based on differing allele frequencies in the two groups.

One of the strategies used in the early genetic mapping studies is the Restriction Fragment Length Polymorphism (RFLP)[2] markers. The DNA is digested by restriction enzymes and run on a gel producing a large set of fragments with a broad spectrum in sizes; the presence or absence of an enzyme restriction site will produce different size fragments. The RFLP marker is therefor a bi-allelic marker, meaning there are only two alleles present at each locus.

Another marker used for Linkage analysis Simple Sequence Length Polymorphism (SSLP) marker, sometimes called a microsatellite[3]. Microsatellites are up to a few hundred base pairs long, many with a repeated core sequence of 2-4 base pair (bp)[4]. The advantage of the SSLP markers over RFLP is its ability to detect the presence of multiple alleles at each locus, since these sites have high variability due to increased deletions and insertions rate[3]. This enabled further differentiation between individuals facilitating linkage studies. As the Polymerase Chain Reaction (PCR) technology emerged, the use of RFLP marker became inconvenient due to the size of the fragments and SSLP became the most commonly used marker. The first linkage map was created including 150 highly polymorphic sites enabling linkage mapping of several Mendelian diseases including Huntington's disease[5] and cystic fibrosis[2,6].

To perform a linkage study, the family structure and the mode of inheritance for the disease, such as recessive, dominant or co-dominant, needs to be assessed. When an individual needs to be homozygous for the disease allele to be affected, it is called a recessive disease. Individuals who are heterozygous for the mutation are phenotypically normal, and considered to be a carrier of the recessive disease. When an individual only needs one disease allele to be affected, it is a dominant trait. A co-dominant trait is when the severity of the phenotype is dependent on the number of disease alleles an individual carry. Sex linked disease is when the mutation is linked to either Y-chromosome or X chromosome, which can be either dominant or recessive, and usually produce different inheritance patterns in males and females.

Multiple statistical approaches have been developed to perform linkage mapping of causal mutations for diseases. They are based on likelihood calculations, and dependent on family structure and the mode of inheritance of the disease.

Genome wide association studies

As array technologies emerged for genotyping, the Single Nucleotide Polymorphism (SNP) marker was introduced. A SNP is a bi-allelic marker similar to the RFLP, but SNPs are more densely distributed across the genome, and thus give some advantages over RFLP markers. As the resolution of the arrays has increased, the regions of interest can be narrowed down further. Whole genome array containing 100 of thousands of SNPs was introduced in the beginning of the 2000, and the initial Genome Wide Association Study (GWAS) was performed [7,8]. Generally GWAS does not consider family structure, and it just compares a large set of markers in cases and controls to identify significantly associated alleles with the disease status[9].

The approach enables studies of complex disease and identification of multiple regions linked to a disease[10]. Since such a large number of SNPs are included, correction for multiple testing and increased significance threshold are needed to filter out false positives[11]. The increased significance threshold is problematic in regards to the detection of rare alleles, so consequently GWAS is more suitable for common variants [12]. Common variants are alleles that occur in the population with a frequency of 1-5%, and rare variants with allele frequency of less than 1%. Some concerns regarding GWAS are the increased rate of genotyping error[12] and susceptibility for population stratification[13]. Population stratification occurs if individuals included in the analysis are related, and levels of relatedness in cases and controls are not well balanced. Population stratification can lead to significant enrichment of certain haplotypes in cases and/or controls, leading to false association signal[13].

Family based GWAS combines GWAS and Linkage analysis, which can detect genotype errors based on pedigree, and the population stratification is no longer an issue. When the transmission of parental alleles to the offspring does not follow the Mendel's law of segregation, the marker is flagged to have Mendel inconsistency, and can be corrected for in the association/linkage analysis[14].

The transmission disequilibrium test (TDT) is a test for both association and linkage. SNP markers with rare alleles that are missed in a GWAS can cluster in a small region of genome and can be detected by linkage analysis[14].

A GWAS will identify a genomic region that is associated with the disease, however, to find the causative mutation sequencing of the associated region is necessary.

Sequencing

The first generation sequencing technique was invented by Fredric Sanger and is referred to as Sanger sequencing[15]. The combination PCR, Sanger sequencing[15] and cloning

technologies such as the bacterial artificial chromosome (BAC)[16] enabled the start of the Human Genome Project[17]. The project started in the 1990s and finished in 2003, costing ~2.7 billion dollars to complete[18].

The major disadvantage of Sanger sequencing is the use of a single amplicon leading to limited throughput and making large-scale sequencing projects laborious and expensive. Several different next generation sequencing technologies have rapidly emerged in the last 6 years, leading to high-throughput sequencing and has revolutionized the field of genetics. The technologies consist of various approaches and involve combinations of sample preparation, sequencing, imaging, sequence assembly and alignment. Currently, the most widely used next generation sequencing technology is the Illumina platform which workflow is described below.

To construct a sequencing library, DNA is fragmented by sonication and selected for a specific size, then sequencing adaptors are ligated onto the fragments containing primer sequences and barcodes for each pool. The library is amplified on a glass slide containing covalently bound forward and reverse primers by the technology called "Bridge PCR". The amplification processes produces about 100-200 million templates clusters across the glass slide[19].

The Illumina platform uses a cyclic reversible terminator approach. When the DNA polymerase incorporates fluorescently modified nucleotide complementary to the template base, the reaction is terminated. The nucleotide incorporated is registered by imaging, and the terminator nucleotide is cleaved off and the processes are repeated[19]. The Illumina platform utilize a paired-end sequence approach where the DNA fragment is selected with a insert size ranging from 200-500 bp and is sequenced 76-150 bp from each side. The orientation of the pairs and deviation from the expected insert size can be used for detection of structural variants.

As whole genome sequencing is still relatively expensive when generating good coverage, and extensive time for assembly and alignment procedures is required, several methods have been developed for targeted enrichment, where specific regions of the genome are selected for sequencing. This approach can generate deeper coverage of the targeted regions, requires less time for data analysis, and can be more cost effective. Several different types of targeted enrichment technologies are available, in array-based format or in solution phase[20]. In most methods used for targeted sequencing probes are designed against the region of interest. The targeted sequencing has enabled the capture the genome-wide coding regions (exom capture), facilitating discovery of causal mutations of several Mendelian disease[21].

Dog as an animal model

Since the domesticated dog shares a broad spectrum of disease with human, receives comprehensive medical care, and is exposed to similar environmental factors as human, this makes the domesticated dog a suitable animal model to study human diseases[22]. In addition, the domesticated dog has a unique genomic structure, as a result of having narrow population bottlenecks, and artificial selection for certain physical and behavioral traits that are desired for each breed characteristics (*fig.1*). These factors have contributed

to the establishment of over 400 distinct breeds that are highly diverse across the breeds, but very homogeneous within a breed[23].

The dog has been of interest for genetic mapping studies for over a century, since the first inherited disease was described in the beginning of the 1900s[24]. In 2005 the sequencing and characterization of the dog genome was completed by the Broad Institute of Harvard and Massachusetts Institute of Technology[25]. A purebred female boxer was selected due to low rates of heterozygosity, which facilitates genome assembly. The female gender was preferred due to the normal coverage of the X chromosome. The dog genome consists of a total of ~2.4 billion bases distributed over 38 autosomal chromosomes and the sex chromosomes (XY). The genome assembly had a coverage of 7.5x including 99% of the genome[25]. Additional whole genome sequences at lesser coverage obtained from a poodle and 9 additional dogs from 9 different breeds were analyzed for SNP discovery[25]. The development of a SNP array for GWAS was made possible by the resources created as part of the dog genome project.

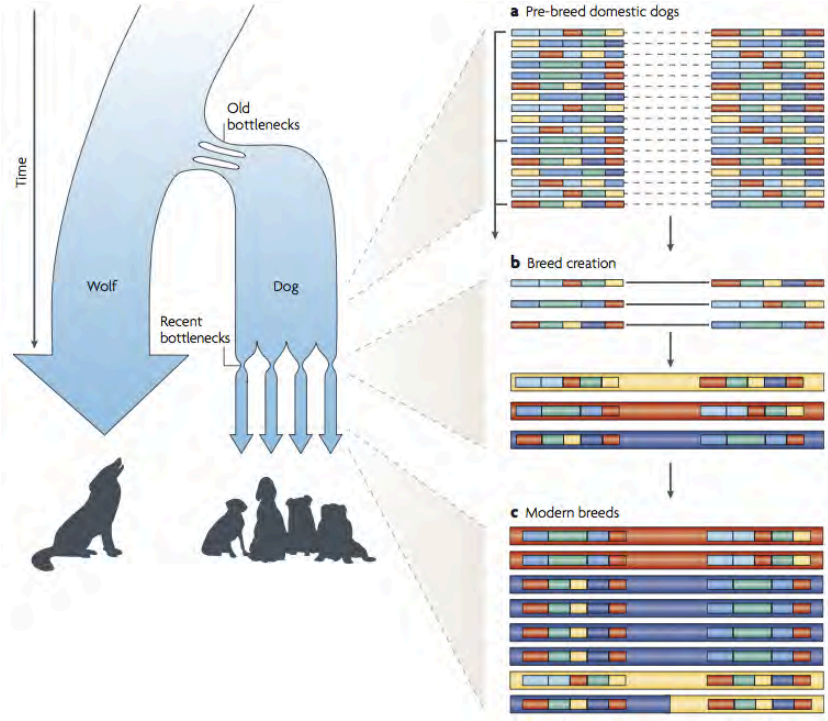


Figure.1 The haplotype structure of the domesticated dog. *(**a**)Two major population bottlenecks have occurred during the dogs population history, one introduced as the dog were domesticated from the wolf population and a more recent population bottleneck ~200years ago at breed creation. Prior the breed creation the population contained short haplotype blocks since the population is relatively old. (**b**)At breed creation specific set of alleles were selected for and each breed are very homogenous in regards of these haplotypes, consequently a few long haplotypes are observed. (**c**)The modern breed has yet undergone extensive recombination and the haplotype remains relatively intact, however by a comparison across the breeds the shorter ancestral haplotype can be distinguished.*

The breeding schemes used to create each breed, and population bottlenecks at the time of domestication and at the time of breed creation has created an exceptional linkage disequilibrium (LD) and haplotype structure in the genome of the domestic dog[25]. The recent bottleneck occurring at breed creation has created long haplotype blocks, with an average length of 0.5-1Mb with 3-5 haplotypes within each breed[24]. Since the breeds

are relatively young ~200 years old, and have not been exposed to extensive recombination events the haplotypes remain largely intact.
However the dog population as a whole is roughly 15 000 years old and cross breed comparison identifies 3-5 ancestral haplotypes every 10kb[25].This history of domestication and breed creation in domesticated dog results in a unique genome structure highly suitable for conducting GWAS.

Since the domesticated dog population has a unique LD and haplotype structure, a two-step genome-wide association mapping approach can be applied. The long LD within each breed enables a small sample set of dogs and fewer markers compared to humans[26], since fewer markers are needed to tag the longer and fewer number of haplotypes. This will result in identification of larger associated segments. Since disease risk alleles are most likely dated prior to breed creation, a fine mapping study is carried out including a larger set of samples from different breeds that share the same phenotype in order to locate smaller associated regions[27]. When different breeds share the same disease-causal haplotype, this approach would increase the resolution to ~10kb similar to human studies[28].

One concern in GWAS is that ~ 6% of the dog genome are homozygous regions larger than 500kb within each breed, which are undetectable in association studies. If the causal mutation is present on a common haplotype shared across the cases and controls, this approach would fail to map the associated region[27]. If the causal mutation is present within these excessively homozygous regions, additional breeds can still be included to carry out homozygosity mapping to narrow down the region of interest. One example is the extremely wrinkled skin of the Chinese Shar-Pie breed[27].

Motor neuron disease
Motor neuron disease involves all diseases involving the degeneration of motor neurons, and the most common diseases are Spinal Muscular Atrophy (SMA) and Amyotrophic Lateral Sclerosis (ALS). There are four neural circuits regulating movement, lower motor neurons and lower circuits are both present in the spinal cord and upper motor neurons located in the motor cortex and cerebellum.

The lower motor neurons have their cell body in the ventral horn of the spinal cord. The cell body has axons that connect with the endplate, forming the neuromuscular synaptic junction (NMJ), which innervates the muscle fibers. The axonal transport occurs from the cell body to the NMJ (anterograde direction) and from the NMJ to the cell body (retrograde-direction). The axonal transport can either be rapid or slow dependent on the material transported. The fast axonal transport[29] is bidirectional while the slow axonal transport[30] only is exhibited in anterograde direction. Vesicles containing neurotransmitters are transported rapid while tubulins and cytoskeletons compartments are transported slowly.

Signaling within the central nervous system as well as at the motor neuron endplate is mediated by synaptic transmission, which is regulated by several neurotransmitters such as GABA, glutamate, glycine, acetylcholine and serotonin. The neurotransmitters are

transported in vesicles within the pre synapse where the vesicles fuse with the presynaptic membrane consequently releasing the neurotransmitter into the synaptic cleft. The release of the neurotransmitters is a very complex process involving a large set of proteins, however a key component in release is the influx of Ca^{2+} into the presynapse[31].

Specific excitatory receptors are located on the post synapse and contribute to induction of an action potential leading to neural signaling. The action potential is an electric pulse caused by a transient change in the membrane potential. During the resting state of neurons a membrane potential is maintained by ion pumps creating a charge of -70mV relative to the outside. Upon activation, the ion channels regulating the permeability of Na^+ and K^+ opens, and the current created by the influx of Na^+ and efflux of K^+ are responsible for depolarization of the cell, creating the action potential that moves from the point of entry to axonal end terminals[32] . In the ventral horn glutamate is the main excitatory neurotransmitter and in the endplate the acetylcholine (ACh) is release enabling the contraction of muscle fibers.

Spinal muscular atrophy

SMA is a recessive autosomal disease[33] in human, with a carrier frequency of 1:35 and an incidence of 1:6000[34]. This makes SMA the major genetic cause of child mortality[34]. The disease is manifested by the degeneration of the lower motor neurons, specifically the a-motor neurons, leading to hypotonia and muscle weakness and eventually respiratory failure. The majority of SMA cases are caused by loss of function of the survival motor neuron 1 (*SMN-1*) by frame-shift and point mutations.[35] The disease is classified into three phenotypes dependent of age of onset and severity of the disease. Humans carry *SMN2*, a nearly identical gene to *SMN1*, where *SMN2* differs from *SMN1* in regards of mRNA splicing, and the majority of *SMN2* transcripts lacks exon 7 and encodes a dysfunctional protein[33]. However, because it can express low levels of correctly spliced transcripts producing a functional protein identical to SMN1, *SMN2* is the major SMA-modifying gene. Increased copy number of *SMN2* has been associated with a milder phenotype[34]. The treatment strategies developed for SMA consist of approaches to abolish the exon skipping in *SMN2* and increase the expression of the full-length transcript[36].

Amyotrophic lateral sclerosis

ALS is characterized by muscles degeneration and progresses to paralysis and subsequently death due to respiratory failure, as results of the degeneration of lower motor neurons in the spinal cord and brainstem or upper motor neurons in motor cortex[37]. The incidence of ALS is uniformly distributed across the world with an incidence of two per 100,000 individuals with an average survival time of 3 years from diagnosis[37].

ALS can be classified into sporadic ALS (SALS), which contributes to about 90-95% of all cases, and familiar ALS (FALS) corresponding to 5-10%. The main factor contributing to FALS is mutations in the *SOD1* gene, which is responsible for ~20% of the cases[38].

Multiple cellular systems are known to be affected in ALS patients, and interactions between these cellular processes are believed to contribute to the pathogenesis of ALS.

Elevated levels of oxidative damage to protein[39], lipids[40], DNA[41] and mRNA have been observed in ALS patients. Reactive oxygen species (ROS) is produced during several cellular processes, resulting in damage to cells. The cellular system have multiple mechanisms, including SOD1, to remove and repair damaged caused by ROS. A dysfunction in the protective mechanisms, such as mutations in SOD1 can lead to extensive exposure to ROS, causing oxidative stress, which contributes to cell death. Protein aggregation of phosphorylated neurofilaments in the proximal dendrites has been observed in mSOD1 mice[42]. It has been suggested that the combination of mitochondrial dysfunction, excitoxocity and aggregation of phosphorylated neurofilaments contributes to impaired axonal transport, a key feature in ALS pathology[43-45].

Several other cellular events have been implicated in ALS such as deregulated transcription[46] and endosomal trafficking[47], neuroinflammation[48] and endoplasmatic stress[49]. ALS is a complex disease and interplay between a large numbers of cellular processes contributes to the diseases. It remains challenging to differentiate the initial causal effects and the secondary effects that accumulate as the disease progresses. Deeper understanding of these cellular processes needs to be investigated to be able to develop treatment strategies

Hereditary Canine Spinal Muscular Atrophy

Hereditary Canine Spinal Muscular Atrophy (HCSMA) is a motor neuron degenerative disease, which spontaneously occurred in a family of pure bread Brittney spaniels in the 1970s[50]. Initially three phenotypes were characterized; accelerated, intermediate and chronic HCSMA, dependent on the rate of disease progression[51]. The chronic phenotype was only observed within the purebred Brittney spaniel population and may have been caused by a modifying or epistatic gene, producing a less sever phenotype. Since it was diminished when the dogs were outcrossed to a beagle population[52]. The hybrids of Brittney spaniel and Beagles are called “BrigglesX”.

An experimentally bred cohort was established by consanguinity mating, and consisted of 125 animals. An autosomal co-dominant inheritance pattern was discovered[52]. The sex ratios of the affected animals appear to be normal, hence no indication for linkage to the sex chromosomes[52]. It was considered to be a co-dominant disease, as severity of the disease correlated with the animal’s disease allele status determined by the pedigree, where two phenotypes were seen; accelerated onset and late onset. The dogs with the accelerated progression of the disease are phenotypically defined as “homozygote mutant”, and the dogs that produced homozygous offspring are defined as obligate “heterozygote mutant”.

Clinical manifestation and pathology

HCSMA shares many of the clinical manifestations and pathology of both human ALS and SMA. The disease targets the lower motor neurons similar to SMA with an age of onset of 6-8 weeks in accelerated form, where weakness in hind limbs are observed, and the disease progresses to quadriplegia by 16 weeks and subsequently death due to respiratory failure approximately in 21-22 weeks[53]. The disease shares similarities with

ALS in regards to pathological observations, such as impaired axonal transport[54], aggregation of phosphorylated neurofilaments[53], and increased oxidative stress[55]. Electrophysiological studies observed that tetanic failure occurred prior to any signs of motor neuron degeneration, implying that the disease starts with impaired endplate current caused by dysfunction of the release of the neurotransmitter ACh[56,57].

It has been suggested that HCSMA targets all types of cellular functions equally, but that the aggregation of neurofilaments are observed later in the disease progression, due to the longer turnover time associated with slow axonal transport. On the contrary, the neurotransmitters are transported with the fast axonal transport, and are therefore impaired earlier[58]. To determine if the human SMA and HCSMA arise from the mutation in the same gene, a linkage mapping study was carried out, which discovered that HCSMA is not linked to *SMN1* and does not share the same genetic basis as human SMA[59]. The *SOD1* gene was also investigated since it corresponds to 20% of the FALS cases in human, however it was not associated with HCSMA[60]. The striking similarities in disease phenotype between human SMA and HCSMA, and in pathophysiology between human ALS and HCSMA indicate HCSMA is a good disease model. Identification of the genetic basis of HCSMA would provide further insight of these motor neuron diseases in human.

Earlier Result

In collaboration with Dr. Marty Pinter of Emory University, the group led by Dr. Kerstin Lindblad-Toh at the Broad Institute conducted a GWAS study of HCSMA to identify the causal mutation of HCSMA.

Genome-wide association study

A family-based genome-wide association study was carried out using the Canine HD Illumina array comprising 170,881 SNPs, and included 16 animals with accelerated disease (homozygote mutant) and 15 obligate heterozygotes that are the parents of the 16 homozygotes. The animals included in the analysis spanned 11 generations with an even gender distribution of 16 females and 15 males (*fig.2)*. The study was design to map the disease as a recessive trait, assigning the late onset animals as carriers and the accelerated animals as homozygous-affected for the mutation.

The GWAS data was analyzed with PLINK analysis software[61]. First, some filters were applied to the data to remove SNPs and individuals with missing data over 10%. A Mendel error check was also performed, and a family trio early in the pedigree with over 10% Mendel errors was excluded from further analysis (B1,FB4, B5) .

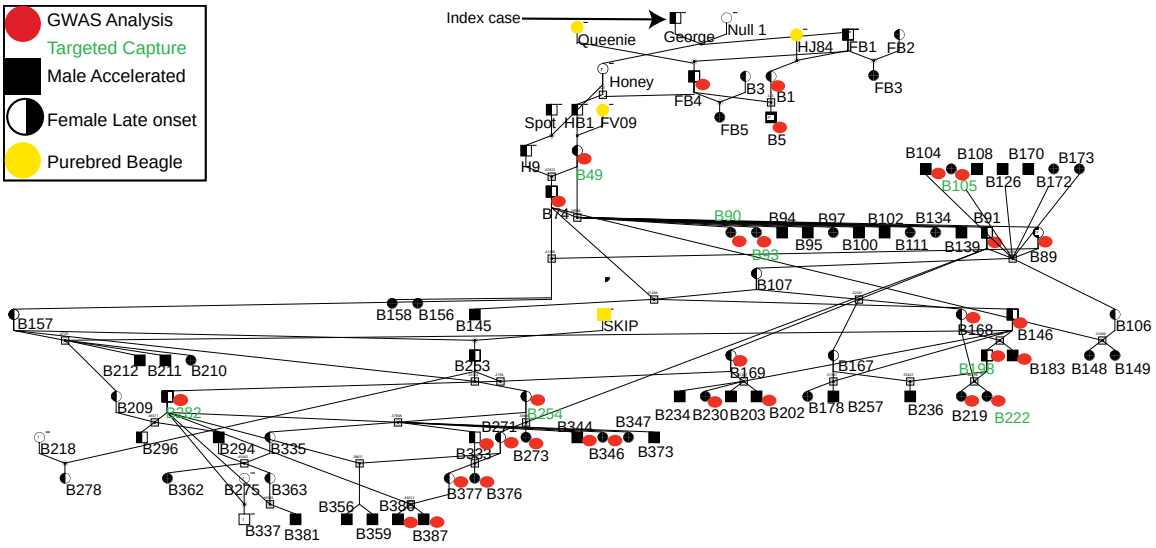


Figure.2 Pedigree of HCSMA. *The yellow fill indicates purebred beagle, and animals selected for the GWAS study are marked with red circle. The eight dogs included in the targeted capture of the ~4Mb region are annotated with green text. The majority of the pedigree was included in the fine mapping study excluding the purebred Beagles and George, Null1, Honey, Spot, HB1 and H9. The B1, FB4 and B5 trios were excluded from the GWAS and further analysis, due to an increased number of Mendel errors, suggesting incorrect identifiers of the sample(s).*

A Transmission Disequilibrium Test (TDT) was performed which identified an associated region of ~10Mb on chromosome 13 (*fig.3*). The allele frequency was also investigated separately in the accelerated and late onset cohort, across the whole genome, to carry out a homozygosity mapping analysis. The region of interest were expected to have a minor allele frequency (MAF) of ~50 % in the late onset cohort, reflecting all animals being heterozygous for the causal mutation. The accelerated cohort was expected to have an MAF at 0% due to full homozygosity for the causal mutation. The only region fulfilling this requirement was observed on chromosome 13 overlapping with the associated region identified by the TDT analysis (*fig.4*). To further narrow down the region, recombination breakpoints were investigated by examining the haplotypes. Within the accelerated disease cohort, two recombination breakpoint events were identified (B105, B222), defining a ~4Mb associated region on chromosome 13 49,000,000-53,000,000 (*fig.3*).

GWAS and breakpoint analysis of the HCSMA pedigree

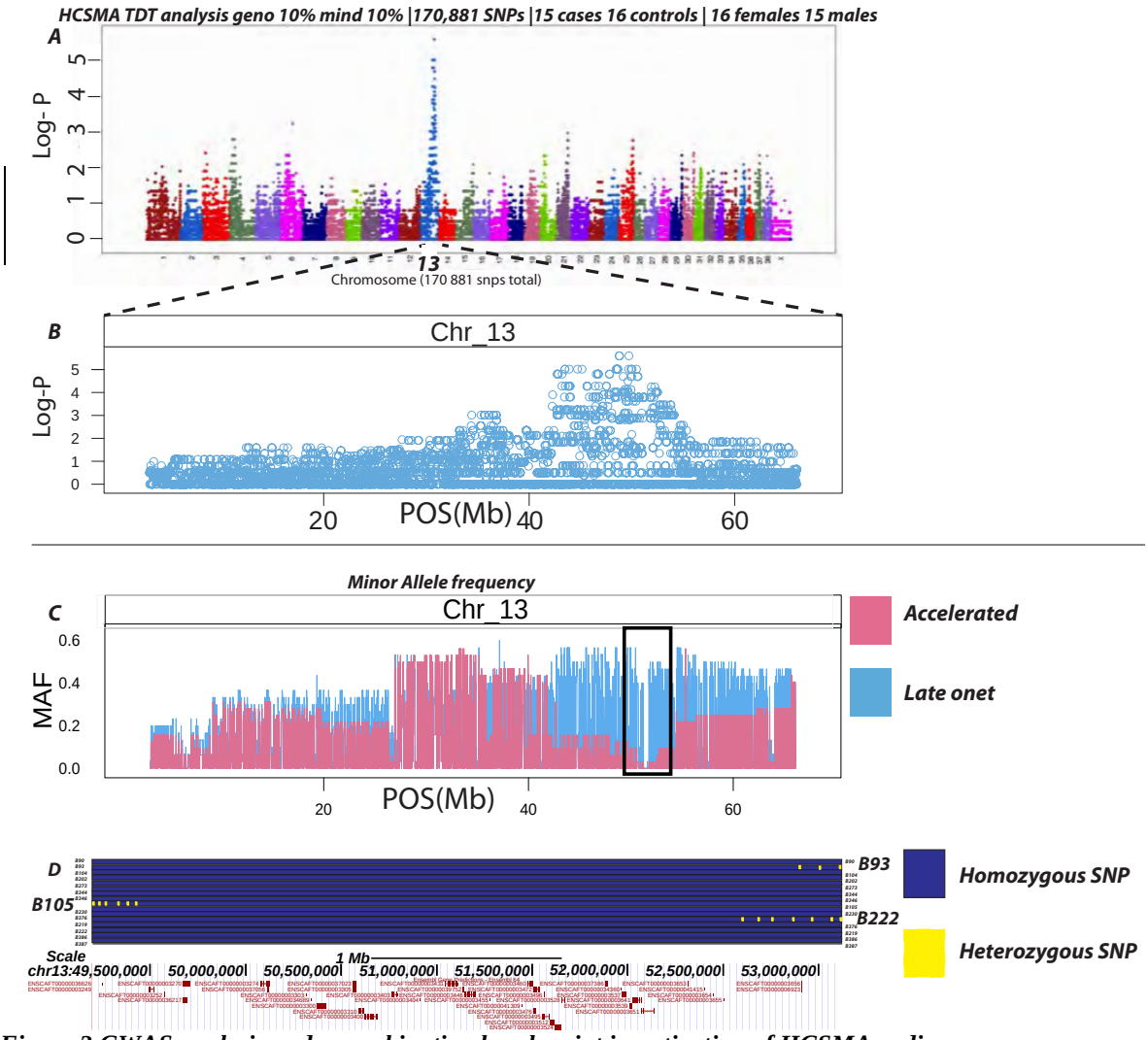


Figure.3 GWAS analysis and recombination break point investigation of HCSMA pedigree*.* ***(A)*** *A GWAS analysis of 15 cases and 16 controls were performed across 170,881 SNPs, represented in a manhattan plot, with log -P values on y-axis and the chromosomes on x-axis, with a significant association on chromosome 13.* ***(B)*** *An enlargement of chromosome 13 with the log- P on the y=-axis and the base pair position on the x-axis, identified a ~10Mb region.* ***(C)*** *The minor allele frequency was calculated separately in the accelerated cohort (blue) and the late onset cohort (red), y-axis indicates minor allele frequency and x-axis base position indicating a region overlapping with previous GWAS results. The black box indicates a fixed region between accelerated cohort and the late onset cohort.* ***(D)*** *Investigation of recombination breakpoints within the accelerated disease cohort, reveal a ~4Mb region on chromosome 13 defined by B105 and B222. The blue fill designates a homozygous SNPs and the yellow fill denotes a heterozygous SNPs.*

Targeted capture and fine mapping

A targeted capture sequencing experiment was performed using Nimblegens array technology. Four animals from the accelerated cohort and four animals from the late onset cohort, including the two animals with the recombination breakpoints defining the

associated region were selected. In parallel with the targeted capture, a fine mapping study was performed including 77 dogs from the same pedigree. A total of 475 SNPs across the associated region of ~4Mb were genotyped in those 77 dogs. The 77 dogs included 52 dogs with the accelerated disease and 25 dogs with the late onset disease spanning 11 generation.

Aim

The aim of the master thesis is to identify the causal mutation of HCSMA. The thesis works includes the downstream data analysis of the fine mapping study and targeted re-sequencing capture and additional functional studies.

Results

Targeted Sequencing results

The Nimblegen targeted capture was carried out on four dogs with the accelerated disease and four dogs with the late onset in the ~4Mb associated region identified by GWAS. The sequencing was performed on the Illumina platform utilizing the paired end technology with a read length of 76bp and an insert size of 200-250bp. The coverage for the dogs included in the analysis is represented in *fig.4.* The majority of the samples had an average coverage above 50x with the exception of two dogs (B105 and B222) with accelerated disease.

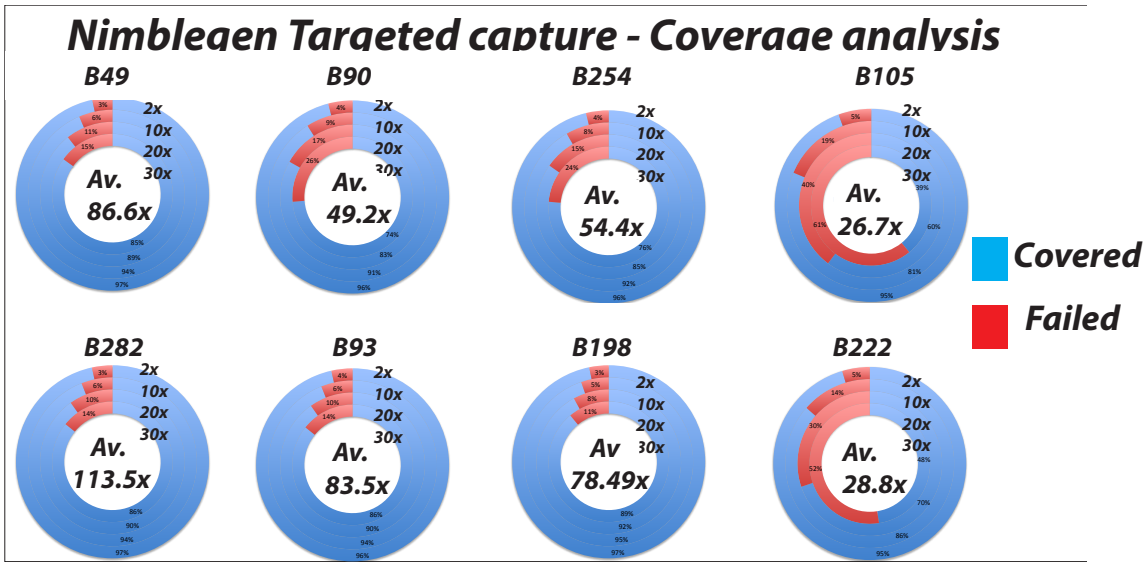


Figure.4 Coverage assessment.
The genomic coverage depth for each sample, red indicating percentage of non-covered regions blue indicates covered region. Each sub-circle corresponds to threshold of 2x, 10x, 20x and 30x genomic coverage.

Sequence analysis

The analysis of the sequences produced by the Illumina technology, often referred to as "next generation sequences" includes several steps of computational analysis; sequencing alignment, base calling, variant discovery, genotyping, annotation, effect prediction and filtering. The sequencing data was processed through the Genome Analysis Tool Kit (GATK) pipeline[62], a pipeline developed at Broad Institute integrating multiple tools for next generation sequencing analysis. The GATK pipeline identified 12,085 variants across the ~4Mb region. To narrow down the potential candidates several filters were applied.

SNP – Analysis

Since HCSMA is a co-dominant trait, the variants were filtered accordingly; the dogs with the accelerated disease had to be homozygous for the mutation and the late onset dogs had to be heterozygous (strict filter). A less stringent filter was applied where one animal from the accelerated cohort and the late onset cohort could have a discordant genotype (loose filter). Roughly 10% of the initial variants passed the loose filter, resulting in ~1,200 variants, which were further ranked for conservation score using the SEQscoring module[63]. A conservation score indicates the preservation of alleles across many mammals, since alleles present across multiple species are more likely to be functional. The SEQscoring module ranks for conservation dependent on three databases; UCSC genome browser, Ensemble and 29 mammals. We included any variant scoring high for conservation by any these three databases, resulting in 79 variants 17 of these fulfilled the strict filter of phenotype to genotype concordance.

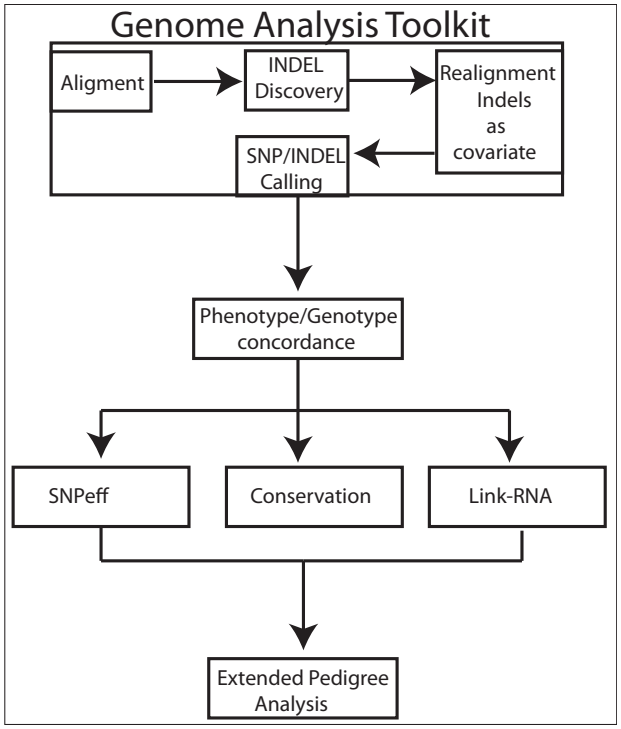


Figure.5 SNP analysis workflow.
GATK analysis of SNPs aligned by BWA aligner,iIndels discovered and genotype used as covariates for realignment procedures, variant discovery and genotyping is performed on the cleaned and realigned bam files. Variants are filtered in regards of phenotype and genotype concordance, Conservation, predicted SNP effect and link RNA intersection. The candidate variants were genotyped in extended pedigree.

In addition we processed all SNPs in a variant prediction database SNPEff[64]. This database rank the variants dependent on several parameters such as coding, synonymous, non-synonymous, being at or near splice site, intronic region, intragenic, and link RNA.

With the assumption that the mutation is novel, any annotated SNPs found in DBSNP[65] database were filtered out. We also intersected the variants with LINK RNA-track and included 2 variants (Broad unpublished data). The workflow of the current analysis is illustrated in in (*fig.5)*.

A total of 81 candidate SNPs were selected across the ~4Mb region (*fig.7),* An extended pedigree analysis was performed by a Sequenom genotype assay of 77 dogs from the same pedigree and additionally 13 dogs from 7 other breeds. The candidate SNPs was tested for concordance with the designated phenotype in the 55 early onset animals and 25 late onset animals. The filter was set so that the additional 13 unrelated dogs had to be homozygous for the reference allele. No mutations were identified concordant with the designated phenotype and genotype.
Therefore, to determine the location of the mutation, a haplotype investigation was carried out.

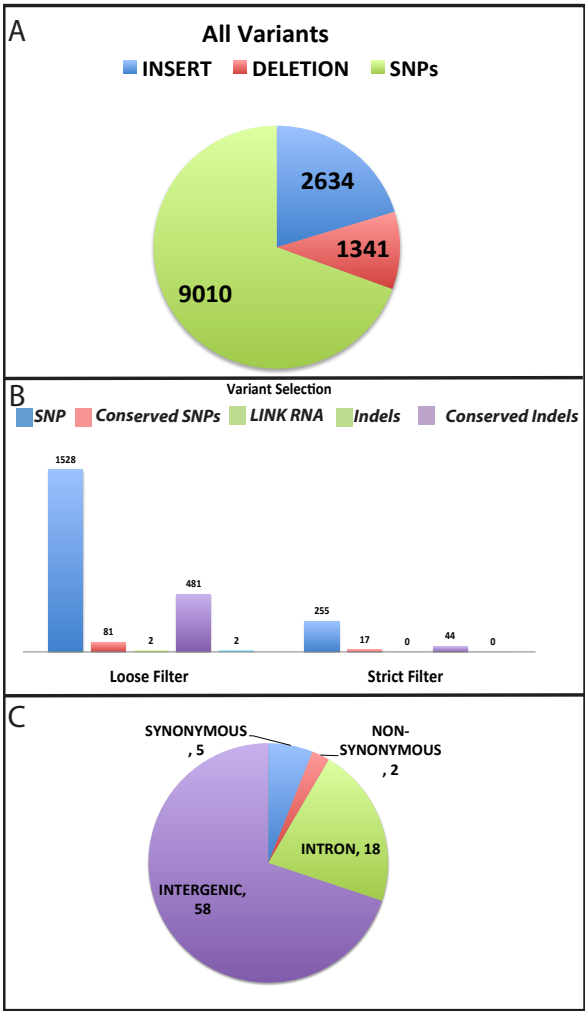


Figure.6 Variant discovery and filtering procedures. **(A)***The distribution of SNPs, insertions and deletions discovered by the Genome Analysis Tool Kit (GATK).* ***(B)*** *The number of SNPs, INDELS, conserved variants and intersected variants located in LINK RNAs regions passing the two filtering procedures.* ***(C)****indicates the candidates SNPs location and coding change.*

Haplotype and Breakpoint analysis
To further narrow down the candidate region, an examination of the haplotype structure was performed, comprising 475 SNPs included in the initial fine mapping study and the additional 81 candidate SNPs. The genotype data were phased and four major haplotypes and several recombinants were identified(*fig.7*). The haplotypes consisted of three healthy haplotypes and one disease haplotype. The disease haplotype was found in all animals and concordance was seen between a co-dominant mode of transmission and the dogs' disease statuses. Within the accelerated cohort the animals were homozygous for the disease haplotype, and within the late onset cohort the animals were carrying one disease haplotype and one healthy haplotype.
The genotype data was easily phased since all breeding pairs of late onset animals had produced offspring with the accelerated disease, who all carry one copy of the disease haplotype without recombination. This haplotype was used as a reference in the phasing procedure of the late onset animals, and any heterozygous non-reference SNP was phased into the healthy haplotype. The recombinant haplotypes present within the accelerated cohort was distinguished by comparison of the phased parents, and identified which healthy haplotypes were recombined. The Brittney spaniels were outcrossed with the beagle

population, which introduced two of the healthy haplotypes. By investigating the recombination breakpoints within the pedigree, a narrower region within which the mutation is located could be identified. The B347 with the accelerated form of the disease shared a large segment with the healthy beagle haplotype. Full homozygosity was not observed until chr13: 50,543,968 thereby defining the left border. The haplotypes introduced by the beagle population deviated from the disease haplotype by 44%-49% of the SNPs across the ~4Mb region. The haplotype introduced earlier in the pedigree was present in B74, B282, B253 and B157 individuals only deviated by ~20% of the SNPs, due to a ~2Mb shared segment between the disease haplotype. In the sequencing data, none of the strictly filtered SNPs observed passed this left breakpoint due to the fact that B282 shared the disease haplotype in this segment. The segment between the left breakpoint defined by B347 and the start point of the shared haplotype between the early onset animals and the late onset animals was ~35kb in size, and spanned the coordinates 50,543,968 bp to 50,578,963 bp on chromosome 13 containing the Steroid 5 Alpha-Reductase 3 (*SRD5A3*) gene.

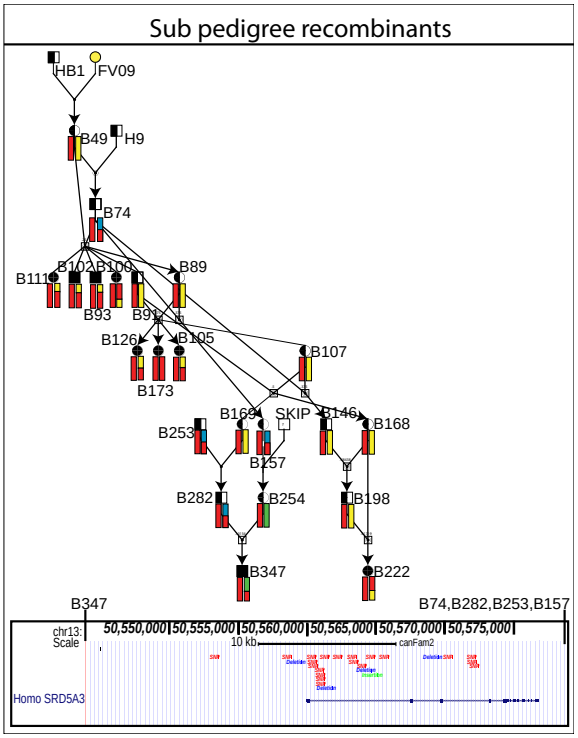


Figure.7 Phased pedigree. *A sub pedigree containing the recombinant offspring. The red haplotype block indicates the disease haplotype, outcrossing with beagle population introduced the yellow and green haplotype, while the (blu/red) haplotype was introduced earlier in the pedigree. The breakpoints are defining the ~35kb region encompassed SRD5A3 containing 21 full concordant SNPs (RED) and 4 Indels (3 deletions (blue) and one insertion (Green).*

Fine mapping to define recombination breakpoints and filling in sequencing gaps

The targeted sequencing had discovered 25 highly concordant variants within the ~35kb region identified by haplotype analysis, and these variants included 21 SNPs, 3 deletions and one insertion. To further investigate these SNPs, a second run of fine mapping was performed comprising the same set of dogs used for the candidate SNPs validation (55 with early onset, 25 with late onset). Highly phenotype-concordant SNPs in the 35kb region and additional SNPs flanking the region were included to define more precise recombination breakpoints. The fine mapping resulted in almost full concordance of the genotypes to phenotypes in all dogs with the exception of the B126 dog and the dogs from the early pedigree, where Mendelian errors had previously been observed. The alleles discovered within the accelerated cohort were however also found in the control dogs.

The highly concordant SNPs were flanking the first exon of SRD5A3, which had no coverage in the sequencing data due to the GC rich promoter region. Since these SNPs were not the causal SNPs, but rather tagging the region of interest. Sequencing of the 1200bp promoter region was carried out, but did not identify the causative mutation.

Expression assay and transcript sequencing

The Indels discovered within this 35kb region were located within simple repeats or SINE elements and were only 1-2 bp long. Sanger sequencing across repeats is challenging, since they are often heterozygous at multiple positions, instead of sequencing across these repeats, an expression assay was designed for *SRD5A3*, and primers were designed across the whole *SRD5A3* transcript to investigate any variants in mRNA splicing.

The tissue included in the analysis was cervical spinal cord from two healthy controls and three accelerated disease dogs. A cross section from the spinal cord was homogenized and levels of *SRD5A3* expression were investigated (*fig.8*).

To be more inclusive, the neighboring genes, *KDR* and *Tmem165* were also included. Since KDR had previously been implicated in ALS[66], and it is possible that an cis-acting element such as an enhancer or silencer may exist within the 35kb region and influence expression levels of the neighboring genes. However, the results show that there is no difference in the expression levels of those genes between the animals with the accelerated form of the disease and the control animals (*fig.8*).

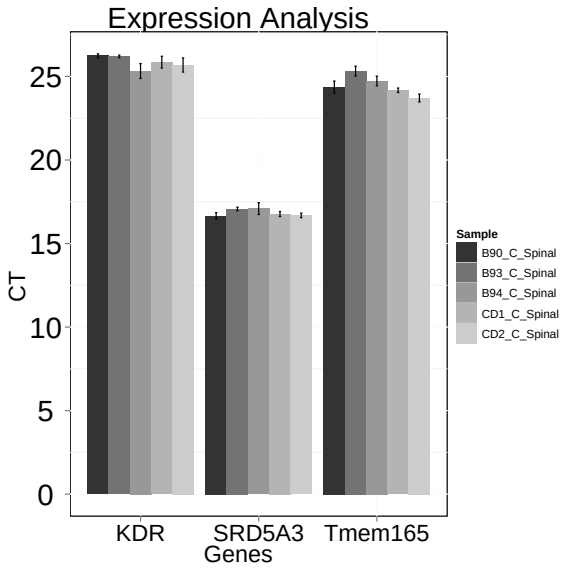


Figure.8 Expression analysis *of KDR, SRD5A3 and Tmem165 performed on cDNA from cervical spinal cord of two healthy controls and three animals with the accelerated form of disease. The y-axis indicates CT-value and x-axis the genes investigated. The error bars denote one standard deviation calculated from three qPCR amplicons. The data were normalized against GAPDH and B-actin genes.*

The transcript sequencing approach of the *SRD5A3* transcript was utilized to see if any alternative splicing might have occurred, which would not have been detected by the expression assay. Sanger sequencing was performed on the whole transcript of SRD5A3 but did not identify any alternative splice variants within the SRD5A3 transcript.

Common haplotype

The earlier founder dogs were excluded from the previous genotype phasing and pedigree analysis, since a trio containing FB4, B1 and B5 exhibited a large number of Mendelian errors and B5 had a high frequency of missing genotypes. The dogs with the accelerated

disease closer to the proband, namely FB3 and FB5, display discordant genotypes to other dogs with accelerated disease in the later generations up to chr13: 50,569,924 after which point share the disease haplotype. Since no mutations concordant with phenotype were discovered by the sequencing in the ~2Mb segment where FB3 and FB5 are concordant with other dogs with accelerated disease, these animals were initially not included in the various analyses. Since mutation was not discovered within the~35kb candidate region defined by B347 and the B74, we had to consider the possibility that the mutation had occurred in the common haplotype, and shared between dogs with the accelerated and late onset cohort. Because the Illumina sequencing data has gaps in the sequences, the mutation could have been missed, there is also a possibility that GATK analysis software, could have failed to identify the causative mutation.

As for haplotype analysis, once FB3 and FB4 were included in the analysis, they were homozygous across the whole ~4Mb segment enabling the phasing of Marks, B3 and Dixie. This phasing identified that the healthy haplotype carried by B74, B253, B282 and B157 is identical to the disease haplotype newly identified in the early pedigree, supporting the hypothesis that the causal mutation must have arisen on a common haplotype.

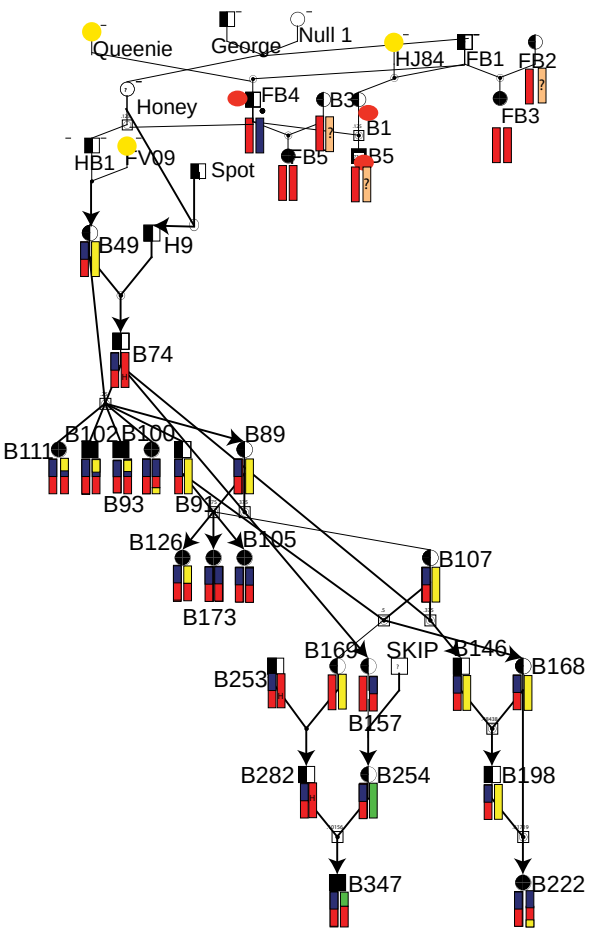


Figure.9 Founder pedigree. *The haplotype denoted with full red fill (blue/red, in the previous pedigree) is the ancestral haplotype carried as the disease haplotype within the early pedigree. Which initially was thought to be a recombined haplotype of the disease haplotype carried by all accelerated dogs, downstream of B49. Since in initial phasing procedure used the disease haplotype carried by animals in later generation as a reference. However when the ancestral haplotype was inherited by HB1, a recombination event occurred producing the haplotype that became common in the later generations in the pedigree (previously denoted with full red, currently signified with red/blue fill). B74, B282, B157 and B253 (red with black lines) carry a healthy haplotype, which is identical to the ancestral disease haplotype across all SNPs across the ~4Mb segment, and was introduced by spot and Honey.*

The disease haplotype previously shown in figure 8 as the "red" haplotype which was found in the majority of the pedigree downstream, is in fact a recombinant haplotype first found in FB4, passed over to HB1 and subsequently to B49 (*fig.10*). We don't have data from H9, but H9 must have carried a healthy ancestral haplotype, identical to the ancestral disease haplotype on all the SNPs across the ~4Mb region, which was passed over to B74.

Taken together, the above haplotype data strongly supports the hypothesis of the causal mutation being located on a common haplotype, and rejects the *SRD5A3* gene as the candidate gene, since the recently identified recombination breakpoint is upstream of *SRD5A3*.Because the mutation is relatively young, and the newly identified haplotype block containing the mutation has not undergone extensive recombination. In addition any recombination event occurring within B282, B74, B253 and B157 passed on to the accelerated disease cohort would go unseen in the due to the shared haplotype. Therefore, the narrowest definition of haplotype block defined by the B74 and B222 is ~2Mb long, located at 50,543,968 to 52,592,547 on chromosome 13. This segment contains 21 genes (*table.1*).

There are several genes expressed within the central nervous system (CNS), such as *REST, TDGF1*, *MSL2* and *HopX*. These genes can be further investigated by Sanger sequencing and by expression approaches to see if the mutation reside within these candidate genes.

Genes	*Expressed*	*Function*
REST	Brain,Lymphocytes	Master regulater of neurogenesis
TDGF1	Brain	Vertebra CNS development,
MSI2	CNS	CNS development
HOPX	Heart,Adult brain,intestine,spleen	Cardiac development
CLOCK	Widely expressed	Circadian rhythmicity
NMU	Brain	Feeding behaviour, energy metabolism
EXOC	Brain,heart,placenta,skeletal muscle	Involved in docking of exocytic vesicles
SRP72	No data	Signal recognition particle,
AASDH	Widely expressed	Acyl-CoA synthetases
CEP135	Widlely expressed	Involved in centrosome
IGFBP7	Widlely expressed	Regulates IGF availability
KIAA1211	No data	Uncharacterized
NOA1	No data	Nitric oxide-associated protein1
PAICS	No data	Purine biosynthesis
POLR2A	Widely expressed	Subunit in RNA polymerase-II
PPAT	No data	Purine biosynthesis
Tmem165	No data	Glycosylation
SFR1	No data	Recombination repair
NUSAP1	No data	Spindle Microtubule Organization
SPINKS2	Sperm	Kazal-type serine protease inhibitor
PDCL2	Testis	Germline specific gene

Table.1 genes located within the shared haplotype. *Column 1 indicates the genes, column 2 in what tissue the genes have been reported being expressed using OMIM database and column 3 the potential function of the genes. Several of the genes are expressed in the central nerve system such as REST, TDGF1 MSI2, HOPX (Unregulated within human ALS patients).*

Insertions and deletions

The GATK pipeline had not discovered any SNPs or indels that are concordant to phenotype within the ~2Mb segment. Therefore, to extend the analysis to detect larger structural rearrangements, we utilized the Pindel software[67], which investigates the orientation of the pairs, and clusters with an increased or decreased insert size. Pindel can detect multiple types of structural variants such as tandem repeats, inversions, and

deletions. We also performed a coverage analysis using CNV-seq and an in house script used for the detection of a duplication event found in Shar-Pei[68].

Several filters were applied to the Pindel analysis, as the majority of the calls were called at/around gaps and repetitive elements. The filter was set so that the calls had to be supported by at least three reads and found in at least six out of eight of the animals to be considered for further analysis. Further more, a repeat masker file was used to filter out any call in proximity to repeats, since these often contribute to artifacts showing extended length of structural variants.

Since Pindel does not detect heterozygous genotypes, the calls had to be manually inspected in integrated genome viewer (IGV)[69] after the filtration process. The called variants were annotated for coding regions by intersecting the variants with Refseq gene track of UCSC-genome browser[70]. Since its hard to define the exact borders of the structural variants, the reads were extracted from the region containing the indel called by Pindel, and de-novo assembled using the Abyss software to determine the exact borders. These analyses detected two large deletions within the *REST* gene of roughly ~6.5kb each.

Initially the deletion appeared to encompass the whole introns and part of the exonic regions, however upon *de-novo* assembly, the accurately defined borders showed the deletion to be only in the intronic regions. This type of signature implies a possible insertion of a pseudo gene, as a pseudo gene for *REST* was not annotated within the reference genome.

To detect the presence of *REST* pseudogene and test for association with the phenotype, the primers were designed within the exons and run on the genomic DNA from the 15 dogs (5 accelerated/5 late onset/ controls) used for validation of the small Indels. The assay confirmed the insertion of a potential REST pseudo gene, since the DNA amplified and produced a product reflecting the size of processed mRNA for REST. The pseudogene was discovered in all dogs thus not associated with the phenotype.

Discussion

At present, the causal mutation of HCSMA remains to be found, even though the characterization and mapping of HCSMA has been ongoing since the 1970s. However the recent technology advances, including genome wide SNP arrays and next generation sequencing, have identified and narrowed down a candidate region on Chromosome 13. The candidate region of 10Mb, initially discovered by GWAS, was further narrowed down to ~4Mb by recombination breakpoints analysis (*Fig.3*). Additional fine mapping assays and haplotype analysis identified a segment of ~2Mb (*Fig.9*).

The targeted sequencing has not been successful in identifying the causal mutation of HCSMA. Even though candidate mutations were found, none of them exhibited full concordance with HCSMA disease phenotypes, nor exclusive to the HCSMA pedigree.

Given the severe phenotype, it is not unreasonable to expect a coding mutation, yet none was discovered after applying the strict filter of 100% concordance between phenotype and genotype in the sequence data. Because the *KDR* gene has expressed similar lesions as the HCSMA dogs in a strain of knockout mouse, it was investigated in details despite the breakpoint in B347 excluded the *KDR* gene. There was no coding mutation within this gene, and since no expression difference were observed between cases and controls, the *KDR* gene was excluded from the analysis

The *SRD5A3* gene was extensively studied, since the initial haplotype analysis indicated the mutation to be located within the 35kB region containing *SRD5A3*. The characterization of *SRD5A3* did not reveal any evidence to link it to HCSMA (*Fig.8*). No expression differences between dogs with accelerated disease and healthy dogs were observed, and no phenotype-concordant mutations were discovered including within the first exon which was filled in by Sanger sequencing as it was initially missing. In addition, sequencing of the *SRD5A3* transcript did not find any abnormal splice variants. Since the mutation did not reside within the ~35kB region identified by the initial haplotype analysis, we hypothesized the mutation to be located on the ~2Mb ancestral haplotype identified by an additional haplotype analysis including dogs in earlier pedigree, which is present in the accelerated disease cohort but also shared in the four of late onset animals (*Fig.9*).

The structural variant pipeline discovered a deletion within the REST gene located within the ~2Mb region. Because deletion was observed only in the introns and also accounted only for a small proportion of all reads, we hypothesized that the deletion was indicative of a processed pseudogene. This hypothesis was validated by a PCR-assay, however, it was present in both accelerated and late onset cases, and healthy controls. In addition three deletions were found within 3' UTR of the REST gene, however, they were not strictly concordant with the disease phenotype.

The two-step mapping approach where additional breeds with the same phenotype normally accelerates mutation discovery within the dog population, is in principle advantageous. However, the two-stage strategy cannot be applied to this study, as the mutation is private to Brittney spaniel population and not shared across multiple breeds. Furthermore the mutation is relatively new, and the haplotype block has not experienced extensive recombination.

No phenotype-concordant mutations have been found in the ~2Mb segment, and it has to be inspected in regards of gaps present within the first exons. The first exons are problematic within the dog population, since dogs have increased GC content in promoter regions, which induces problems in PCR and sequencing procedures. Initially the genes expressed in central nervous system such as *REST, TDGF1, MIS2* and *HOPX* should be investigated (*Table.1*).

Next generation sequencing has revolutionized the field of genetics and greatly improved variant discovery. SNPs can easily be identified, and smaller Indels can be found with decent accuracy. The problems with next generation sequencing is the identification of

larger structural rearrangements, normally defined as >1kb. There are several types of rearrangements including; deletions, insertions, tandem duplication, interspersed duplication, inversion and translocation. Four approaches are currently used for identification of structural rearrangements in next generation sequencing, including analysis of the read-pairs, coverage depth, split reads and *de novo* assembly.

We employed several Structural Variants (SV) algorithms trying to identify the causative mutation of HCMSA. The Pindel[67] algorithm was used for analysis of paired-end data and split-reads. The Pindel algorithm identifies regions with increased insert size signifying a deletion event. Some fragments will only have one mapped read, which can be used as an anchoring point, and the second read remains unmapped. These unmapped reads are likely to contain information of the exact location, since the breakpoint is likely to have occurred within the unmapped read. Pindel uses a "pattern of growth" approach to detect substrings within the read, and try to map these identified substrings of the read to a restricted region containing 2-3 times of the normal insert size from the anchoring point (the mapped read), and map split-read and exact breakpoints. Pindel also detects all the other structural rearrangements, however, with varied accuracy. The sequence coverage was analyzed by CNV-seq[71] and the in house algorithm developed by Dr. Evan Mauceli, which was used for the identification of a duplication event in Shar-pei[68]. Both algorithms were unable to find any differences in coverage between the accelerated cohort and the late onset cohort. De novo assembly was performed using the Abyss[72] but were only used for validation of REST gene and was not performed across the whole ~4Mb.

Several factors complicate the identification of larger structural variants such as repetitive regions, which are hard to align and assemble since they can map to multiple locations in the reference genome. Therefore, de novo insertions of repetitive elements remain hard to be identified, and might also be filtered out during the aligning procedures.
A good example of this problem is illustrated in the exome sequencing of induced pluripotent stem cells, where a homozygous insertion of an ALU element was accidently observed within an exon, and not called within sequencing analysis[73].

The causal mutation for HCSMA might be a larger structural variant missed by the SV-analysis. Since these are complicated to be characterized by the current sequence technology and available software. Currently, a *de novo* assembly across the ~2Mb region is on going, which might reveal new information. Also, filling in the gaps and assessing expression of the genes found within the ~2Mb region should be performed. If these attempts all fail to discover the mutation, then a new sequencing approach may be feasible, including whole genome sequencing with a mate pair library for detecting larger structural variants.

Acknowledgement

I would like to thank professor Kerstin Lindblad-Toh for helping me through this project, and giving me this great opportunity and experience to spend time in such an inspiring and knowledgeable environment as the Broad Institute.

I'm grateful for the support from my immediate supervisor Dr. Noriko Tonomura who has performed the GWAS study, Finemapping analysis and laboratory work regarding the targeted sequencing. I also want to thank my opponent Carl-Johan Rubin, by providing a great discussion and giving good feedback on my presentation.
I would also like to acknowledge the people involved in Kerstin's group at Broad Institute who has helped with elaborating thoughts on the project; Dr. Evan Mauceli, Dr. Andrew Lundquist, Ross Swofford, Michele Perloski, Dr. Ruqi Tang and Hyun Ji Noh.

Materials and methods

Experimental cohort
The animals included in the Fine mapping, GWAS analysis and targeted sequencing provided in the *Figure.2*

Next generation sequencing data analysis
The variants were processed by the GATK pipeline, including alignment by the BWA aligner[74] to the CanFam2 May 2005 reference genome. The candidate SNPs were ranked for conservation by seq scoring module[63] and the effect were predicted by SNPEff[64] using the CanFam 2.61 database. Removal of previously annotated SNPs using DB-snp database[65], SNPs intersected with the LINK RNA track. Larger structural variants were processed with Pindel software according to default settings[67]. Artifacts generated by the Structural variants pipeline were clean in proximity of repeats using a repeat masker from CanFam-2.61 using the BED tools suite[75]. The candidates were intersected with the non-reference gene track at UCSC genome browser[70]. The candidates were manually inspected using the Integrative Genomics (IGV) browser[69]. The Indels were *De novo* assembled by the Abyss-pe assembler to identify the borders[72]

Finemapping
Candidate genotyping and fine mapping SNPs were genotyped using the Sequenom MassArray technology, and the data were processed in PLINK analysis pipeline.

Primer design
Primers were designed with the primer3[76], transcript sequencing and primer walks were designed implementing the primer3 algorithm in a Perl script, which stepwise walking across the segment of interest.

PCR- Protocol
The 5-10 ng of genomic DNA was mixed with 1 U AccuPrime-Taq HiFi (Invitrogen) 2.5 uL of 10x AccuPrimer Buffer II and corresponding primers at a final concentration of 100-300nM and diluted with Gibco RNase free H20 to a final volume of 25 uL. In GC rich templates the reaction mixture were supplemented with DMSO at a final concentration of 10%.
The reaction mixture were amplified in a Master Cycler Pro (Eppendorf) with the following thermocycling protocol 95C 5 min, 95C 30s – ramp 62C-52C 1C increment 68C 1min for 10 cycles, followed by 95C 30s, 54C 30s, 68C 1min for 30 cycles. Gel electrophoresis was carried out on a 1.5% agaros gel for 30min on 110V (PowerPac,

BioRAD). The PCR-products were purified according to manufactures specifications (Mini Elute PCR Purification kit, QIAGEN)

Sanger sequencing
First exon sequencing were bidirectional sequenced using dye termination chemistry on a ABI 3730 sequencer.

Sanger Data analysis
The trace files were analyzed using a pipeline consisting of polyphred[77] for SNP calling and CHILD algorithm[78] for detection of heterozygous insertions and deletions. The variants were manually inspected using CodonCode aligner (V3.7.1.2).

RNA extraction
Tissues included in analysis were cross sections of cervical, lumbar spinal cord and also liver from 3 cases and 5 controls. The tissues were homogenized using (Rotor Stator Generator, OMNI International) on ice in 1 mL TRizol ((TriRegent, Molecular Research Center), incubated at RT for 5 min followed by adding 100uL BCP (BCP,Molecular Research Center). Samples were vortexed for 10s and incubated for 8min at RT, centrifuged at 12000g for 15min at 4C, the aqueous phase was collected and mixed with 0.5mL isopropyl and incubated for 15 min at RT. The pellet was cleaned in 1mL 75% EtOH and centrifuged for 10 min at 12000g in 4C, the pellet was dissolved in 100 uL RNase free H20 (Gibco,Life technologies).

DNAse Treatment
The tissues were mixed with 11uL of 10x Turbo DNase buffer 1ul Turbo DNAse and incubated for 30 min at 37C, an additional 1uL of Turbo DNase was added followed by 30 min incubation at 37C. The samples were incubated with 11uL DNase inactivation reagent for 5 min at RT followed by centrifugation at 10000g for 1.5 min. The supernatant was collected and the RNA was precipitated according to previously described protocol. The concentration was measured and by nanodrop 8000 (Thermo Scientific) and the quality of RNA was assessed by bioanalyzer (2100 Bioanalyzer , Agilent technologies).

cDNA Synthesis
The cDNA synthesis was carried out according to manufactures specification (Superscript II RT, Invitrogen) using 250ng random primers and a RNA concentration ranging from 0.5-3ug total RNA. Subsequently transformed into double stranded cDNA using second strand synthesis protocol provided by manufacture (Invitrogen).

Quantitative-PCR
The cDNA was mixed 2x QuantiFast- Syber Green PCR (QIAGEN) and corresponding primers (*Supplementary data*) at a final concentration of 100nM in a volume of 10uL. Each sample was run in three replicates and quantified by qPCR (Light Cycler 480 II,Roche) with the following thermocycling protocol 1 cycle 2min 95C 40cycles 95C 15s,60C 1min.The qPCR data was analyzed in R- 2.13.1 using both GAPDH and beta-Actin for normalization of the data, ggplot2 package were used for graphical illustrations.

References

1. Watson, J.D. & Crick, F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).
2. Botstein, D., White, A.M., Skolnich, M. & David, R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphism. *Am J Hum Genet* **32(3)**, 314-331 (1980).
3. Jefferson, A., WilSon, V. & Lay Thein, S. Hypervariable 'minisatellite' regions in human DNA. *Nature* **314**, 67 - 73 (1985).
4. Litt, M. & Luty, J. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am J Hum Genet* **44**, 397-401 (1989).
5. Gusella, J.F.*, et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234-238 (1983).
6. Riordan, J.R.*, et al.* Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**, 1066-1073 (1989).
7. Klein, R.J.*, et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385-389 (2005).
8. Maragonore, D.*, et al.* High-Resolution Whole-GenomeAssociation Study of Parkinson Disease. *Am J Hum Genet* **77**, 685-693 (2005).
9. Souied, E.H.*, et al.* Y402H complement factor H polymorphism associated with exudative age-related macular degeneration in the French population. *Mol Vis* **11**, 1135-1140 (2005).
10. Hardy, J. & Singleton, A. Genomewide association studies and human disease. *N Engl J Med* **360**, 1759-1768 (2009).
11. Manolio, T.A. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* **363**, 166-176 (2010).
12. Ott, J., Kamatani, Y. & Lathrop, M. Family-based designs for genome-wide association studies. *Nature reviews. Genetics* **12**, 465-474 (2011).
13. Hao, K., Chudin, E., Greenawalt, D. & Schadt, E.E. Magnitude of stratification in human populations and impacts on genome wide association studies. *PLoS One* **5**, e8695 (2010).
14. Kong, A.*, et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868-874 (2009).
15. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-5467 (1977).
16. Shizuya, H.*, et al.* Cloning and stable maintenance of 300-kilobase-pair fragment of human DNA in Escherichia coli using an F-factor-based vector. *Proc Natl Acad Sci U S A* **89**, 8794-8797 (1992).
17. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945 (2004).
18. News, N. International Consortium Completes Human Genome Project. Vol. 2012 (2003).
19. Metzker, M.L. Sequencing technologies - the next generation. *Nature reviews. Genetics* **11**, 31-46 (2010).

20. Mamanova, L.*, et al.* Target-enrichment strategies for next-generation sequencing. *Nat Methods* **7**, 111-118 (2010).
21. Ng, S.B.*, et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**, 30-35 (2010).
22. Ostrander, E.A. & Kruglyak, L. Unleashing the canine genome. *Genome Res* **10**, 1271-1274 (2000).
23. Club, A.K. *The Complete Dog Book*, (Howell Book House, New York, 1998).
24. Galibert, F. & Andre, C. The dog genome. *Genome Dyn* **2**, 46-59 (2006).
25. Lindblad-Toh, K.*, et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-819 (2005).
26. Peltonen, L., Palotie, A. & Lange, K. Use of population isolates for mapping complex traits. *Nat Rev Genet* **1**, 182-190 (2000).
27. 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).
28. Lindblad-Toh, K.*, et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-819 (2005).
29. Grafstein, B. & Forman, D.S. Intracellular transport in neurons. *Physiol Rev* **60**, 1167-1283 (1980).
30. Droz, B. & Leblond, C.P. Migration of proteins along the axons of the sciatic nerve. *Science* **137**, 1047-1048 (1962).
31. Simon, S.M. & Llinas, R.R. Compartmentalization of the submembrane calcium activity during calcium influx and its significance in transmitter release. *Biophys J* **48**, 485-498 (1985).
32. Hodgkin, A.L. & Huxley, A.F. Action Potentials Recorded from Inside a Nerve Fibre. *Nature* **144**, 710-711 (1939).
33. Lefebvre, S.*, et al.* Identification and characterization of a spinal muscular atrophy-determining gene. *Cell* **80**, 155-165 (1995).
34. Feldkotter, M., Schwarzer, V., Wirth, R., Wienker, T.F. & Wirth, B. Quantitative analyses of SMN1 and SMN2 based on real-time lightCycler PCR: fast and highly reliable carrier testing and prediction of severity of spinal muscular atrophy. *Am J Hum Genet* **70**, 358-368 (2002).
35. Lorson, C.L., Hahnen, E., Androphy, E.J. & Wirth, B. A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. *Proc Natl Acad Sci U S A* **96**, 6307-6311 (1999).
36. Baughan, T.*, et al.* Stimulating full-length SMN2 expression by delivering bifunctional RNAs via a viral vector. *Mol Ther* **14**, 54-62 (2006).
37. Wood-Allum, C. & Shaw, P.J. Motor neurone disease: a practical update on diagnosis and management. *Clin Med* **10**, 252-258 (2010).
38. Bruijn, L.I., Miller, T.M. & Cleveland, D.W. Unraveling the mechanisms involved in motor neuron degeneration in ALS. *Annu Rev Neurosci* **27**, 723-749 (2004).
39. Shaw, P.J., Ince, P.G., Falkous, G. & Mantle, D. Oxidative damage to protein in sporadic motor neuron disease spinal cord. *Ann Neurol* **38**, 691-695 (1995).
40. Shibata, N.*, et al.* Morphological evidence for lipid peroxidation and protein glycoxidation in spinal cords from sporadic amyotrophic lateral sclerosis patients. *Brain Res* **917**, 97-104 (2001).

41. Sathasivam, S., Grierson, A.J. & Shaw, P.J. Characterization of the caspase cascade in a cell culture model of SOD1-related familial amyotrophic lateral sclerosis: expression, activation and therapeutic effects of inhibition. *Neuropathol Appl Neurobiol* **31**, 467-485 (2005).
42. Sobue, G.*, et al.* Phosphorylated high molecular weight neurofilament protein in lower motor neurons in amyotrophic lateral sclerosis and other neurodegenerative diseases involving ventral horn cells. *Acta Neuropathol* **79**, 402-408 (1990).
43. De Vos, K.J.*, et al.* Familial amyotrophic lateral sclerosis-linked SOD1 mutants perturb fast axonal transport to reduce axonal mitochondria content. *Hum Mol Genet* **16**, 2720-2728 (2007).
44. Ackerley, S.*, et al.* Glutamate slows axonal transport of neurofilaments in transfected neurons. *J Cell Biol* **150**, 165-176 (2000).
45. Brownlees, J.*, et al.* Phosphorylation of neurofilament heavy chain side-arms by stress activated protein kinase-1b/Jun N-terminal kinase-3. *J Cell Sci* **113 (Pt 3)**, 401-407 (2000).
46. Kirby, J.*, et al.* Mutant SOD1 alters the motor neuronal transcriptome: implications for familial ALS. *Brain* **128**, 1686-1706 (2005).
47. Lai , C.*, et al.* Amyotrophic lateral sclerosis 2-deficiency leads to neuronal degeneration in amyotrophic lateral sclerosis through altered AMPA receptor trafficking. *Journal of Neuroscience* **26(45)**, 11798-11806 (2006).
48. Beers, D.R., Henkel, J.S., Zhao, W., Wang, J. & Appel, S.H. CD4+ T cells support glial neuroprotection, slow disease progression, and modify glial morphology in an animal model of inherited ALS. *Proc Natl Acad Sci U S A* **105**, 15558-15563 (2008).
49. Atkin, J.D.*, et al.* Induction of the unfolded protein response in familial amyotrophic lateral sclerosis and association of protein-disulfide isomerase with superoxide dismutase 1. *J Biol Chem* **281**, 30152-30165 (2006).
50. Lorenz, M.D., Cork, L.C., Griffin, J.W., Adams, R.J. & Price, D.L. Hereditary spinal muscular atrophy in Brittany Spaniels: clinical manifestations. *J Am Vet Med Assoc* **175**, 833-839 (1979).
51. Cork, L.C., Griffin, J.W., Choy, C., Padula, C.A. & Price, D.L. Pathology of motor neurons in accelerated hereditary canine spinal muscular atrophy. *Lab Invest* **46**, 89-99 (1982).
52. Sack, G.H., Jr., Cork, L.C., Morris, J.M., Griffin, J.W. & Price, D.L. Autosomal dominant inheritance of hereditary canine spinal muscular atrophy. *Ann Neurol* **15**, 369-373 (1984).
53. Cork, L.C.*, et al.* Changes in neuronal size and neurotransmitter marker in hereditary canine spinal muscular atrophy. *Lab Invest* **61**, 69-76 (1989).
54. Cork, L.C., Griffin, J.W., Munnell, J.F., Lorenz, M.D. & Adams, R.J. Hereditary canine spinal muscular atrophy. *J Neuropathol Exp Neurol* **38**, 209-221 (1979).
55. Green, S.L., Bouley, D.M., Pinter, M.J., Cork, L.C. & Vatassery, G.T. Canine motor neuron disease: clinicopathologic features and selected indicators of oxidative stress. *J Vet Intern Med* **15**, 112-119 (2001).

56. Pinter, M.J., Waldeck, R.F., Cope, T.C. & Cork, L.C. Effects of 4-aminopyridine on muscle and motor unit force in canine motor neuron disease. *J Neurosci* **17**, 4500-4507 (1997).
57. Rich, M.M.*, et al.* Reduced endplate currents underlie motor unit dysfunction in canine motor neuron disease. *J Neurophysiol* **88**, 3293-3304 (2002).
58. Carrasco, D.I., Rich, M.M., Wang, Q., Cope, T.C. & Pinter, M.J. Activity-driven synaptic and axonal degeneration in canine motor neuron disease. *J Neurophysiol* **92**, 1175-1181 (2004).
59. Blazej, R.G., Mellersh, C.S., Cork, L.C. & Ostrander, E.A. Hereditary canine spinal muscular atrophy is phenotypically similar but molecularly distinct from human spinal muscular atrophy. *J Hered* **89**, 531-537 (1998).
60. Green, S.L.*, et al.* Structure, chromosomal location, and analysis of the canine Cu/Zn superoxide dismutase (SOD1) gene. *J Hered* **93**, 119-124 (2002).
61. Purcell, S.*, et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575 (2007).
62. McKenna, A.*, et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
63. Katarina Truvé, O.E., Martin Norling, Maria Wilbe, Evan Mauceli, Kerstin Lindblad-Toh, Erik Bongcam-Rudloff. SEQscoring: a tool to facilitate the interpretation of data generated with next generation sequencing technologies. *EMBnet journal* **17**, 38-45 (2011).
64. De Baets, G.*, et al.* SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res* (2011).
65. Sherry, S.T.*, et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-311 (2001).
66. Oosthuyse, B.*, et al.* Deletion of the hypoxia-response element in the vascular endothelial growth factor promoter causes motor neuron degeneration. *Nat Genet* **28**, 131-138 (2001).
67. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z.M. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871 (2009).
68. Olsson, M.*, et al.* A Novel Unstable Duplication Upstream of HAS2 Predisposes to a Breed-Defining Skin Phenotype and a Periodic Fever Syndrome in Chinese Shar-Pei Dogs. *Plos Genet* **7**(2011).
69. Thorvaldsdottir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* (2012).
70. Kent, W.J.*, et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).
71. Xie, C. & Tammi, M.T. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *Bmc Bioinformatics* **10**(2009).
72. Simpson, J.T.*, et al.* ABySS: A parallel assembler for short read sequence data. *Genome Res* **19**, 1117-1123 (2009).

73. Tucker, B.A.*, et al.* Exome sequencing and analysis of induced pluripotent stem cells identify the cilia-related gene male germ cell-associated kinase (MAK) as a cause of retinitis pigmentosa. *Proc Natl Acad Sci U S A* **108**, E569-E576 (2011).
74. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
75. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
76. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**, 365-386 (2000).
77. Nickerson, D.A., Tobe, V.O. & Taylor, S.L. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res* **25**, 2745-2751 (1997).
78. Zhidkov, I., Cohen, R., Geifman, N., Mishmar, D. & Rubin, E. CHILD: a new tool for detecting low-abundance insertions and deletions in standard sequence traces. *Nucleic Acids Res* **39**, e47 (2011).