Advances in Genetic Mapping and Sequencing Techniques:

A Demonstration using the Domestic Dog Model

Oliver Forman BSc

Submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

The University of Glasgow

School of Veterinary Medicine

September 2013

Abstract

Over the past ten years huge advances have been made in the field of genetics and genomics. Genetic mapping has evolved from laborious linkage and homozygosity based approaches to high-throughput genome-wide association studies using whole genome SNP array technology. Through massively parallel sequencing technology, gigabases of sequencing data can now be produced in a single experiment. The domestic dog has been increasingly recognised as a model for human disease and mapping of inherited disease in the domestic dog is facilitated by fixed and genetically isolated populations.

The aims of this thesis were to demonstrate advances in mapping and sequencing techniques by investigating the genetics of five inherited disorders, representing significant welfare issues in the purebred dog. An additional aim was to develop diagnostic DNA tests to identify affected individuals and asymptomatic carriers.

A parallel mapping approach was used to map two autosomal recessive conditions in the Cavalier King Charles Spaniel. The use of a single common set of controls for two independent genome-wide association studies was demonstrated as an efficient mapping strategy when studying two conditions affecting a single breed. Newly available target enrichment and massively parallel sequencing methodology was used to simultaneously sequence both disease-associated loci, with one condition acting as a control for the other.

A genome-wide homozygosity mapping approach using microsatellite markers was used to investigate spinocerebellar ataxia in the Italian Spinone. The disorder was successfully mapped to a single chromosome using six cases and six controls, and fine mapped with additional microsatellite markers. Subsequently, a progression of sequencing techniques were used to identify the disease-associated mutation, with the study highlighting the potential difficulties of using massively parallel sequencing technologies.

Spinocerebellar ataxia (or late onset ataxia) in the Parson Russell Terrier was investigated using a genome-wide association study followed by a target enriched massively parallel sequencing approach. Further sequencing was performed to reduce the large number of potential causal variants, with the entire workflow achieved in-house.

The final experimental chapter describes the use of a genome-wide mRNA sequencing (mRNA-seq) approach as a method of candidate gene sequencing of a single case of neonatal cerebellar cortical degeneration in a Beagle dog. The mRNA-seq approach demonstrates a simple, fast and cost effective method of targeted resequencing of expressed genes when a suitable tissue resource is available.

For all five disorders under investigation, disease-associated mutations were identified leading to the development of diagnostic tests. Three of the mutations were in genes not previously associated with similar conditions in humans or other model organisms.

Contents

List of Tables

List of Figures

List of Appendices

Acknowledgments

I would like to thank the Animal Health Trust for giving me the opportunity to study towards my PhD and all the amazing people that work there. In particular Louise Pettitt for being a great friend and for making work such good fun. I would like to thank Dr Mike Boursnell for introducing me to Linux and Perl, and being enormously patient and helpful, and Graham Newland for IT support.

I would like to thank my University of Glasgow supervisor Professor Jacques Penderis for helping to initiate my PhD and for help and support throughout my studies. I am particularly thankful to my supervisor at the Animal Health Trust, Dr Cathryn Mellersh, for her leadership and being incredibly patient and supportive throughout.

I would like to thank my parents and family for their love and support and I am especially thankful to my wife Roxanna for being loving and supportive, believing in me and always being there for me. I'm proud of everything we have achieved together.

Publications

Forman, O. P., Penderis, J., Hartley, C., Hayward, L. J., Ricketts, S. L. & Mellersh, C. S. (2012) Parallel mapping and simultaneous sequencing reveals deletions in *BCAN* and *FAM83H* associated with discrete inherited disorders in a domestic dog breed. *PLoS Genetics*, 8, e1002462.

Forman, O. P., De Risio, L., Stewart, J., Mellersh, C. S. & Beltran, E. (2012) Genome-wide mRNA sequencing of a single canine cerebellar cortical degeneration case leads to the identification of a disease associated *SPTBN2* mutation. *BMC Genetics*, 13, 55.

Forman, O. P., De Risio, L., & Mellersh, C. S. (2013) Missense mutation in *CAPN1* is associated with spinocerebellar ataxia in the Parson Russell Terrier dog breed. *PLoS One,* 8, e64627.

Forman, O. P., De Risio, L., Matiasek, K., Platt, S. R. & Mellersh, C. S. (2013) Spinocerebellar ataxia in the Italian Spinone dog breed is associated with an intronic GAA repeat expansion in *ITPR1*. *Acta Neuropathologica* (*In preparation*).

Declaration

I declare that this thesis is my own original work, and has not been submitted for an award at any other university. Contributions from service providers and individuals are acknowledged in the text.

Oliver Forman
9th September 2013

Abbreviations

ADHAI	Autosomal-dominant hypocalcification *amelogenesis imperfecta*
AHT	Animal Health Trust
BAC	Bacterial artificial chromosome
BAEPs	Brainstem auditory-evoked potentials
BAER	Brainstem auditory-evoked responses
BAM	Binary sequence alignment/map
BCAN	Brevican
BED	Browser extensible data
BLAST	Basic local alignment search tool
BLOSUM	Blocks substitution matrix
BWA	Burrows-Wheeler aligner
C2L	C2 like domain
CAPN1	Calcium dependent cysteine protease, calpain1
ChIP-seq	Chromatin immunoprecipitation sequencing
CKCS	Cavalier King Charles Spaniel
CKCSID	Congenital keratoconjunctivitis sicca and ichthyosiform dermatosis
CNV	Copy number variation
CSKDD	Committee for the standardisation of the karyotype in the domestic dog
Ct	Threshold cycle
DENND4B	DENN/MADD domain containing 4B
DMD	Dystrophin gene
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
EB	Empty basket
EDTA	Ethylenediaminetetraacetic acid
EF	Episodic falling
ENCODE	Encyclopedia of DNA elements
ESTs	Expressed sequence tags
FFPE	Formalin fixed paraffin embedded
FISH	Fluorescent *in situ* hybridisation
FMM	Fast mixed model
FXN	Frataxin
GA	Genome Analyser
GABA	Gamma-aminobutyric acid
GATK	Genome Analysis ToolKit
GBS	Genotyping-by-sequencing
GENO	SNP genotyping frequency
GSP	Gene specific primer
GWAS	Genome-wide association study

HCRTR2	Hypocretin (orexin) receptor 2
IBS	Identity-by-state
IGV	Integrative Genomics Viewer
Indel	Insertion deletion polymorphism
IP3	Inositol triphosphate
IPTG	Isopropyl β-D-1-thiogalactopyranoside
ITPR1	Inositol 1,4,5-trisphosphate receptor, type 1
JRT	Jack Russell Terrier
KID	Keratitis-ichthyosis-deafness syndrome
LB	Lysogeny broth
LD	Linkage disequilibrium
LOA	Late onset ataxia
LOD	Logarithm of odds
MAF	Minor allele frequency
MDS	Multidimensional scaling
MIND	Percentage SNP missingness per individual
MPSS	Massively parallel signature sequencing
mRNA-seq	Genome-wide mRNA sequencing
NCCD	Neonatal cerebellar cortical degeneration
NGS	Next generation sequencing
NHGRI	National Human Genome Research Institute
OMIA	Online Mendelian Inheritance in Animals
OMIM	Online Mendelian Inheritance in Man
PBS	Phosphate-buffered saline
PBS-T	Phosphate-buffered saline/0.1% Tween 20
PCD	Protease core domain
PC	Purkinje cells
PCR	Polymerase chain reaction
PEF	Penta EF-hand calcium binding domain
PGM	Personal Genome Machine
Polyphen	Polymorphism Phenotyping
PPi	Pyrophosphate group
PRCD	Progressive rod-cone degeneration
PRT	Parson Russell Terrier
PTP	PicoTitre plate
QPCR	Quantitative polymerase chain reaction
QRT-PCR	Quantitative reverse transcription polymerase chain reaction
QQ	Quantile-quantile
RACE	Rapid amplification of cDNA ends
RH-mapping	Radiation hybrid mapping
RIN	RNA integrity number
RNA	Ribonucleic acid

RPE65	Retinal pigment epithelium-specific 65 kDa protein
RT-PCR	Reverse transcription PCR
SAM	Sequence Alignment/Map
SBS	Sequencing by synthesis
SCA	Spinocerebellar ataxia
SCAR	Spinocerebellar ataxia recessive
SDS	Sodium dodecyl sulphate
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SHFT	Smooth-Haired Fox Terrier
SIFT	Sorting Intolerant from Tolerant
SINEs	Short interspersed nuclear elements
SNP	Single nucleotide polymorphism
SOC medium	Super optimal broth with catabolite repression medium
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
SPTBN2	Beta-III spectrin
TAE	Tris-acetate-EDTA
TE	Tris-EDTA
UTR	Untranslated region
w/t	Wild-type

Chapter

1. Introduction

1.1. The domestic dog

The domestic dog, *Canis lupus familiaris*, has ancient origins and descended from the grey wolf around 33 thousand years ago based on the most recent archaeological evidence (Ovodov et al., 2011). The location of domestication is highly debated. Mitochondrial and Y chromosome evidence suggests an East Asian origin (Ding et al., 2011, Pang et al., 2009), whereas genome-wide single nucleotide polymorphism (SNP) analysis and archaeological findings imply a European and Middle Eastern origin (Vonholdt et al., 2010), although it is agreed that multiple origins are likely, with interbreeding between early domestic dog and wolf populations occurring through a period of domestication.

Humans share a unique relationship with the domestic dog. From the point of domestication, the dog has been bred by humans for companionship, novelty and numerous working attributes such as herding, hunting and retrieving. Dog breeding as a hobby first became fashionable in the middle of the 19th Century. The rapid rise in the popularity of breeding and showing dogs led to the formation of the Kennel Club on the 4th April 1873 as a regulatory governing body. Today hundreds of dog breeds have been established worldwide, with the British Kennel Club alone recognising 210 breeds and registering over 200,000 purebred dogs each year (The Kennel Club 2012).

1.2. Disease in the purebred dog

Collectively, purebred dogs suffer from a wide range of diseases. The development of pure breeds produces closed populations with fixed and limited gene pools. Individual breeds have closed stud books and dogs can only be registered as a particular breed if both its parents were also registered as the same breed. This means that each breed is genetically isolated. The problem is amplified by the existence of breed standards, detailing a “perfect” example of a breed. Breed standards provide a strong artificial selection pressure, resulting in only a small subset of dogs being used as breeding stock, which causes a further reduction in genetic diversity. Within the reduced breeding stock there are often dogs that are particularly fashionable because of their close resemblance to the breed standard. These dogs are often extensively used for breeding purposes which leads to an overrepresentation of particular gene variants within a breed. This is commonly known as the popular sire effect. Closed populations and limited breeding stocks, often containing popular sires, can lead to random deleterious mutations reaching very high levels. For example, if a popular sire is a carrier of an autosomal recessive

disease, on average the causal mutation will be passed on to 50% of the offspring, rapidly increasing the frequency of the mutant allele within the breed population.

1.3. The dog as a model organism

Dogs suffer from a range of disorders that have comparable human conditions, and therefore provide naturally occurring models for human disease. The total number of canine genetic diseases was estimated at 479 in 2004 (Sargan, 2004), and the Online Mendelian Inheritance in Animals (OMIA) database suggests at least 286 disease phenotypes are potential models for human disease (OMIA 2012). In addition to showing disease phenotypes comparable with human disorders, the dog is considered to be physiologically similar to humans and physically more similar in size than traditional model organisms such as the mouse, making them more appropriate as a comparative tool and for trialling therapies. Dogs are relatively long lived and disease treatments for both canine and human conditions are often very similar. Even the domestic dog's living environment and lifestyle often mirrors that of humans.

One example of the suitability of the dog as a comparative model is its current use in the field of gene therapy, with the reversal of blindness in dogs with a homozygous mutation in the *RPE65* (retinal pigment epithelium-specific 65 kDa protein) gene (Acland et al., 2001, Le Meur et al., 2007). *RPE65* mutations cause a form of Leber congenital amaurosis in humans, a severe retinal dystrophy which progresses before the age of one (Cremers et al., 2002). It was recently reported that three human Leber congenital amaurosis patients showed significant improvements in vision after being subjected to two rounds of gene therapy with an adeno-associated virus vector, illustrating successful transition from the dog model through to use in human medicine (Bennett et al., 2012).

The discovery of the genetic cause of canine narcolepsy is an example of how a breakthrough in the dog has made a major contribution to human studies. Two independent mutations identified in the hypocretin-2-receptor gene (*HCRTR2*) were shown to cause aberrant splicing patterns, resulting in the skipping of exon 4 in the narcoleptic Dobermann Pinscher and of exon 6 in the narcoleptic Labrador Retriever (Lin et al., 1999). Following the discovery of the gene causing narcolepsy in the dog, hypocretin deficiencies and a reduction in the number of hypocretin neurons were subsequently described in association with human narcolepsy, showing how the study in the dog had helped accelerate the understanding of human disease (Nishino et al., 2000, Peyron et al., 2000, Thannickal et al., 2000).

The research into canine narcolepsy was a long term study. The investigation started with a linkage-based genome scan using a genome-wide set of microsatellite markers, identifying one marker on chromosome 12 in complete linkage with the disorder for both the Dobermann Pinscher and Labrador Retriever. A canine BAC (bacterial artificial chromosome) library was then constructed and probed for the initially linked marker. Chromosome walking by BAC end sequencing was used to build a 1.8 Mb region around the linked marker consisting of 77 contiguous BACs. Eleven polymorphic microsatellite markers identified by probing the BACs were all concordant with Labrador and Dobermann Pinscher cases, and the mutation containing region could not be narrowed at this stage. Basic Local Alignment Search Tool (BLAST) searches of BAC end sequences against human databases revealed close sequence similarity with *MYO6*, helping to identify the syntenic human chromosome. Human expressed sequence tags (ESTs) in the *MYO6* region were then used to probe for further BACs, which could be used in fluorescent *in situ* hybridisation (FISH) against metaphase spreads of canine chromosomes to confirm their existence on canine chromosome 12. Backcross breeding was undertaken in parallel with positional cloning experiments in the hope of producing affected animals with recombinations between the initially linked marker and the narcolepsy locus to narrow down the associated region. Over 100 backcross animals were used in conjunction with a litter of narcoleptic Dachshunds to narrow the critical region to approximately 800 kb, which contained just one previously identified gene, *HCRTR2*. Degenerate primers for cDNA sequencing were designed using published human and rat *HCRTR2* cDNA sequences, which were successfully used to sequence canine *HRCTR2* mRNA and identify the causal exon skipping events. Although the study into canine narcolepsy was one of the first examples showing the potential of the dog model for identifying novel disease-associated genes, the complexity of the investigation clearly showed that the canine genome would need to be sequenced in order to make the dog a more realistic model that could be used routinely.

1.4. Canine genomics

1.4.1. The canine karyotype

The canine karyotype was first characterised in 1928 as 2n = 78 when a study of meiotic cells revealed 38 pairs of acrocentric autosomes and one pair of metacentric sex chromosomes (Minouchi, 1928). Although the number of chromosomes in the canine genome was discovered at an early date, allocating numbers to chromosomes using standard banding techniques alone proved difficult, despite the efforts of several investigators. The Committee for the Standardisation of the Karyotype in the Domestic

Dog (CSKDD) was set up to find a definitive answer, and in 1996 an agreement was reached for allocating numbers 1-21 to canine chromosomes using standard cytogenetic techniques (Switonski et al., 1996). The use of FISH, with whole-chromosome paint probes, helped define the entire canine karyotype, and resulted in a consensus being reached over chromosome nomenclature in 1998 (Breen et al., 1998).

1.4.2. The development of genetic maps

To allow canine geneticists to utilise genetic markers in mapping studies, an initial linkage map of the canine genome was constructed in 1997, by genotyping 150 microsatellite markers across 17 three-generation pedigrees to define 30 linkage groups (Mellersh et al., 1997). As genetically close loci are inherited together, linkage mapping works by analysing the inheritance of marker alleles through generations of a pedigree to assess whether they are in close proximity to one another by generating a statistical score known as a logarithm of odds (LOD). Scores of greater than three indicate significant linkage implying that two markers are close together on a chromosome. Scores of between two and three are defined as being suggestive of linkage. This may mean that markers are on the same chromosome, but linked inheritance is often disrupted by recombination events between the two loci. Scores less than two exclude linkage, indicating that markers are on separate chromosomes or are sufficiently distant to effectively be inherited independently (Lander and Kruglyak, 1995). Just prior to publication of the initial linkage map, copper toxicosis in the Bedlington Terrier became the first canine disease loci to be mapped using the linkage approach (Yuzbasiyan-Gurkan et al., 1997). Radiation hybrid mapping (RH-mapping) approaches were used to further build on previous linkage mapping efforts to produce integrated genome maps (Priat et al., 1998, Vignaux et al., 1999). In the RH-mapping approach a culture of canine fibroblast cells is gamma irradiated to fragment the genetic material. Irradiated canine donor cells are then fused with thymidine kinase-deficient hamster cells. During the fusion process donor chromosome fragments are incorporated into recipient chromosomes by insertion or translocation. Fused cells are then grown in selective media, so only cell lines that have acquired the thymidine kinase gene survive, indicating successful fusion and incorporation of donor genetic material. Colonies are tested by polymerase chain reaction (PCR) to determine which genetic markers are present, with closely linked markers more likely to be contained in the same cell lines. By using microsatellite markers from previous linkage mapping studies in RH-mapping experiments, unlinked regions were joined to form larger regions. Known canine genes and human ESTs were also used in early RH-mapping experiments, allowing regions of synteny between the canine and human genomes to be considered for the first time. These initial linkage and RH maps provided the foundations for a second generation

of linkage maps (Neff et al., 1999) and combined RH-linkage maps (Mellersh et al., 2000). With the completion of the human and mouse genomes, further resources became available for building subsequent canine genome maps, culminating in a map consisting of 1,800 markers, covering all canine chromosomes (Breen et al., 2001).

1.4.3. The canine genome project

The first attempt to sequence the canine genome was undertaken by Celera Genomics. Genomic DNA from a Standard Poodle was used to produce a total of 6.22 million sequence reads, equivalent to 1.5x coverage of the dog genome and sufficient to give 77% base coverage of the genome (Kirkness et al., 2003). Over 25% of the sequence data generated aligned uniquely to the human genome, allowing a full map of synteny between canine, human and murine chromosomes to be produced. Over 18 thousand sequences aligned to annotated genes in the human genome. Building on this information a high resolution RH map was created using 10,348 markers, of which 9,850 were canine orthologue sequences of human genes (Hitte et al., 2004).

The National Human Genome Research Institute (NHGRI) also saw the potential of the dog as a model of genetic disease and invested $30 million towards sequencing the canine genome, which was previously only commercially available. The first open-access assembly of the canine genome was completed and became available in 2004 (Lindblad-Toh et al., 2005). A female Boxer, named Tasha, was chosen for the project as preliminary analysis indicated the sequence of this particular dog to be highly homozygous. The project produced 7.8x coverage of the canine genome from 35 million sequence reads, and the assembly is freely available online via the Ensembl and UCSC genome browsers (Hubbard et al., 2002, Kent et al., 2002). The sequence is annotated to show predicted genes, gene alignments and syntenic regions between species. This annotation was an essential prerequisite for the development of gene arrays for the analysis of genome-wide gene expression, such as the Affymetrix Canine GeneChip, which has probes to monitor over 18,000 canine mRNA/EST-based transcripts and over 20,000 non-redundant predicted genes. By comparing the genome sequence against the 1.5x partial poodle genome sequence and 100,000 sequence reads generated from nine other breeds of dog, a dense SNP map consisting of 2.5 million SNPs was formed. This SNP resource was the key requirement for the development of early high throughput SNP genotyping arrays such as the Affymetrix Canine Platinum sets v1 and v2, the Illumina CanineSNP20 beadchip, and most recently the Illumina CanineHD beadchip, which assays for over 170,000 genome wide SNPs and is most commonly used in today's genome-wide association studies (GWAS).

1.4.4. Genome-wide association studies

In the GWAS approach, case-control sample cohorts are collected and genotyped on high throughput genotyping arrays such as the Illumina CanineHD beadchip. Illumina beadchips consist of microscope slides that are covered in thousands of tiny wells, which are designed to hold 3 µm wide silicon beads. Each bead is covered in a clonal population of oligonucleotides, which act as specific capture probes for the SNP genotyping assays (Figure 1.1).

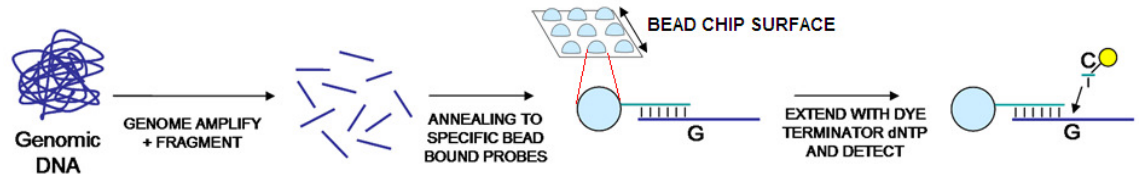


Figure 1.1 The Illumina Infinium assay

In the Infinium assay, genomic DNA is fragmented, annealed to bead bound probes and allele determined by detection of the dye-terminator nucleotide incorporated.

The genomic DNA under investigation is whole genome amplified, fragmented and specifically hybridised to the bead bound oligonucleotide sequences. Annealed fragments then undergo a single base extension at the polymorphism loci, using a DNA polymerase and fluorescent dye terminator nucleotides. The identity of the incorporated base, which is dependent on the SNP allele, is identified by fluorescence.

The genotyping data generated from SNP arrays is subjected to statistical testing known as allelic association analysis, with the null hypothesis that there is no significant difference in allele frequency between cases and controls for a particular SNP marker. Because of the huge number of SNP markers that are under investigation in GWAS approaches, there is a high probability that significant association signals are produced by chance. Permutation analysis is therefore often performed to adjust the results for multiple testing. Results that are significant at the 5% level after permutation testing are often referred to as being genome-wide significant.

Careful selection of controls for GWAS is critically important to avoid the potential effects of population stratification. Population stratification refers to the variable level of relatedness of individuals within a cohort. Population stratification may have a negative impact on association studies when the individuals in the control group are from a different sub-population to those in the case group. Controls from a different population subset may be genetically different because of artificial selection or genetic drift, and will result in false

positive statistical signals being produced. The potential effects of population stratification can be avoided by using controls that are closely related to cases, but in situations where this is not possible computational methods exist to adjust for population stratification in datasets.

1.4.5. Candidate gene studies

Candidate gene studies can be performed when similar diseases occur in humans and dogs. *In silico* literature searches can be used to identify genes that are associated with disease in humans, and using the available online genome browsing facilities, canine orthologues of these genes can be found. Candidate genes can be investigated for their role in canine disease, typically by testing for association between closely linked markers and the disease under investigation. Microsatellite markers flanking candidate genes can be located directly from the genome sequence, and genotyped using PCR followed by capillary electrophoresis, to look for a pattern of linkage disequilibrium (LD) (Mellersh, 2008, Mellersh et al., 2006).

1.4.6. Linkage disequilibrium in the dog

Linkage disequilibrium is defined as the non-random association of alleles at two or more loci, ie the region of shared haplotype around a particular genotype or a disease mutation. Dogs of the same breed show a high level of LD across the genome, which in some breeds is up to 100 times more extensive than in humans (Sutter et al., 2004). The extensive LD is the result of most breed populations being formed within the last 200 years from small numbers of founding individuals. Gene mapping studies can be aided by extensive LD as broad shared haplotypes around the disease-associated mutation are often seen, making it easier to identify disease-associated critical regions in GWAS, as fewer markers need to be analysed. In some instances it is possible to find a disease-associated region from very small numbers of starting cases and controls (Hillbertz and Andersson, 2006). The extent of LD in the dog does however present a more challenging task once the critical region for a disease has been isolated. Disease-associated regions are typically long, as a result of the extensive LD, and if the mapped critical region happens to be gene dense, prioritising candidate genes can be difficult. The challenging task of pinpointing precise causal mutations can be alleviated in some cases by utilising the close relationships which are present between many breeds. If related breeds suffer from clinically similar diseases, there is a high probability that the diseases have an identical genetic cause. Since the similar breeds diverged, independent recombination events will have occurred around the causal mutation, resulting in each breed having slightly different disease-associated haplotypes. The overlapping critical regions of the

different breeds can be combined to identify a single shared critical haplotype, helping to home in on the faulty gene. Examples of using breed relationships and overlapping regions of homozygosity to aid disease mapping include the studies of progressive rod-cone degeneration (PRCD), a late-onset autosomal recessive photoreceptor disorder, and Collie eye anomaly, a complex ocular development disorder (Parker et al., 2007, Zangerl et al., 2006). Interestingly, Collie eye anomaly was initially mapped in 2003, but the causal mutation was not identified until 2007. Fine mapping the disease haplotype using the multi-breed approach, helped to reduce the critical region from an initial 3.9 Mb region to a 691 kb region on chromosome 37. Even after fine mapping the disorder, no exonic mutations were found in the four known genes in the region. The causal mutation was eventually found to be a 7.8 kb deletion in a partially conserved intronic region.

1.4.7. Complex disease mapping in the dog

With the dog proving to be a useful model organism, the Lupa project was set up in January 2008 funded by the European Commission (Copeland et al., 2008). The aim of the Lupa project was to use canine genetics to unravel common human diseases such as cancer, cardiovascular disease, inflammatory disorders and neurological diseases. These diseases have traditionally been difficult to study in humans due to their genetic complexity and phenotypic heterogeneity. In the dog, cases are collected from genetically isolated breed populations and subjected to GWAS using SNP array technologies. The high frequency of particular disorders within certain breed populations allows for quick collection of case samples, and the reduced phenotypic and genetic heterogeneity between cases means that significant results can often be obtained with far fewer samples than in human studies. For example, there is a high incidence of histiocytic cancers in the Flat Coated Retriever and the Bernese Mountain Dog, allowing large sample cohorts to be collected in a reasonable timeframe. Dual investigation of the genetics of histiocytic cancers in these two breeds may help to unravel genes associated with cancer formation and also the effects of genetic background on tumour development (Hedan et al., 2011).

The dog model could help in the study of epilepsy; a term which is used to cover a broad spectrum of complex neurological conditions. This heterogeneity makes it a particularly difficult disease to study in the human population, as each individual idiopathic epilepsy case may be caused by a number of mutations in several genes, all which have a small contribution towards the disease phenotype. This means that within a case-control cohort there could be many epilepsy genes represented, and as a result a huge number of samples would be required to give the study sufficient power. Because of the complexity

of human epilepsy, finding the underlying genetic causes is not possible on a case by case basis and the inroads that have been made to date have largely been made through the study of rare autosomal recessive epilepsy syndromes (Sanchez-Carpintero Abad et al., 2007, Steinlein, 2008). The genes identified as causal in these syndromes help to improve the understanding of the molecular pathways involved and can be screened for mutations in idiopathic epilepsy cases. Because dog breeds exist as genetically isolated populations, epilepsy in a single dog breed is likely to be less genetically complex with far fewer genes involved, meaning a much smaller number of samples would be required to map the associated genes. A recent success story is the identification of a novel epilepsy locus in the Belgian Shepherd, which is currently being sequenced with the hope of identifying a novel genetic risk factor (Seppala et al., 2012). Identification of an associated gene will provide a new candidate for human idiopathic epilepsy cases.

1.5. The development of massively parallel sequencing techniques

The development of new massively parallel sequencing techniques, nicknamed “next generation sequencing” (NGS), has revolutionised the speed at which large disease-associated regions of the genome can be investigated. With NGS, genomic libraries are created consisting of millions of short strands of DNA, which are sequenced in parallel to produce enormous datasets. Sequencing of megabases of DNA from a single experiment is easily achievable, and is a huge advance on Sanger sequencing reactions which generate a maximum of around one kilobase of sequence data. The four main next generation sequencing platforms currently available are Illumina (Solexa), ABI SOLiD, Roche 454 and most recently Ion Torrent.

1.5.1. Illumina sequencing

Illumina sequencing is probably the most established next generation sequencing technology currently available. The technology has the greatest availability for outsourcing and is the machine that the Wellcome Trust Sanger Institute has most heavily invested in at their genome campus.

Illumina sequencing technology originally stemmed from two companies; Lynx therapeutics, which developed a technique known as Massively Parallel Signature Sequencing (MPSS) (Brenner et al., 2000) and Solexa who developed solid phase sequencing using fluorescent reversible bases, now commonly known as sequencing by synthesis (SBS). In 2005 Solexa merged with Lynx therapeutic to form an international company, whose combined expertise culminated in the development of Solexa’s first SBS machine, the Genome Analyser.

This SBS machine was capable of generating 1 Gb of sequencing data in a single run, and subsequently Illumina have gone on to increase capacity through the development of the Genome Analyser IIx and the HiSeq 2000, which has the capacity to generate over 600 Gb of sequencing data per run. In the fourth quarter of 2011 Illumina launched a benchtop machine, the MiSeq, marketed as a user friendly personal genome machine for use in smaller laboratories and for point of care work.

Illumina sequencing is widely regarded as the most accurate massively parallel sequencing technology, although has not been able to produce the long reads obtainable through some other next generation platforms. Maximum read lengths for Illumina machines currently stand at 250 bp, however, chemistry to support 350 bp reads on the MiSeq platform is currently being developed, which is projected to become commercially available in the fourth quarter of 2013.

Illumina sequencing experiments start with the preparation of a DNA sequencing library (Figure 1.2). The starting DNA, which may be genomic, plasmid, PCR amplicon or otherwise, is fragmented either enzymatically or mechanically before end repair and dA tailing reactions are carried out. A DNA adapter is then ligated to the ends of fragments allowing the fragments to be amplified and an index sequence incorporated if libraries are to be combined for multiplexed sequencing protocols.

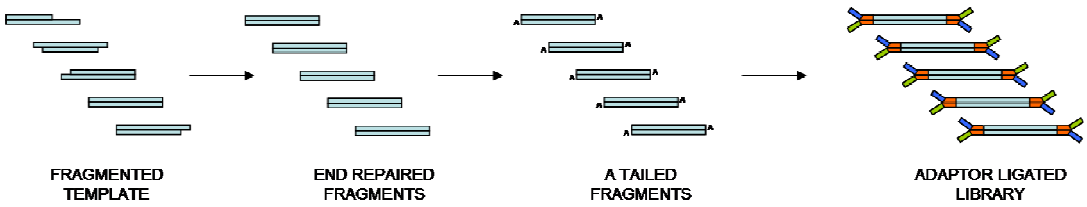


Figure 1.2 Stages of the Illumina DNA library preparation

In the Illumina library preparation DNA is fragmented, end repaired and dA tailed. Adapters are ligated onto fragment ends, before amplification by PCR if required.

The next stage of Illumina sequencing is to bind library fragments to the surface of a flowcell, for the generation of clonal clusters in a process known as bridge amplification. Flowcells have a lawn of oligonucleotides bound to their surface which are complementary to adapter sequences. As the sequencing library is passed across the surface, fragments randomly anneal to the bound oligonucleotides, before a process of bridge amplification occurs (Figure 1.3). Resulting clusters contain approximately 1,000 clonal sequences.

Cluster density is dependant on library concentration. If the concentration is too high clusters will be too close together resulting in mixed signals and sequencing inaccuracies. If the concentration is low data output will be low as a consequence.

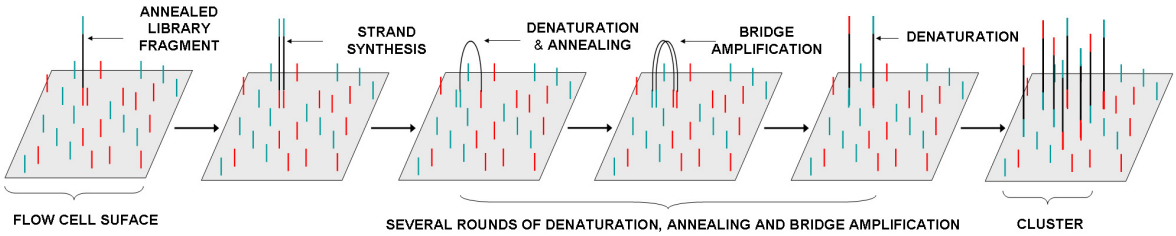


Figure 1.3 Cluster generation by bridge amplification.

The generation of clonal clusters on flow cell surfaces by bridge amplification PCR (Shendure and Ji, 2008).

On the completion of cluster formation the process of sequencing by synthesis can begin. Firstly reverse strands are cleaved to ensure that all sequenced strands are in the same orientation. A sequencing oligo is then bound to fragments and the SBS process is started by the addition of reversible dye terminator nucleotide bases (Figure 1.4). After a nucleotide is incorporated, clusters are excited by laser and photographed. The colour of the photographed cluster is dependant on the nucleotide base that is incorporated. The dye terminator groups are removed from the bases allowing the next dye terminator nucleotide to be incorporated and the process repeated up to 250 times to build the sequence of each cluster.

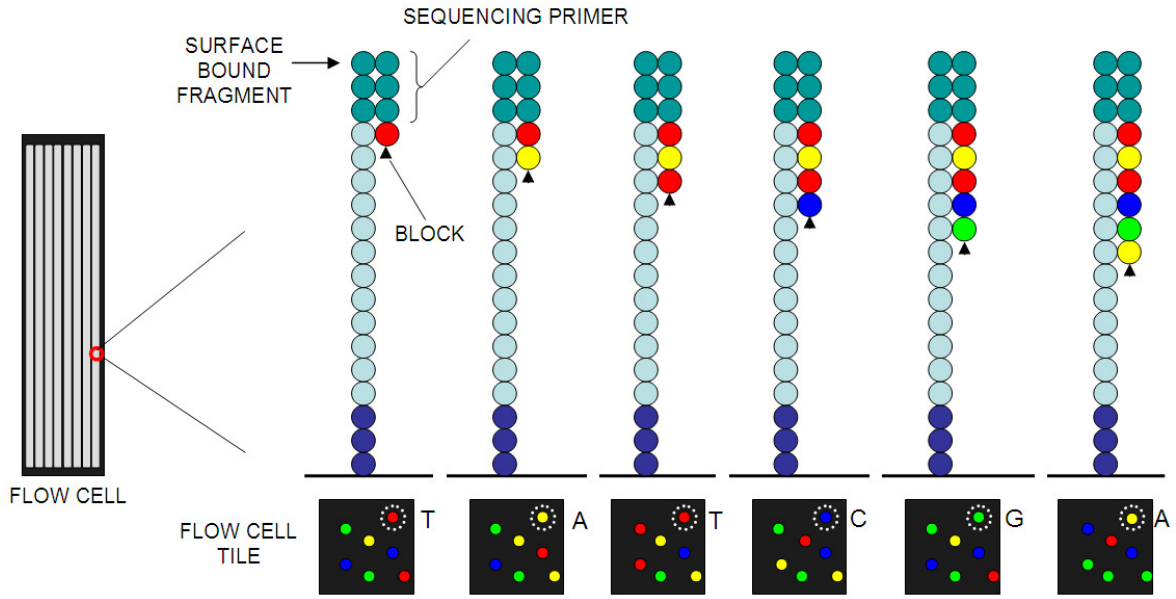


Figure 1.4 Sequencing by synthesis

On the surface of the flowcell a sequencing primer is annealed to bound fragments. Dye-terminator nucleotides are then incorporated and detected by fluorescence. Removal of the “block” between cycles allows the next dye-terminator nucleotide to be incorporated and identified. Several cycles are completed to enable the sequence of the fragments in the clusters to be determined.

Other applications of Illumina sequencing include RNA sequencing (RNA-seq) and chromatin immunoprecipitation sequencing (ChIP-seq). RNA sequencing can be used as a method of transcriptome sequencing, which may be total or just coding if messenger RNA is isolated before the library preparation. Library preparation is essentially similar in theory to the DNA library preparation, but includes a reverse transcription stage. RNA-seq can be used as an alternative method of expression analysis, and can also be used to help enhance genome annotation by filling in missing genes and improving accuracy of gene predictions, and has recently been suggested for the canine genome (Derrien et al., 2012). Protocols are also available for the isolation of microRNAs and other small non-coding RNAs.

With ChIP-seq, protein-DNA interactions can be investigated, such as those seen between transcription factors and promoter DNA sequences. ChIP-seq works by binding and cross linking proteins to DNA sequences. Proteins are then captured by magnetic particle bound antibodies. The DNA strands bound to the capture protein are then unlinked, releasing them for sequencing. ChIP-seq may be a potential method of investigating the cause of gene down-regulation in complex disease studies.

1.5.2. 454 sequencing

The 454 sequencing technology is based on a technique known as pyrosequencing developed by Pål Nyrén in the 1990s (Ronaghi et al., 1996), and was introduced commercially in 2005 (Margulies et al., 2005). The pyrosequencing process relies on a three enzyme system consisting of DNA polymerase, sulfurylase and luciferase (Figure 1.5). As the DNA polymerase incorporates a nucleotide, a phosphate group is released which is converted to ATP by the sulfurylase. The luciferase then catalyses the conversion of luciferin to oxyluciferin using the ATP generated, a reaction which releases a burst of light which can be measured.

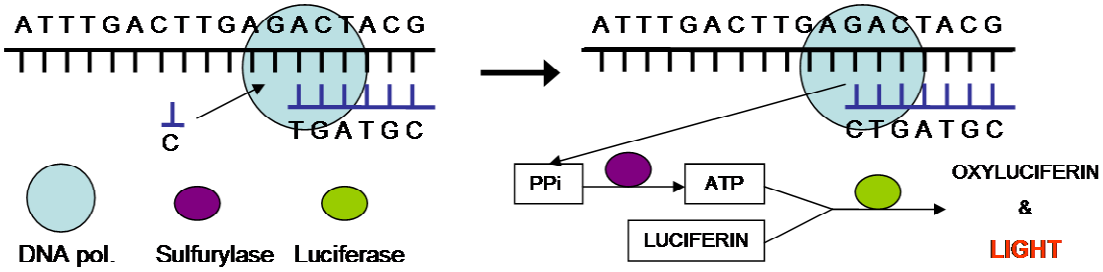


Figure 1.5 The pyrosequencing process

As a nucleotide is incorporated by DNA polymerase a pyrophosphate group (PPi) is released, which is converted to ATP by sulfurylase. In the presence of ATP, luciferin is converted to oxyluciferin by luciferase, producing a burst of light.

Adenine, cytosine, thymine and guanine are added sequentially and cyclically to the reaction, with the intensity of the burst of light produced being proportional to the number of bases incorporated and can be used to calculate the sequence.

The library preparation for 454 sequencing is essentially similar to the method used for Illumina sequencing. DNA is first fragmented and then short adapter sequences are added. Adapter linked fragments then undergo a process of emulsion PCR to produce clonal populations on bead surfaces in a process similar to bridge amplification, with individual fragments isolated in a bead containing droplet rather than on the surface of a flowcell (Figure 1.6).

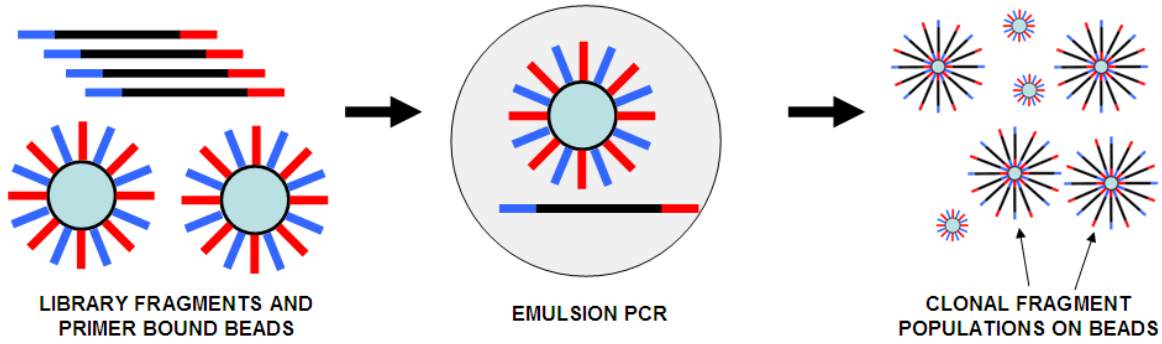


Figure 1.6 Emulsion PCR

Emulsion PCR to produce a clonal population of a library fragment on the surface of a bead for 454 sequencing (Shendure and Ji, 2008).

Beads from the emulsion PCR process are pre-incubated with DNA polymerase and flooded onto a picotitre plate (PTP), with each well having a single bead capacity (Figure 1.7). Bead immobilised sulfurylase and luciferase are also added to the wells. As reagents are flowed onto the PTP, the pyrosequencing reaction begins, with bursts of light being produced in wells where nucleotide bases have been incorporated.

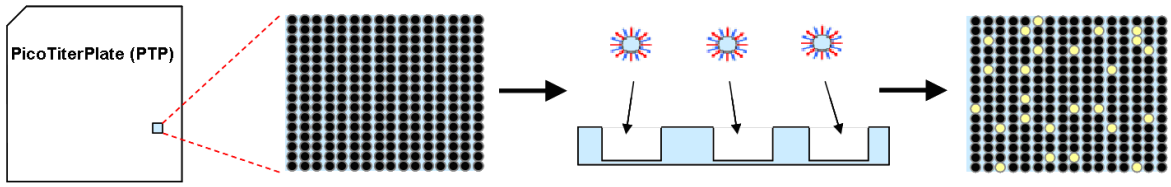


Figure 1.7 454 sequencing in picotitre plates (PTPs)

Each well of the PTP is filled with a single clonally populated bead. As nucleotides are sequentially introduced, wells of the PTP light up if a base is incorporated, enabling the sequence to be determined.

The main advantage of 454 sequencing is the length of read that can be produced by the technique. Reads of 1,000 bp can be produced using the latest chemistry, although 454 sequencers have difficulty accurately determining the length of homopolymer repeat sequences. There are currently two 454 systems on the market; these are the GS FLX+ which can generate 700 Mb of data in a single run, taking approximately a day, and the GS Junior, which can generate around 30 Mb of data in 10 hours.

1.5.3. ABI SOLiD sequencing

The Sequencing by Oligonucleotide Ligation and Detection (SOLiD) platform, uses a sequencing method first described by Shendue and colleagues in 1995 and developed by Agencourt Personal Genomics before they were acquired by Applied Biosystems in 1996 (Shendure et al., 2005). The SOLiD system is unusual in the fact that nucleotide sequences are not identified consecutively and several rounds of sequencing using synthesis by ligation are required to build the consensus sequence of a particular strand.

Library preparation for SOLiD sequencing requires DNA fragmentation and adapter linking, followed by a bead-based emulsion PCR approach. At the end of the emulsion PCR, bead attached amplicons are 3' modified, allowing covalent bonding to the surface of a slide. Sequencing begins by annealing the first of five sequencing primers to the bead attached strands. Eight nucleotide probes, consisting of two identifier bases, three degenerate bases and three universal bases linked to one of four different fluorescent tags, are then specifically ligated to the sequencing primer, depending on the template sequence. In total there are 16 identifier combinations and a total of 1,012 possible probes (two identifier bases + three degenerate bases = 4^5). After the first probe has been incorporated and identified by laser, cleavage occurs between bases five and six to remove the universal bases and fluorescent tag, allowing the second probe to be incorporated. This sequencing by ligation process is repeated for several cycles. Base positions 1 and 2, 6 and 7, 11 and 12, etc, are identified in this round of sequencing, before the synthesised strand is removed and a second sequence primer is annealed at position (n-1). In the second round bases 0 and 1, 5 and 6, 10 and 11, etc, are read. A total of five rounds of sequencing by ligation are completed, with each nucleotide base being read twice to enable the sequence of the DNA strand to be determined (Figure 1.8).

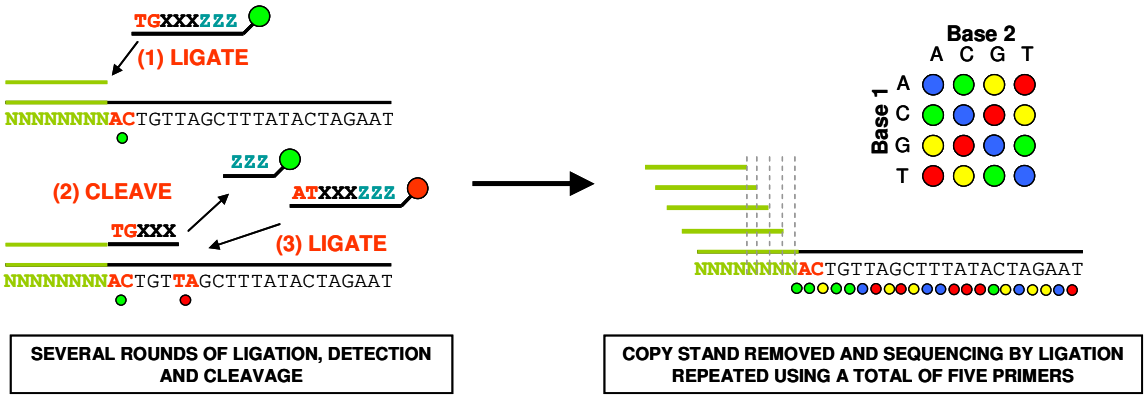


Figure 1.8 SOLiD sequencing

Probes consisting of two identifier bases (red), three degenerate bases (XXX) and three universal bases (ZZZ) are specifically incorporated at the priming site by ligation and detected by fluorescence. The universal bases are cleaved ready for the second cycle of ligation and detection. Several cycles are completed before removal of the synthesised strand and repetition of the process with four other sequencing primers at alternative start positions (green).

The current SOLiD system is the 5500xl, which has maximum reads lengths of 75 bp and can generate up to 20 Gb of data per day.

1.5.4. Ion Torrent sequencing

The Ion Torrent platform is a non optical method of sequencing, which utilises semi-conductor technology. Ion Torrent was launched in 2011, with a paper describing the methodology published by J. Rothberg and colleagues (Rothberg et al., 2011). Rather than relying on chemistry that requires the use of lasers to detect incorporated nucleotides such as the Illumina and SOLiD technologies, or the need for additional enzymatics in the case of 454 pyrosequencing, the Ion Torrent system works by detecting the shift in pH that naturally accompanies the incorporation of a nucleotide into a DNA sequence via the release of a hydrogen ion.

Similar to other sequencing technologies, library preparation consists of DNA fragmentation followed by the addition of adapters to the fragment ends, which are then amplified onto beads in an emulsion PCR process. Beads containing clonally amplified products are then placed into the wells of a chip. Perfectly aligned underneath each well is an ion-sensitive field-effect transistor-based sensor (Figure 1.9) (Rothberg et al., 2011).

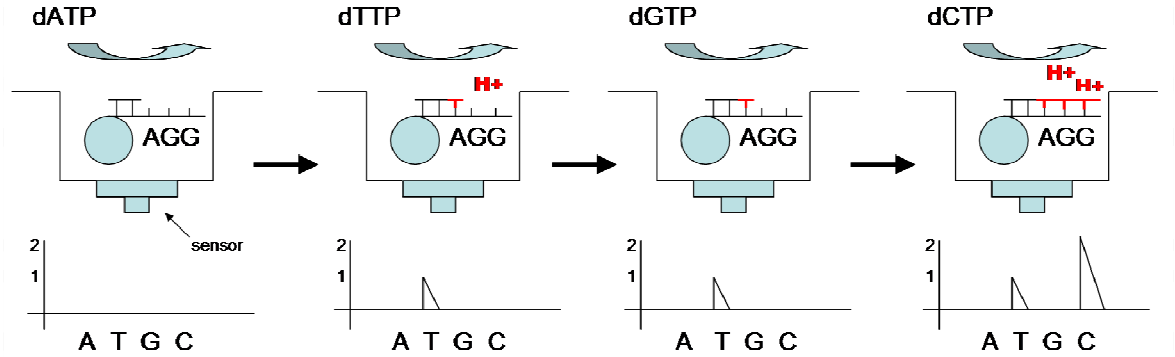


Figure 1.9 Ion Torrent sequencing

Nucleotides are sequentially flooded across the sequencing chip. If the bases are incorporated hydrogen ions are released causing a shift in pH which is detected by the sensor.

The simple chemistry of the Ion Torrent system allows for fast sequencing times and because the sequencing reactions and detection are done on the chip, different options on well capacity are available making sequencing reactions easily scalable to match requirements. Ion Torrent sequencing is also capable of producing long reads, but (like 454 sequencing) is inaccurate over homopolymer tracts.

Two machines are available for the Ion Torrent platform; the Personal Genome Machine (PGM) and the Ion Proton. The Ion Proton was launched in the first quarter of 2012, with the manufacturers aiming to launch a second version of the Proton sequencing chip (Proton II) capable of sequencing the human genome in one day.

1.6. Target enrichment

The amount of data produced by next generation sequencing is so vast that large genomic regions spanning several megabases can be sequenced in a single experiment. Generating a template of this size using standard PCR techniques is possible, but can be time consuming, expensive and technically difficult. With a need for fast and large-scale template generation in next generation sequencing, several array based and in-solution target enrichment techniques have been developed to allow specific capture of genomic regions (Mamanova et al., 2010). These include Roche Nimblegen SeqCap, Illumina TruSeq Enrichment Kits and Agilent SureSelect. Agilent, with their SureSelect system, were the first company to offer a custom in-solution kit, allowing capture designs to be tailored to the dog genome (Figure 1.10).

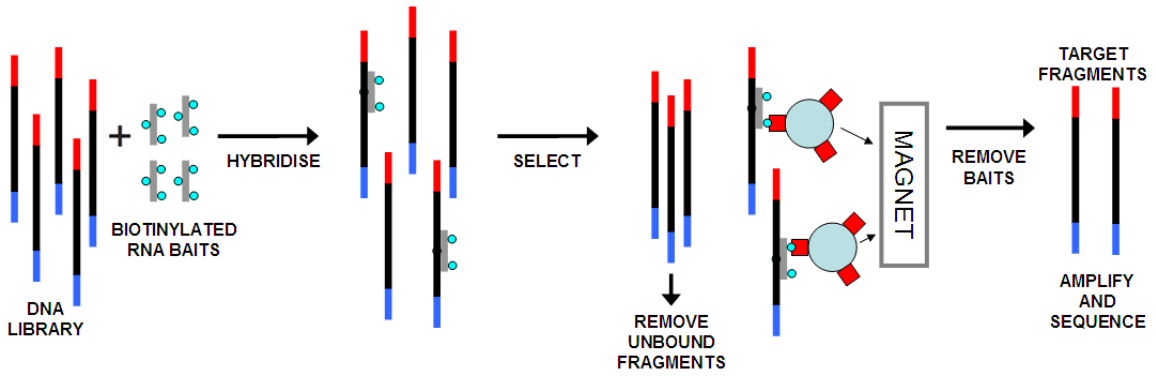


Figure 1.10 SureSelect target enrichment

Library fragments hybridise with biotinylated RNA baits, which are captured by streptavidin coated magnetic bead.

The great advantage of an in-solution kit is that capture experiments can be performed in the standard laboratory environment, without the need for specialist equipment. In-house sequencing library preparation is desirable as it is far more cost effective than outsourcing. Capture in the SureSelect system is achieved using a system of RNA based probes, also known as baits (Gnirke et al., 2009). Using the online tool, e-array, baits can be designed for specified regions of a genome, totalling up to 6.6 Mb (55,000 RNA baits of 120 bp in length). The DNA library is prepared in a similar way to a standard method using fragmentation and the addition of adapter sequences to the ends. RNA baits, which are also biotinylated, are then hybridised to the library fragments to capture the sequences of interest. Biotinylated RNA baits which are hybridised to target DNA fragments are then selected for using streptavidin coated magnetic beads, which strongly bind to the biotin molecules, allowing unbound non-target fragments of the library preparation to be removed. RNA baits can then be digested to leave target fragments of DNA that can be amplified and used as a template for sequencing.

Pre-designed capture enrichment products are also available including exome enrichment kits and cancer panel enrichment kits. Agilent technologies launched a canine exome capture kit in the fourth quarter of 2012 and although not commercially advertised is available on request. An alternative method to probe-based target enrichment is ultra-high multiplex PCR (Porreca et al., 2007). Custom kits for this method include the Illumina TruSeq Custom Amplicon kit and the Ion Ampli-seq kit.

1.7. Summary

It is an exciting time to be a molecular geneticist as the toolset of technologies available is rapidly expanding for both broad spectrum and canine specific applications. With these technological advances there is an inevitable increase in the complexity of techniques and

a considerable challenge is to select the most appropriate research methodology relevant to a particular disorder. In particular, smaller laboratories need to find ways to explore these techniques and optimise the procedures where possible to make them more financially and technically accessible. Use of the most up to date techniques is absolutely critical to ensure that any laboratory remains competitive with other groups within the same field and stays an attractive proposition in terms of collaborative work. Staying in touch with available technologies also helps to reduce the timescale of proposed work, increasing the chance of goals being reached on time. This may convince funding bodies that the group is at the forefront of research in the field and will give confidence that project aims are achievable, maximising the chance that a grant proposal will be accepted.

Historically technological restrictions have meant that knowledge would flow in a single direction from the human to dog field, with the dog being a rather theoretical model. With the recent availability of new technologies, and the ability to adapt these technologies for use in canine genetics, the dog model is likely to become a more applied model for identifying novel genes that could shed light on human disease.

1.8. Aims

In this study molecular techniques have been used to investigate five inherited conditions afflicting four canine breeds. The study demonstrates how current research methodologies can be optimised to best suit the particular characteristics of each disease. The disorders investigated in this study included episodic falling (EF) and congenital keratoconjunctivitis sicca and ichthyosiform dermatosis (CKCSID), both in the Cavalier King Charles Spaniel (CKCS), spinocerebellar ataxia in the Italian Spinone (IS), cerebellar ataxia in the Parson Russell Terrier (PRT) and neonatal cerebellar cortical degeneration in the Beagle. The phenotypes under investigation demonstrated a wide variety of clinical signs, and had different ages of onset, progressions and prognoses. These factors meant that each disease had to be treated very much on an individual basis when investigating the molecular biology. The type of investigation performed also depended on the availability of samples and strength of diagnosis. For instance, when clinical signs were non-uniform between cases and diagnosis of disease was difficult because a spectrum of disease from a mild to more severe phenotype was displayed, then a greater case-control cohort was needed to gain significant results.

Outsourcing of some work was unavoidable, where facilities were neither available at the University of Glasgow or the Animal Health Trust (AHT). However, wherever possible, the

aim was to perform the bench work and analysis in-house. This had the advantage of saving considerable amounts of money and also helped to improve understanding of the techniques being utilised and the results being obtained. Some techniques needed to be customised for in-house use.

The conditions investigated are highly debilitating so any advances in the molecular genetics made as part of this study also helped the understanding of both canine and human forms of the diseases. Neurological diseases make up four of the five disorders under investigation. The nervous system is relatively inaccessible for investigative work and for treatment, so building an understanding of the molecular pathways involved in nervous system disease is vital and may aid the development of novel therapeutic approaches for treatment.

The primary experimental aims of this study were to A) demonstrate the research methodologies available and to show how these could be best optimised to investigate the molecular biology of five different disorders in the dog, B) to attempt to identify the disease causing mutations in these disorders, and C) where the disease causing mutations were identified, to develop diagnostic tests for these disorders.

Chapter

2. Materials and Methods

2.1. Definition of cases and controls

2.1.1. Definition of EF cases

Episodic falling cases were defined as dogs that were reported by their owner or veterinarian to have a clinical phenotype consistent with EF as described in the veterinary and scientific literature. Cases were confirmed in consultation with a veterinary neurologist, using video evidence where possible. Full neurological work-ups were also performed in some cases to rule out other possible causes of the clinical signs. There is no definitive way of diagnosing EF, and diagnosis is therefore made on presentation of a consistent clinical phenotype with exclusion of all other possible underlying causes. Age of onset was confirmed for all cases on DNA sample submission.

2.1.2. Definition of CKCSID cases

All cases of CKCSID were clinically assessed by a single veterinary ophthalmologist (Claudia Hartley). Diagnosis was based on the presentation of clinical signs consistent with previously seen cases and reports in the veterinary literature. Age of onset was recorded by Claudia Hartley for all confirmed cases.

2.1.3. Definition of Italian Spinone spinocerebellar ataxia cases

Cases of spinocerebellar ataxia in the Italian Spinone were diagnosed by a veterinary neurologist, based on clinical signs of ataxia presenting at 4 months of age and exclusion of all other possible causes by means of a neurological work-up. Age of onset was determined and recorded after consultation with dog owners once the clinical status had been determined by means of a neurological examination.

2.1.4. Definition of Parson Russell Terrier spinocerebellar ataxia cases

Cases of spinocerebellar ataxia in the Parson Russell Terrier were reported by owners, veterinarians or veterinary neurologists, and diagnosis made using the available veterinary notes and video footage, in consultation with a veterinary neurologist. Age of onset was determined by analysis of veterinary notes and by liasing with individual dog owners.

2.1.5. Definition of Beagle neonatal cerebellar cortical degeneration cases

A single case of Beagle neonatal cerebellar cortical degeneration was seen at the AHT which was diagnosed by Elsa Beltran and Dr Luisa De Risio when the dog was four weeks of age. Clinical signs of ataxia were consistent with previous reports in the veterinary literature. A full neurological work-up and histopathology ruled out other

possible causes and confirmed the diagnosis. A second case was diagnosed retrospectively after re-assessment of the reported clinical signs and cerebellum tissue histopathology.

2.1.6. Definition of controls

Controls were defined as individuals reported by owners as clinically normal. Individuals selected as controls for the episodic falling and Parson Russell Terrier ataxia studies were free from clinical signs and at least four years of age, with no history of suspected inherited disease. All controls were selected from the same breed group as the case individuals. Age of individual dogs on sample submission was recorded for all controls. Owners were contacted for a health update if DNA from their dog was to be used in an investigation, but the dog was less than four years of age on sample submission.

2.2. Sample collection

Genomic DNA samples were collected from the general pet dog population, either in the form of residual ethylenediaminetetraacetic acid (EDTA) blood, collected as part of a veterinary procedure, or by collection of buccal cells on cytology brushes. Genomic DNA was also extracted from formalin fixed paraffin embedded (FFPE) tissue for two IS ataxia cases and one neonatal cerebellar cortical degeneration case and from formalin fixed spleen for an additional IS cerebellar ataxia case.

2.3. DNA extraction

2.3.1. Extraction of DNA from whole blood

Genomic DNA was extracted from whole blood using an adaption of the Nucleon BACC2 genomic DNA extraction kit (Tepnal Life Science). In 50 ml tubes whole blood (<10 ml) and Reagent A (Appendix 1) were combined to a total volume of 50 ml. Tubes were mixed by inversion 20 times and centrifuged at 4,500 x g. Supernatants were removed and discarded, and an additional 25 ml Reagent A added to tubes, which were mixed by vigorous shaking for 20 seconds. White blood cells were pelleted by centrifugation at 4,500 x g for 5 minutes. After removal of the supernatant, 2 ml of Reagent B (Appendix 1) were added to pellets, and incubated at 37°C for 16 hours to lyse cells. The lysis mixtures were transferred to 15 ml polypropylene tubes containing 800 µl 5 M sodium perchlorate solution and mixed by inversion 10 times. To the mix 2 ml of ice cold chloroform were then added, mixing by inversion 20 times. Phases were separated by centrifugation at 4,500 x g for 5 minutes, before the upper aqueous phases were transferred to new 15 ml tubes. DNA was precipitated by addition of 5 ml of ice cold 100% ethanol. If a visible precipitate could be seen, the DNA was then spooled onto glass hooks, which were dried

for 5 minutes, before dissolving DNA into 300 µl tris-EDTA (TE) solution (Appendix 1). If no visible DNA precipitate could be seen tubes were centrifuged at 4,500 x g for 30 minutes, supernatant removed, pellets washed in 70% ethanol and tubes re-centrifuged for 10 minutes at the same speed. The 70% ethanol was then removed, DNA dried at 37°C for 1 hour, before rehydrating in 100 µl TE. DNA samples were stored at -20°C.

2.3.2. Extraction of DNA from freshly frozen tissue samples

Small sections of tissue (<100 mg) were dounce homogenised before addition of 2 ml of Nucleon Reagent B and 50 µl proteinase K (600 mAU/ml, solution) (Qiagen). Samples were incubated at 37°C for 16 hours to allow full digestion of the tissue. Proteins in the lysis mix were precipitated by addition of 800 µl sodium perchlorate and DNA purified using the chloroform method outlined in section 2.3.1.

2.3.3. Extraction of DNA from buccal swabs

Genomic DNA was extracted from buccal swabs using the QIAamp Midi kit (Qiagen). Up to 4 buccal swabs were placed in 15 ml tubes and 3 ml of lysis mix added, consisting of 1 ml phosphate-buffered saline (PBS) (Appendix 1), 1 ml AL buffer (Qiagen) and 80 µl protease solution (Qiagen). Tubes were incubated at 56°C for 15 minutes, with occasional vortex mixing. To each tube 1.5 ml of 100% ethanol were added and mixed by shaking. Sample mixes were applied to QIAamp Midi columns, which were centrifuged at 1,850 x g for 3 minutes, discarding the flow-through. To wash the column filters 2 ml of Qiagen AW1 were applied to the columns which were centrifuged at 4,500 x g for 1 minute, discarding the flow-through. A second wash was performed by applying 2 ml of Qiagen AW2 to each column and centrifuging at 4,500 x g for 15 minutes, discarding the flow-through. Columns were placed on new 15 ml tubes and 150 µl Qiagen AE buffer applied directly to the filter. Columns were incubated for 1 minute before centrifuging at 4,500 x g for 1 minute. The elution stage was repeated twice to maximise yield, giving a final elution volume of approximately 450 µl.

2.3.4. Extraction of DNA from FFPE tissue

Genomic DNA was extracted from FFPE tissue using the Nucleospin FFPE DNA kit (Macherey Nagel). Sections of embedded tissue were cut using a microtome. Samples were deparaffinised by adding 400 µl of paraffin dissolver and heating at 60°C for 3 minutes, followed by immediate vigorous vortexing. To the cooled mixture 100 µl of buffer FL were added, followed by vigorous vortexing and centrifugation at 11,000 x g for 1 minute. The upper organic layer was removed. To the sample 10 µl of proteinase K (2.5 U/mg) were added, mixed by pipetting and incubated at room temperature for 3 hours. To de-crosslink the DNA 100 µl of D-link buffer were added to the samples,

incubated at 90℃ for 30 minutes. Samples were allowed to cool to room temperature before addition of 200 µl of 100% ethanol, and vortex mixing of tubes. DNA was bound to a NucleoSpin FFPE DNA column by applying the mixture to the column, and centrifuging at 2,000 x g for 30 seconds. The column was washed twice by applying 400 µl buffer B5 and centrifuging at 11,000 x g for 2 minutes, discarding the flow-through. The DNA was eluted by applying 20 µl of buffer BE directly to the membrane of the column and centrifuging at 11,000 x g for 2 minutes. Extraction of FFPE Beagle cerebellum tissue was performed by Louise Pettitt.

2.4. DNA quantification

Initial quantification of DNA was performed using a Nanodrop spectrophotometer (Thermo Fisher). Purity was assessed using 260/230 and 260/280 measurements. Accurate assessment of DNA concentration was performed using a Qubit Fluorometer (Invitrogen). For high throughput accurate quantification Picogreen methodology was adopted. For Picogreen quantification DNA samples were diluted in TE to between 500 and 1,500 pg/µl based on Nanodrop spectrophotometry results. In a transparent low profile PCR plate 10 µl of diluted DNA were combined with 10 µl of 1x Picogreen reagent (concentration not specified by manufacturers) (Invitrogen) and mixed by pipetting. Plates were sealed using qPCR sealing film (Thermo Scientific). DNA standards of 0, 400, 800, 1,200, 1,600 and 2,000 pg/µl were made by diluting a λ DNA standard in TE buffer (both Invitrogen). Plates were analysed on a Quantica qPCR machine (Techne). Analysis to assess the molecular weight of the DNA (100 ng) was performed by 1% agarose gel electrophoresis.

2.5. Standard PCR

Standard PCR was carried out in 12 µl reactions consisting of 0.2 mM dNTPs (NEB), 1x PCR buffer (Qiagen), 0.5 µM forward primer, 0.5 µM reverse primer, 0.5 units HotStarTaq plus DNA polymerase (Qiagen), template DNA/cDNA and ultrapure water to a volume of 12 µl. For PCR of GC rich amplicons, Q solution (Qiagen) was also added to a final 1x concentration. Cycling parameters for standard PCR were 95℃ for 10 minutes, followed by 35 cycles of 95℃ for 30 seconds, 57℃ for 30 seconds and 72℃ for 60 seconds, and completed with a final elongation stage of 72℃ for 10 minutes. Primer sequences are shown in Appendix 2.

2.6. Agarose gel electrophoresis

For standard 1.5% agarose gel electrophoresis, 1.5 g of agarose were added to 100 ml of 1x tris-acetate-EDTA (TAE) buffer. Solutions were heated in a microwave on full power (800 W) for 2 minutes to dissolve agarose. Liquid gels were cooled to ~60℃ before addition of 2 µl of 10 mg/ml ethidium bromide solution (Gibco) and cast into a sealed 12 x

14 cm tray with an appropriate comb inserted. Set gels were placed into a tank filled with 1x TAE, combs removed and DNA solutions containing 20% loading buffer loaded into wells. Gels were run at 100 V for 1 hour and DNA bands visualised using a UV transilluminator.

2.7. QIAquick PCR product and gel extract purification

PCRs were diluted in 5 volumes of buffer PB (Qiagen). Agarose gel slices were dissolved in 3 volumes of buffer QC (Qiagen) at 50°C for 10 minutes, before adding 1 volume of isopropanol. Dissolved gel extracts/PCRs were mixed well before applying to QIAquick columns (Qiagen) which were centrifuged at 17,000 x g for 1 minute, discarding the flow-through. Residual agarose was removed by applying 500 µl buffer QC to gel extract columns and centrifuging for 1 minute at 17,000 x g, discarding the flow-through. Columns were washed by applying 600 µl buffer PE (Qiagen) and centrifuging for 1 minute at 17,000 x g. Columns were placed on new collection tubes and re-centrifuged for 3 minutes at 17,000 x g to remove any residual ethanol. Purified DNA was eluted by placing columns on 1.5 ml tubes, applying 50 µl buffer EB (Qiagen) and centrifuging at 17,000 x g for 1 minute.

2.8. Multiscreen PCR product purification

To purify PCR products for Sanger sequencing, Multiscreen PCR purification plates (Millipore) were used. Reaction volumes were adjusted to 200 µl by addition of ultrapure water and loaded into wells of the Multiscreen plate. Plates were placed on a manifold and a vacuum of -10 psi applied using a pump to draw liquid through the wells until the nitrocellulose membranes were completely dry. The pump was turned off and 20 µl of ultrapure water applied to the wells, which were incubated at room temperature for 2 minutes to dissolve the DNA.

2.9. Sanger sequencing

Cycle sequencing was performed in 6 µl reactions consisting of ~30 ng purified PCR product template, 0.5 µl Big Dye v3.1 (Applied Biosystems), 1 µl of SBDD buffer, 0.27 µM forward or reverse oligo, and ultrapure water. Sequencing thermal cycling parameters were 96°C for 30 seconds, followed by 44 cycles of 92°C for 4 seconds, 55°C for 4 seconds and 60°C for 90 seconds. Sequencing products were precipitated by the addition of 60 µl of 80% isopropanol to wells, followed by centrifugation at 4,500 x g. Supernatants were discarded and pellets washed by applying 100 µl of 60% isopropanol to wells, and the plate centrifuged at 4,500 x g for 10 minutes. Supernatants were discarded and the plate centrifuged at 150 x g upside-down onto tissue paper for 1 minute to remove any large remaining droplets of supernatant. Sequencing plates were placed at 37°C for 30

minutes to fully dry and purified reactions resuspended in 10 µl of HiDi Formamide (Applied Biosystems). Sequencing fragments were separated and detected using ABI3130xl genetic analysers using the parameters stated in Appendix 3 and data analysed using Staden Gap software version 4 (Bonfield et al., 1995).

2.10. RNA extraction

RNA was extracted from skin, footpad, cerebellum tissue, and buccal cells using the Qiagen RNeasy Mini kit. Up to 20 mg of tissue were added to 600 µl of RLT buffer (Qiagen) containing 1% beta-mercaptoethanol (Sigma-Aldrich) and disrupted using a mortar and pestle. Complete homogenisation of samples was achieved using QIAshredder columns (Qiagen). Tissue lysates were centrifuged for 3 minutes at 17,000 x g and supernatants transferred to new microcentrifuge tubes. To the lysates 600 µl of 70% ethanol were added, which were mixed immediately by pipetting. RNeasy columns were loaded with 600 µl of the extracts and centrifuged for 15 seconds at 8,000 x g. Flow-through was discarded and a further 600 µl of each sample were applied and the columns re-centrifuged. Columns were washed by applying 700 µl of buffer RW1 (Qiagen) and centrifuging at 8,000 x g for 15 seconds, discarding the flow-through. Two further washes using 500 µl of buffer RPE (Qiagen) were performed using the same method. Columns were placed on a fresh waste collection tube and centrifuged at 17,000 x g for 3 minutes to completely dry the column membranes. The RNA on the columns was eluted using 50 µl of RNase-free water, applied directly onto the silica membranes, and columns centrifuged for 1 minute at 8,000 x g. The RNA-containing flow-through was passed through the columns a second time to increase the final yield and concentration.

2.11. Reverse transcription

A mixture of 1 µl 500 µg/ml Oligo $(dT)_{15}$ (Promega), 10-200 ng total RNA, 1 µl of 10 mM dNTP mix (Amersham) and ultapure water to a total volume of 12 µl were incubated on a thermal cycler at 65°C for 5 minutes and chilled on ice. To the mix 2 µl of 5x first strand buffer (Invitrogen), 2 µl of 0.1 mM DTT (Invitrogen) and 1 µl of RNasin (Promega) were added, gently mixed by pipetting and heated at 42°C for 2 minutes. To start the reaction 1 µl of SuperScript II Reverse Transcriptase (200 U/µL) (Invitrogen) was added and mixed by pipetting. The contents were incubated at 42°C for 50 minutes, followed by inactivation at 72°C for 15 minutes. To remove RNA complementary to the cDNA 1 µl of RNase H (Invitrogen) was added to the mixture, which was incubated at 37°C for 20 minutes. Targeted regions of the cDNA were amplified using the standard PCR protocol.

2.12. Rapid amplification of cDNA ends (RACE)

Amplification of 5' ends was performed using the Roche 5' RACE kit, 2nd generation. First strand cDNA synthesis was performed using a gene specific primer (GSP), followed by degradation of complementary RNA by RNaseH activity. The cDNA was purified using the High Pure PCR purification kit (Roche) according to the manufacturer's instructions. The cDNA was then tailed with dATP and TdT with a recombinant terminal transferase, allowing PCR with an oligo(dT) anchor primer and a nested GSP. A second nested PCR was performed using an anchor primer and an internal GSP. PCR products were Sanger sequenced. 5' RACE reactions were carried out by Louisa Hayward. Amplification of 3' ends was carried out using the Clontech SMARTer RACE cDNA amplification kit. First strand synthesis was carried out using 3'-RACE CDS Primer A (Tailed and anchored oligo $(dT)_{30}$). The 3' ends were amplified using 2 µl RACE ready cDNA, using a standard PCR protocol with a GSP and the SMARTer universal primer mix.

2.13. Episodic falling candidate gene selection

Candidate genes were selected by *in silico* searching of the literature for diseases belonging to the muscle hypertonicity category using the PubMed scientific literature searching facility (PubMeb National Center for Biotechnology Information, U.S., 2010) and Online Mendelian Inheritance in Man. Genes were identified as good candidates based on their association with muscle hypertonicity or due to their gene function. A list of candidate genes is shown in Table 2.1.

Table 2.1 Candidate genes for EF

Gene	Human Condition or Gene Function
ATP2A1	Recessive Brody myopathy
CACNL1A3	Dominant hypokalemic periodic paralysis
CLCN1 (CIC1)	Myotonia congenita
GLRA1	Hyperekplexia
GLRA3	Glycine receptor subunit alpha-3 precursor
GLRB	Hyperekplexia
GPHN	Hyperekplexia
KCNE3	Thyrotoxic hypokalemic periodic paralysis
SCN4A	Non-dystrophic myotonia
SLC32A1	Vesicular inhibitory amino acid transporter (GABA and glycine transporter)
SLC6A5	Hyperekplexia (sodium- and chloride-dependent glycine transporter 2 (GlyT2))
SLC6A9	Sodium- and chloride-dependent glycine transporter 1 (GlyT1)

2.14. Microsatellite genotyping

2.14.1. Microsatellite identification and primer design

Canine orthologues of the human genes were identified using the Ensembl genome browser (www.ensembl.org). Dinucleotide repeat microsatellite markers were searched for in the 400 kb of sequence surrounding each gene. Primers for microsatellite amplification were designed using the online tool Primer3 (http://frodo.wi.mit.edu/) (Rozen and Skaletsky, 2000). At the 5' end of the forward primers, an 18 bp tail sequence (TGACCGGCAGCAAAATTG) was added to allow the amplification of a 5' fluorescently labelled "third primer" of the same 18 bp sequence for visualisation of products on ABI3130xl genetic analysers (Oetting et al., 1995).

2.14.2. PCR amplification of microsatellites using tailed primers

Microsatellites were amplified by PCR. Reactions consisted of 0.2 mM dNTPs (NEB), 1.5 mM $MgCl_2$ (Applied Biosystems), 1x PCR Gold buffer (Applied Biosystems), 0.17 µM forward primer with 18 bp extension, 0.42 µM reverse primer, 0.5 µM fluorescently labelled (6FAM, VIC, NED or PET) third primer, 1.2 units AmpliTaq Gold DNA polymerase (Applied Biosystems) and ultrapure water to a volume of 12 µl. Cycling parameters are listed in Appendix 4.

2.14.3. Genotyping by capillary electrophoresis

Genotyping of microsatellite markers was performed by capillary electrophoresis. On a non skirted PCR plate, 1 µl of PCR product was combined with 10 µl HiDi formamide (Applied Biosystems) containing 5% Rox 400HD size standard (Applied Biosystems). Double stranded products were denatured by heating at 95°C for 1 minute before chilling on ice. Plates were run on ABI3130xl genetic analysers using parameters stated in Appendix 3 and results visualised and analysed using ABI Genemapper software.

2.15. Genome scanning

2.15.1. Homozygosity mapping

Genome scanning for the IS spinocerebellar ataxia study was performed by homozygosity mapping using 6 cases and 6 obligate carriers (parents of affected individuals). The genome-wide set of markers used for homozygosity mapping consisted of ~300 genome-wide microsatellites arranged in multiplex panels of between 3 and 10 markers for PCR. Forward primers were fluorescently tagged to allow separation and visualisation of products by capillary electrophoresis. PCRs were performed in 17 µl volumes consisting of 1x PCR Gold buffer (Applied Biosystems), 2.5 mM $MgCl_2$ (Applied Biosystems),

0.15 mM dNTPs (NEB), 1.75% DMSO (Sigma Aldrich), 0.65 units AmpliTaq Gold DNA polymerase (Applied Biosystems) and 10-20 ng of genomic DNA. Cycling parameters are listed in Appendix 4. Products were separated by capillary electrophoresis on ABI3130xl genetic analysers and genotypes analysed and scored using Genemapper v4.0.

2.15.2. Fine mapping

Fine mapping was performed using microsatellite markers surrounding the boundaries of the disease-associated haplotype. Microsatellites were identified using the Ensembl genome browser and genotyped as described in section 2.14. Additional fine mapping was performed by SNP analysis at the boundaries of disease-associated regions. The disease-associated genomic regions were interrogated for SNPs using the BioMart database in Ensembl. Genotyping of SNPs was performed by PCR followed by Sanger sequencing.

2.15.3. Linkage analysis

Linkage analysis was used to confirm homozygosity mapping results. An extended IS pedigree was drawn in Cyrillic and later Progeny software packages. Two-point linkage analysis was performed using MLINK (part of the LINKAGE software package) (Lathrop and Lalouel, 1984).

2.15.4. Genome-wide SNP analysis

A total of 96 DNA samples were prepared for the first batch of genotyping on the Illumina CanineHD SNP array. The DNA samples were from 31 EF cases, 19 CKCSID cases and 39 CKCS controls. Four additional CKCSs with a CKCSID-like phenotype were included (2 CKCSs with a dry eye phenotype and 2 CKCSs with rough coat phenotype). One previously genotyped Golden Retriever DNA was included as a genotyping control and two Italian Spinoni were genotyped for fine mapping and copy number variation analysis. The second batch of samples for genotyping on the Illumina CanineHD SNP array consisted of 16 cases and 16 controls for the Parson Russell Terrier ataxia project.

The concentration of the DNA samples was initially measured using a Nanodrop spectrophotometer. Samples with concentrations lower than 50 ng/µl were concentrated using Multiscreen PCR purification plates as described in section 2.8. Samples with a 260/280 or 260/230 of lower than 1.70 were selected for re-extraction using the Nucleon procedure (Section 2.3.1.). Final quantification of DNA was performed using PicoGreen (Section 2.4.).

DNA samples were sent to the Cambridge Genome Service, Department of Pathology, University of Cambridge, for processing on the Illumina CanineHD SNP genotyping array, comprising 173,662 genome-wide SNPs.

2.15.4.1. Raw SNP genotyping data quality control and handling

Raw SNP genotyping data were visualised in the Genome Studio software package (Illumina). SNPs were re-clustered to improve call rates. SNPs were then sorted by cluster separation, call frequency, mean normalised intensity of the heterozygous cluster and normalised theta value (relative location) of the heterozygous cluster. Cluster boundaries were then adjusted to improve the call rate for the SNP or if the clusters were poorly defined the SNP was zeroed. Quality checked and edited SNP genotyping data were exported from Genome Studio and imported into the pedigree drawing programme Progeny. In Progeny cases and control sets were assigned and SNP genotyping data exported in a format accepted by the statistical package PLINK (Purcell et al., 2007).

2.15.4.2. SNP analysis

Allelic association analysis of SNP genotyping data were performed in the statistical package PLINK. Binary files were initially made to define the SNP set. Samples with a genotyping call rate of less than 90% (mind 0.1) were excluded. Any SNPs with a genotyping call rate of less than 95% (geno 0.05) and a minor allele frequency of less than 5% (maf 0.05) were removed. Binary file preparation was performed by implementing the following command in Linux:

```
> plink --noweb --dog --allow-no-sex --file (file name) --make-founders
--make-bed --mind 0.1 --maf 0.05 --geno 0.05 -- out (name)
```

Allelic association analysis was performed using the following command:

```
> plink --noweb --dog --allow-no-sex --bfile (binary file name) --assoc -
- out (name)
```

Allelic association analyses were corrected for multiple testing using the Max(T) permutations procedure, with the number of permutations set to 100,000 using the command --mperm 100000.

Quantile-quantile (QQ) plots, of observed versus expected probability values, were constructed to assess datasets for genomic inflation. A value for genomic inflation was also obtained by including the --adjust command. Data points were plotted in Excel.

```
> plink --noweb --dog --allow-no-sex --bfile (binary file name) --assoc --adjust --qq-plot --log10 --out (name)
```

Multidimensional scaling (MDS) plots were constructed to assess the relatedness of individuals used in GWAS. Data points, which were subsequently plotted on a graph using Excel, were calculated using the following command.

```
> plink --noweb --dog --allow-no-sex --bfile (binary file name) --cluster --mds-plot 2 --out (name)
```

Data sets with high genomic inflation values, as indicated by QQ plotting and separated case and control clusters in MDS plots were adjusted for population stratification using the Fast Mixed Model (FMM) (Astle and Balding, 2009). Files in PLINK format were converted to FMM format using a Perl script written by Dr Mike Boursnell, which was implemented in Linux. The FMM was then executed in the 'R' package for Linux.

2.16. Massively parallel sequencing

2.16.1. Target enrichment by long range PCR

The template for sequencing the entire *ITPR1* gene was generated by long range PCR, using a high fidelity DNA polymerase. Reactions consisted of 0.2 mM dNTP mix (ABgene), 1x HF PCR buffer (or 1x GC buffer when amplifying GC rich sequences), 0.5 µM forward primer, 0.5 µM reverse primer, 1 unit Phusion hot start polymerase (Finnzymes), 20 ng genomic DNA and ultrapure water to a final volume of 50 µl. Cycling parameters are listed in Appendix 4. Genomic regions which failed to amplify with the high fidelity polymerase, such as AT or GC rich regions were amplified using standard PCR.

A 10% sample of each PCR product was analysed using 1.5% agarose gel electrophoresis, to verify product size and assess the reaction specificity. PCR products were purified using Multiscreen PCR purification plates (section 2.8) and DNA quantified using a Nanodrop spectrophotometer. Purified PCR products were run on a 1% agarose gel as a final check of concentration and for artefact band consideration. The 60 PCR products forming the contiguous template for sequencing were combined to form an equimolar mixture which was Nanodrop quantified and 1,320 ng (40 µl at 33 ng/µl) were sent to Fasteris Life Science for library preparation.

2.16.2. DNA fragmentation methods for sequencing library preparation

Sonication, nebulisation, enzymatic and Covaris methods of DNA fragmentation were investigated for sequencing library preparation, and results analysed by running DNA on a 1% TAE agarose gel.

2.16.2.1. Sonication

Genomic DNA was sonicated using an adaption of the method described by J. Sambrook (Sambrook, 2001). Sonication was performed using 100 µl of DNA (50 ng/µl) at full amplitude, for either 10, 15 or 20 cycles of 20 seconds on followed by 20 seconds off, using a Misonix XL 2020 Ultrasonic Liquid Processor (cup horn arrangement). The tubes containing the DNA were held in an ice bath during sonication. The sonicator was tuned according to the manufacturer's instructions before use.

2.16.2.2. Nebulisation

Nebulisers (Invitrogen) were assembled by inserting the provided vinyl tubing over the central atomiser, as illustrated in Figure 2.1. The DNA solution was made by adding 5 µg of stock DNA in 50 µl TE to 700 µl nebulisation buffer (TE containing 50% glycerol). The DNA solution was mixed and placed in the nebuliser and the lid screwed on, before placing on ice to chill. Using compressed nitrogen, a pressure of 24 psi was applied to the unit for 10 minutes. The DNA solution was recovered from the sides of the nebuliser by centrifugation at 450 x g. Recovered DNA was purified using a Qiagen QIAquick PCR purification column, with an elution volume of 30 µl.

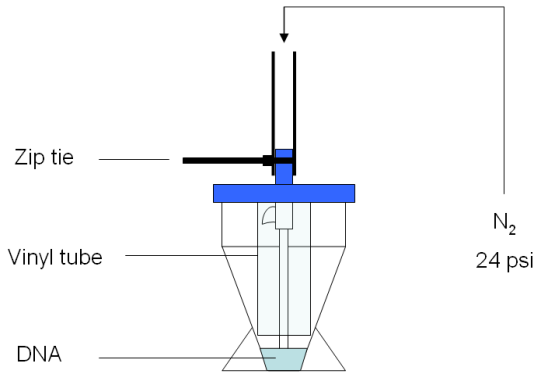


Figure 2.1 Arrangement of apparatus for nebulisation of DNA

As pressure is applied the solution is forced up the tubing and out of the central atomiser, shearing the DNA.

2.16.2.3. Enzymatic fragmentation

This approach used DNA Fragmentase (New England Biolabs) to fragment genomic DNA. Genomic DNA (5 µg) was combined with 10 µl Fragmentase reaction buffer, 1 µl 100x BSA (10 mg/ml) and ultrapure water to a final volume of 90 µl. The mix was vortexed

thoroughly and incubated on ice for 5 minutes. To the reaction mix 10 µl dsDNA Fragmentase enzyme solution was added and vortexed thoroughly, before incubating at 37℃ for 30 minutes in a water bath. The reaction was stopped by the addition of 5 µl of 0.5 M EDTA. Digested DNA was purified using a Qiagen QIAquick PCR purification column, with an elution volume of 30 µl.

2.16.2.4. Covaris shearing

Covaris shearing of DNA using adaptive focused acoustics was performed by the Eastern Sequence and Informatics Hub. Settings for Covaris shearing are shown in Table 2.2.

Table 2.2 Covaris settings for DNA fragmentation

Setting	Value
Duty cycle	10%
Intensity	5
Cycles or Burst	200
Time	6 cycles of 60 seconds each
Temperature	4℃

2.16.3. End repair of fragmented DNA

End repair was performed using the NEBnext end repair module, by combining 30 µl of fragmented DNA with 10 µl end repair buffer, 5 µl NEBnext repair enzyme mix and ultrapure water to a final volume of 100 µl. Reactions were incubated on a thermal cycler at 20℃ for 30 minutes. For repair of DNA fragmented using NEB DNA Fragmentase, 1 µl of *E.coli* DNA ligase was also added to the mix. Repaired DNA was purified using a Qiagen QIAquick PCR purification kit, eluting in 37 µl of buffer EB.

2.16.4. dA tailing of repaired DNA fragments

Fragment dA tailing was performed using the NEBnext dA tailing module, by combining 37 µl end repaired blunt DNA, 5 µl NEBnext dA-tailing reaction buffer (10x), 3 µl Klenow fragment (5,000 units/ml) (3' to 5' exo-), and 5 µl ultrapure water. Reactions were incubated on a thermal cycler for 30 minutes at 37℃. DNA was purified using a Qiagen QIAquick PCR purification kit, eluting in 25 µl of buffer EB.

2.16.5. Adapter ligation to dA tailed DNA library

DNA adapters for the multiplexed pair end module were made by annealing the Illumina adapter oligonucleotides, which were synthesised and HPLC purified by Integrated DNA Technologies. In a 96 well microplate 15 µl of each adapter oligonucleotide (100 µM) were mixed and heated on a thermal cycler at 95℃ for 2 minutes, followed by cooling at 0.1° per second to 25℃. The double stranded adapter was then snap cooled on ice and

diluted to 15 µM by the addition of 70 µl TE buffer. Adapters were stored at -20℃ and thawed on ice when used.

Adapter ligation was performed using the NEBnext quick ligation module, by combining 25 µl of dA tailed DNA, 10 µl quick ligation reaction buffer (5x), 10 µl of 15 µM DNA adapters and 5 µl quick T4 DNA ligase (400,000 units/ml), and incubating on a thermal cycler at 20℃ for 15 minutes. DNA was purified using a Qiagen QIAquick PCR purification kit, eluting in 15 µl of buffer EB.

2.16.6. Gel size selection of adapter ligated library

To the adapted fragments 2 µl of gel loading buffer were added and the entire sample run on a 2% Microsieve Low Melt agarose gel (Flowgen) containing 0.3 µg/ml ethidium bromide, alongside a 100 bp ladder (NEB). Fragments in the 200 to 300 bp range were excised from the gel and purified using the QIAquick Gel extraction kit, eluting in 32 µl of buffer EB.

2.16.7. Amplification of adapted DNA library by PCR

Size selected fragments were amplified using Finnzymes Phusion DNA polymerase (Thermo-Fisher). Reactions contained 3 to 25 ng of template DNA, 1x Phusion mastermix, Illumina primer InPE 1.0 and Illumina index primer at a final concentration of 0.5 µM, and Illumina primer InPE 2.0 at a final concentration of 0.01 µM. Reactions were made up to a final volume of 50 µl with ultrapure water. Cycling parameters are listed in Appendix 4. PCR products were purified on a QIAquick PCR purification column and DNA eluted in 50 µl of EB.

2.16.8. SureSelect target enrichment

The SureSelect target enrichment system was used to enrich DNA sequencing libraries for genomic regions of interest, in a process of capture by hybridisation using custom RNA baits. Two enrichment experiments were carried out. Slightly different methods were adopted due to a change in kit reagents and the introduction of a Covaris service at the Eastern Sequencing and Informatics Hub, both of which became available for the second target enrichment experiment.

2.16.8.1. RNA bait design for SureSelect target enrichment

RNA baits were designed for the SureSelect target enrichment system using the online tool, e-array (https://earray.chem.agilent.com/earray/). Tiling of baits was set to 2x, and repeat sequences were masked during the design process.

2.16.8.2. SureSelect target enrichment library preparation

For the first SureSelect experiment genomic DNA was fragmented using the enzymatic procedure outlined in section 2.16.2.3. Library end repair, dA tailing and adapter ligation reactions were carried out as detailed in sections 2.16.3. to 2.16.5. Size selection was performed as outlined in 2.16.6. The second SureSelect experiment used an enzymatic approach for fragmenting DNA from whole blood samples and Covaris shearing for fragmenting DNA from buccal swabs. End repair, dA tailing and adapter ligation reactions were performed using additional reagents added to the SureSelect kit, comparable to those found the NEBnext library preparation kit.

2.16.8.3. Pre-hybridisation amplification of libraries

Adapter ligated libraries were subjected to pre-hybridisation amplification reactions for use in the SureSelect procedure. Reactions were performed in 50 µl volumes consisting of 15 µl adapter ligated library, 21 µl nuclease-free water, 1.25 µl InPE primer 1.0 (25 µM), 1.25 µl SureSelect Indexing Pre-Capture PCR primer (25 µM), 10 µl Herculase 5x reaction buffer (Agilent), 0.5 µl dNTP mix (provided with Herculase polymerase) and 1 µl Herculase II polymerase (concentration not specified by manufacturer) (Agilent). Thermal cycling parameters are listed in Appendix 4. Reactions were purified using AMPure XP beads.

2.16.8.4. Agencourt AMPure XP bead DNA clean-up

At room temperature AMPure beads (Beckman Coulter) were homogenised by vigorous shaking of the stock bottle, before combining 90 µl with a 50 µl reaction mix in a 1.5 ml microtube. Tubes were then mixed and incubated at room temperature for 5 minutes, before placing on a magnetic stand to separate. The cleared solution was discarded and beads washed twice with 70% ethanol, allowing the beads to settle for 2 minutes before ethanol removal. Bead containing tubes were then removed from the magnetic stand and dried for 5 minutes at 37°C, before addition of 10 µl nuclease free water. Tubes were vortex mixed and placed back on the magnetic stand for bead separation and removal of the supernatant containing the purified DNA.

2.16.8.5. Assessment of pre-hybridisation libraries

Pre-capture libraries were quantified using a Qubit fluorometer, normalised to 147 ng/µl, and checked for expected size range by 2% agarose gel electrophoresis. One library was also outsourced to Cambridge Genomic Services for analysis on the Agilent Bioanalyser DNA 12000 Series II chip, for size checking purposes.

2.16.8.6. Hybridisation

SureSelect hybridisation buffer was assembled according to the manufacturer instructions and heated to 65°C for 5 minutes to prevent precipitate formation. In a 1.5 ml microtube

5 µl of SureSelect bait library and 2 µl of 25% RNase block were combined and placed on ice. To a 96 well PCR plate (row B) 3.4 µl of DNA library at a concentration of 147 ng/µl were added, combined with 5.6 µl SureSelect block mix, heated at 95°C for 5 minutes and then held at 65°C on a thermal cycler with the heated lid set to 105°C. Whilst holding the temperature at 60°C, 40 µl of hybridisation buffer were added to row A, the plate resealed and temperature held at 65°C for a further 5 minutes. To row C 7 µl of bait library were added, the plate resealed and heated for a further two minutes. Maintaining the temperature at 65°C, 13 µl of hybridisation mix were swiftly taken from row A and placed in row C, mixing by pipette. Without delay, the entire 8 µl of the DNA library mix was transferred to row C, mixed by pipetting and the plate sealed with two layers of thermal sealing film. The hybridisation mixture was incubated for 24 hours at 65°C.

2.16.8.7. Streptavidin bead capture

For each capture a 50 µl volume of Dynal magnetic streptavidin bead solution (Invitrogen) was added to a 1.5 ml microtube. Beads were washed by adding 200 µl SureSelect binding buffer and vortexing tubes for 5 seconds. Tubes were placed on a magnetic rack and beads allowed to settle before removing and discarding the supernatant. Beads were washed 3 times, before being resuspended in 200 µl of SureSelect binding buffer.

After the 24 hour incubation period at 65°C, the hybridisation mixture was pipetted directly off the plate and into the streptavidin bead solution, inverting 5 times to mix. Tubes were placed on a rotary mixer held at a 45 degree angle to mix for 45 minutes at room temperature, and pulse centrifuged. Beads were separated by placing tubes on a magnetic rack and the supernatant removed. Beads were washed using 500 µl SureSelect wash buffer #1, incubating at room temperature for 15 minutes with occasional vortexing. After pulse centrifuging, tubes were placed on a magnetic rack to separate the beads and the supernatant was removed. The second washing stage consisted of bead resuspension in 500 µl SureSelect wash buffer #2 (prewarmed to 65°C) and placing tubes on a heat block at 65°C for 15 minutes with occasional vortexing. Beads were separated on a magnetic rack and all the supernatant was carefully removed. The second washing stage was repeated twice. A 50 µl volume SureSelect elution buffer was then added to the beads, which were vortex mixed and incubated at room temperature for 5 minutes before separating the beads with the magnetic rack and placing supernatant (captured DNA) in a separate tube. Eluted DNA was combined with 50 µl of SureSelect neutralisation buffer and purified using AMPure beads.

2.16.8.8. Post-capture amplification

In a PCR plate 14 µl captured DNA, 22.5 µl nuclease-free water, 5 µl Herculase II reaction buffer, 0.5 µl dNTP mix, 1 µl Herculase II Fusion DNA polymerase (Agilent), 1 µl SureSelect Indexing Post-Capture PCR primer (25 µM), and 1 µl Index PCR reverse primer (25 µM) were combined. Thermal cycling parameters are listed in Appendix 4. Reactions were purified using AMPure beads, with a final elution volume of 15 µl.

2.16.9. mRNA-seq library preparation

2.16.9.1. RNA extraction

RNA was extracted from 40 mg of canine cerebellum preserved in RNAlater solution (Life Technologies) using the Qiagen RNeasy midi kit. Cerebellum tissue was disrupted in 2 ml RLT buffer (stabilised by the addition of 40 µl 2 M dithiothreitol (DTT)) using a mortar and pestle. Lysates were transferred to a Qiashredder column, and centrifuged at 17,000 x g for 2 minutes to fully homogenise. Lysates were then transferred to a 15 ml tube and centrifuged at 4,500 x g for 10 minutes to pellet the cell debris, and the supernatant was transferred to a clean 15 ml tube. One volume of 70% ethanol (~2 ml) was added to the supernatant, which was mixed by inversion, applied to an RNeasy midi column and centrifuged for 5 minutes at 4,500 x g. After discarding the flow-through, 2 ml RW1 were added to the column, which was centrifuged at 4,500 x g for 5 minutes, discarding the flow-through. The RNA was DNase treated on-column by applying 160 µl DNase (20 µl DNase stock (60 Kunitz units), 140 µl buffer RDD) to the column filter and incubating at room temperature for 15 minutes. A further 2 ml RW1 were added to the column, which was centrifuged at 4,500 x g for 5 minutes, discarding the flow-through. The column was then washed twice using 2.5 ml RPE buffer, centrifuging at 4,500 x g for 2 minutes for the first wash and at 4,500 x g for 5 minutes for the second wash, discarding the flow-through between washes. The column was then placed on a clean 15 ml collection tube and 150 µl RNase free water added directly to the column filter and incubated at room temperature for 5 minutes. The RNA was eluted by centrifuging the column at 4,500 x g for 3 minutes.

2.16.9.2. mRNA isolation

Isolation of mRNA was performed using Sera-mag oligo(dT) magnetic beads (Thermo-Fisher). Total RNA (4.9 µg) was diluted to a 50 µl volume with nuclease-free water. 15 µl of Sera-mag beads were placed in a 1.5 ml tube, washed twice in 100 µl of hybridisation buffer (Thermo-Fisher) and resuspended in 50 µl of hybridisation buffer. All washing stages were performed by placing bead containing tubes onto a magnetic rack, removing supernatant, and adding the buffer before vortexing tubes for 5 seconds. Tubes were then placed back on the rack to allow beads to settle, before removing the supernatant and proceeding to the next stage. Tubes containing total RNA were placed on a heat block at

65℃ for 5 minutes to disrupt secondary structures and chilled on ice before transferring the total RNA to the resuspended washed beads. The bead solutions were rotary mixed for 5 minutes at room temperature to hybridise, before washing beads twice in 200 µl wash buffer (Thermo-Fisher) and resuspending in 50 µl elution buffer (Thermo-Fisher). The mRNA was eluted from the beads by placing in a heat block at 80℃ for 2 minutes, before immediately placing the tubes back on the magnetic rack and transferring the supernatants to tubes containing 50 µl of hybridisation buffer. A second enrichment stage was performed to further increase the mRNA:rRNA ratio. Beads from the first enrichment stage were washed twice in 200 µl wash buffer ready for re-use. Eluted mRNA was heated for 2 minutes at 65℃, and then quickly chilled on ice and transferred to the washed beads. The bead solutions were rotary mixed for 5 minutes at room temperature to hybridise, before washing the beads twice in 200 µl wash buffer and resuspending the beads in 18 µl elution buffer. The mRNA was eluted from the beads by placing in a heat block at 80℃ for 2 minutes, before immediately placing the tubes back on the magnetic rack and transferring the supernatants to clean tubes.

2.16.9.3. mRNA fragmentation

Libraries for mRNA-seq were prepared using the NEBnext Library Prep Master Mix Set for Illumina. The messenger RNA was fragmented by adding 18 µl purified mRNA to 2 µl RNA fragmentation buffer in a 96 well plate and heated in a thermal cycler for 5 minutes at 94℃. The plate was then chilled on ice and 2 µl of 10x RNA fragmentation stop solution added to reactions. Fragmented mRNA was cleaned-up using an RNeasy mini column. Reaction mixes were transferred to a 1.5 ml tube and adjusted to 100 µl by adding 78 µl nuclease free water. To the diluted RNA, 350 µl of buffer RLT were added and mixed by vortexing, followed by addition of 250 µl of absolute ethanol, before mixing by inversion. The RNA mixture was applied to an RNeasy mini column, which was washed and RNA eluted according to section 2.10. The elution volume was 20 µl.

2.16.9.4. First strand cDNA synthesis

In a 96 well plate 13.5 µl fragmented mRNA and 1 µl random primers were combined. Plates were sealed and incubated at 65℃ for 5 minutes before chilling on ice. To the reaction mix 4 µl first strand synthesis buffer (Invitrogen) and 0.5 µl murine RNase inhibitor (NEB) were added, mixed and incubated at 25℃ for two minutes. To the reaction 1 µl SuperScript II Reverse Transcriptase (200 U/µL) (Invitrogen) was added and incubated at 25℃ for 10 minutes, 42℃ for 50 minutes, 70℃ for 15 minutes, and then held at 4℃.

2.16.9.5. Second strand cDNA synthesis

To the first strand synthesis reaction, 48 µl of nuclease free water, 8 µl second strand synthesis buffer (NEB) and 4 µl second strand enzyme mix (NEB) were added and mixed by pipetting. Reactions were incubated at 16℃ for 2.5 hours before purifying cDNA using AMPure beads (Section 2.16.8.4.).

2.16.9.6. End repair, dA tailing and adapter ligation

Library end repair, dA tailing and adapter ligation were performed as described in sections 2.16.3 to 2.16.5

2.16.9.7. Size selection and library amplification

Adapter ligated libraries were separated on a 3% 1:1 low melt:standard agarose gel and bands excised at the 200-250 bp (main library) and at 300-350 bp (reserve library) position using an X-tracta gel extraction tool (Sigma Aldrich). Library DNA fragments were purified from excised bands using the QIAquick gel extraction protocol (section 2.16.6). Size selected libraries were amplified for multiplexed sequencing as described in section 2.16.7 and purified using AMPure beads as described in section 2.16.8.4.

2.16.10. Quantification of sequencing libraries

The KAPA library quantification kit was used to accurately quantify the libraries for Illumina sequencing. Libraries were diluted 1:100,000 with nuclease free water to put the concentrations within the range of the DNA standards (0.0002 - 20 pM). Reaction mixes containing 4 µl diluted library or standard, 4 µl nuclease free water and 12 µl KAPA SYBR Fast qPCR master mix (containing primer premix), were run on an Illumina Eco qPCR machine in triplicate with the cycling parameters defined in Appendix 4. Based on the threshold cycle (Ct) values of the standards, a standard curve was drawn allowing reaction efficiency and r^2 values to be calculated. The concentration of the libraries was calculated using Ct values and the standard curve equation. Stock concentrations were calculated by first adjusting for the size of the library (452 / expected fragment length) and then multiplying by the dilution factor.

2.16.11. Blunt end cloning

Blunt end cloning was used to check that libraries contained the expected fragment sequences using the pT7Blue Perfectly Blunt Cloning Kit (Novagen). Library fragments were phosphorylated prior to ligating into the plasmid using the end conversion enzyme mix provided. In a 96 well microplate 2 µl of sequencing library, 3 µl nuclease-free water and 5 µl end conversion mix were combined, mixed with a tip and incubated on a thermal cycler at 22℃ for 15 minutes. The reaction was inactivated by heating at 75℃ for 5 minutes, and chilled on ice. For the ligation, 1 µl (50 ng) PT7Blue Blunt Vector and 1 µl T4

DNA Ligase (4 Weiss units/µl) were added to the reaction and mixed with the end of a tip. The reaction mix was incubated at 22℃ for 15 minutes.

The cells used for transformations were NovaBlue Singles™ Competent Cells. Competent cells were removed from -80℃ storage and placed on ice to thaw. Cells were evenly resuspended by gently flicking tubes 3 to 4 times. To the cells 1 µl of ligation reaction was added which were then stirred gently to mix and returned to ice for 5 minutes. Heatshock was performed by heating the tubes for exactly 30 seconds in a 42℃ water bath, and returning to ice for 2 minutes. To the tubes 250 µl of room temperature super optimal broth with catabolite repression (SOC) medium were then added.

Lysogeny broth (LB) agar plates were made containing 50 µg/ml ampicillin, 12.5 µg/ml tetracycline, 50 µg/ml X-gal and 50 µM IPTG (see Appendix 1). Glass Pasteur pipettes were sterilised and shaped into spreading tools by heating, and used to evenly spread 50 µl of the cell mixture over plates. Plates were inverted and incubated overnight at 37℃. Once colonies reached a 0.5 to 1 mm diameter, plates were placed at 4℃ to further develop the colour for blue/white colony screening.

2.16.12. Direct colony PCR

White colonies were picked using a sterile cocktail stick and deposited into a microcentrifuge tube containing 50 µl of TE buffer by twirling. Tubes were vortexed thoroughly, and pulse centrifuged to collect the cell solution, which was then transferred to a single well of a 96 well plate. The plate was sealed and heated at 95℃ for 5 minutes to lyse the cells and chilled on ice. The plate was then centrifuged at 4500 x g to pellet the debris and 2 µl of supernatant used for standard PCR followed by Sanger sequencing. Primers for PCR in the 5' to 3' direction were TGCAGGTCGACTCTAGAGGAT and GTTTTCCCAGTCACGACGTT, with an expected product size of 105 bp for an uncut plasmid.

2.16.13. Illumina sequencing (outsourced)

The sequencing template generated by long range PCR was sequenced, after library preparation, at Fasteris Life Science. Single-end sequencing was performed to generate reads of 33 bp.

SureSelect indexed libraries were pooled to a final concentration of 10 nM, and sent for sequencing on an Illumina GAIIx or Illumina HiSeq 2000 at the Wellcome Trust Centre for Human Genetics, University of Oxford. Paired-end sequencing with reads of 51 bp was performed.

2.16.13.1. Illumina MiSeq sequencing

Sequencing on the Illumina MiSeq platform was performed in-house. Stock libraries were diluted to 2 nm with Tris-Cl 10mM, pH 8.5 containing 0.1% Tween 20. Libraries were denatured by combining and mixing 10 µl freshly made 0.2 N NaOH solution with 10 ul of 2 nM library, and incubating for exactly 5 minutes. To the 20 µl of denatured library, 980 µl of prechilled HT1 were added to give a library concentration of 20 pM. The library was further diluted by combining 500 µl of library with 500 µl HT1 to give a library concentration of 10 pM. The MiSeq reagent cartridge was prepared by thawing in a room temperature waterbath for one hour before inverting 10 times to mix reagents. The sample well was then pierced with a clean pipette tip and a 600 µl volume of 10 pM library loaded into the cartridge, before inserting into the sequencer. The flow cell was removed from the storage solution before thoroughly flushing with ultra-pure water and drying completely with a lint-free tissue, ensuring surfaces were streak free. The flow cell, PR2 bottle and emptied waste bottle were installed onto the machine, before linking the sample sheet and starting the sequencing run. On run completion a post-run wash was initiated by replacing the reagents cartridge with an ultrapure water filled wash cartridge, and the PR2 bottle with an ultrapure water filled wash bottle.

2.16.14. Illumina sequencing data analysis

The dataset generated from the long range PCR template was analysed using Maq assembly software (Li et al., 2008), with the canine genome sequence CanFam2 as a reference. *De novo* assembly was performed at Fasteris Life Science using Edena software (Hernandez et al., 2008). Staden gap software (Bonfield et al., 1995) was used to align both Edena and Maq assemblies and capillary reads from exon sequencing, for comparison with the CanFam2 canine genome build.

For the SureSelect sequencing datasets a Perl script entitled “NGS analysis”, executable from the Linux command prompt, was written to handle data in FASTQ format outputted from Illumina sequencing experiments. Key commands of the Perl script are listed in Appendix 5 and a text file of the full script found on the Supplementary CD. Reads were aligned to the canine genome using the program BWA (Burrows-Wheeler Aligner) (Li and Durbin, 2009). Aligned reads from BWA were in SAM format (Sequence Alignment/Map). The read file manipulation program Samtools (Li et al., 2009) and Picard tools (http://picard.sourceforge.net) were used to sort aligned reads, convert read files into binary format (BAM), remove PCR duplicate reads and index the aligned reads. The Genome Analysis ToolKit (GATK) was used to further improve read alignment, adjust quality scores, and make SNP and indel calls (McKenna et al., 2010). Variant calls were

annotated using genomic information available on the Ensembl website using the Variant Effects Predictor (McLaren et al., 2010). Metrics data files and histograms of GC bias and insert size were created by Picard tools. An option to run Pindel, a structural variant detection tool was also included in the Perl script (Ye et al., 2009). Sequence alignments were visualised in the Integrative Genomics Viewer (IGV) (Robinson et al., 2011).

2.16.15. *De novo* assembly of next generation sequencing reads

Reads from Illumina sequencing experiments were aligned against the dog genome and visualised in IGV. Reads aligning across the *BCAN* transcript were manually extracted in FASTA format and *de novo* assembled using the Staden Gap software package.

2.17. Quantitative PCR

Quantitative PCR used to measure gene expression was carried out on an Illumina Eco qPCR machine in 10 µl volumes, consisting of 4 µl KAPA Probe Fast, 2 µl ultra-pure water, 1x PrimeTime primer mix and 2 µl template cDNA. Cycling parameters are listed in Appendix 4. Doubling serial dilutions of cDNA across 7 points were performed to construct standard curves and to calculate reaction efficiencies. All reactions were performed in triplicate.

2.18. Western blotting

2.18.1. Protein extraction

Canine cerebellum samples (~30 mg) were homogenised in 1 ml ice cold RIPA lysis buffer (Sigma-Aldrich), containing one complete protease inhibitor cocktail tablet per 10 ml (Roche), using an electric homogeniser. Protein lysates were 4°C centrifuged for 10 minutes at 14,000 x g, and supernatants removed and stored at -20°C. Protein concentrations were measured using a Qubit fluorometer (Invitrogen).

2.18.2. Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE)

Protein extracts (6 µl at 60 ng/µl) were combined with 6 µl 2x Laemmli buffer and denatured at 98°C for 5 minutes, before chilling on ice. Protein samples were separated by denaturing 6% SDS-PAGE (National Diagnostics) at 200 V for 1 hour.

2.18.3. Coomassie staining

Coomassie staining was used to confirm the presence of protein on SDS gels. Gels were covered in Coomassie solution and heated in a microwave at 800 W for 10 seconds 3

times, shaking between each round of heating. Gels were then destained in 30% methanol/10% acetic acid solution on a plate shaker set at 100 rpm for 1 hour.

2.18.4. Blotting

Unstained separated proteins on gels were transferred to a nitrocellulose membrane. In a blotting tank frame the SDS gel and a piece of nitrocellulose paper were held between blotting paper soaked in blotting buffer. Blotting buffer soaked sponges were placed outside of the blotting paper and the frame clamped, placed in the blotting tank and 100 V applied for 1 hour to complete the transfer.

2.18.5. Primary probing

Nitrocellulose membranes were blocked for 16 hours with 5% non-fat dried milk in phosphate-buffered saline/0.1% Tween 20 (PBS-T). Blocked nitrocellulose membranes were incubated for one hour in 1:200 goat anti-SPTBN2 (Santa Cruz Biotechnology) or 1:1,000 mouse anti-ACTB (Camlab) primary antibody in blocking buffer.

2.18.6. Secondary probing

Primary probed blots were washed in PBS-T, and incubated in 1:10,000 HRP-conjugated donkey anti-goat or 1:1,000 HRP-conjugated goat anti-mouse secondary antibody in blocking buffer.

2.18.7. Detection

Immunoreactive proteins were detected using HRP-conjugate substrate kit for enhanced chemiluminescence (GE Healthcare). After rinsing 3 times in PBS-T, nitrocellulose membranes were placed protein side up on Saran wrap and surfaces covered with 5 ml of combined detection reagent. Excess reagent was poured off and the membrane placed protein side down onto a new piece of Saran wrap. Blotting paper was placed on top of the membrane to absorb excess reagent, and Saran wrap folded to form a neat flat parcel. In a dark room a high performance chemiluminescence film (GE healthcare) was placed on top of the wrapped membrane and secured in a dark frame for 2 minutes. Exposed films were developed in 1:5 x-ray developer solution (Polycon) and upon image appearance placed in non-hardening fixing solution (Photosol), before rinsing in water, drying and photographing.

2.19. DNA tests

2.19.1. Italian Spinone spinocerebellar ataxia DNA test

A multiplex PCR amplifying two microsatellite markers was used as a linkage-based diagnostic DNA test. PCRs were performed in 12 µl reactions consisting of 0.2 mM dNTP mix (ABgene), 1x GC PCR buffer (Finnzymes), 0.06 µM microsatellite A primers, 0.11 µM microsatellite B primers, 1 unit Phusion hot start polymerase (Finnzymes), 2 µl genomic DNA (concentration range: 1-100 ng/µl) and ultrapure water to a final volume of 12 µl. Primer sequences and cycling parameters are listed in Appendices 2 and 4 respectively. Fragment analysis on an ABI 3130xl was performed using 1 µl of PCR product.

2.19.2. Cavalier King Charles Spaniel DNA test

A multiplex PCR was developed to assay for both EF and CKCSID associated mutations simultaneously. The EF deletion was assayed using one forward primer 5' of the deletion, one forward primer inside the deleted region and a fluorescently labelled reverse primer 3' of the deletion. The CKCSID one base pair deletion was assayed for using a single primer pair flanking the deletion, with the forward primer fluorescently labelled. Fragments were amplified using standard PCR. Primer sequences are listed in Appendix 2. Fragment analysis on an ABI 3130xl was performed using 1 µl of PCR product.

2.19.3. Beagle neonatal cerebellar cortical degeneration DNA test

The Beagle NCCD 8 bp deletion was assayed for using a single primer pair flanking the deletion, with the forward primer fluorescently labelled. Fragments were amplified using standard PCR. Fragment analysis on an ABI 3130xl was performed using 1 µl of PCR product. Primer sequences and cycling parameters are listed in Appendices 2 and 4 respectively

2.19.4. Parson Russell Terrier late onset ataxia DNA test

The *CAPN1* SNP associated with LOA was assayed for using an allelic discrimination (TaqMan) approach. Reactions consisted of 4 µl KAPA Probe Fast mastermix, 2 µl genomic DNA (concentration range: 1-100 ng/µl), 0.2 µl of 40x primer/probe mix (Appendix 2), and 1.8 µl of ultrapure water. Cycling, detection and analysis were carried out on an ABI Step One Plus real-time PCR machine. Primer sequences and cycling parameters are listed in Appendices 2 and 4 respectively.

2.19.5. PCR amplification of GAA triplet repeat expansion

Amplification across the GAA triplet repeat expansion was performed using the Qiagen Long Range PCR Kit. PCRs consisted of 1.25 µl dNTP mix (2 mM), 2.5 µl PCR buffer,

0.4 µM forward and reverse primers, 0.2 µl Qiagen LR enzyme mix (1 unit), 2 µl genomic DNA (100 ng) and ultrapure water to a final volume of 25 µl. Primer sequences and cycling parameters are listed in Appendices 2 and 4 respectively.

Chapter

3. Investigation of two disorders in the Cavalier King Charles Spaniel

3.1. Background

The Cavalier King Charles Spaniel (CKCS) originated from the King Charles Spaniel in the early part of the 20th century, with a dedicated breed club forming in 1928. The popularity of the CKCS gradually increased and the breed was granted separate Kennel Club registration from the King Charles Spaniel in 1945 (Cunliffe, 2004). Today the CKCS is one of the UK's most popular breeds with 5,970 Kennel Club registrations in 2012 (The Kennel Club 2013).

In common with many other purebred dog populations, the CKCS suffers from a high incidence of inherited disease. Diseases affecting the CKCS are well documented in the scientific literature and breed health websites, with a notable focus on neurological disorders (Rusbridge, 2005) and mitral valve disease (Beardow and Buchanan, 1993) (http://www.cavalierhealth.org/, http://www.aboutcavalierhealth.com/). The health of the breed was brought to the attention of the public through two BBC documentaries entitled Pedigree Dogs Exposed, which were broadcast in August 2008 and February 2012 highlighting the painful neurological disorder syringomyelia.

This chapter describes the genetic investigation into two CKCS specific conditions, episodic falling (EF) and congenital keratoconjunctivitis sicca and ichthyosiform dermatosis (CKCSID).

3.1.1. Episodic falling

Episodic falling in the CKCS, also known as sudden collapse, muscle hypertonicity and hyperekplexia, is an exercise, excitement or stress induced syndrome caused by an increase in muscle tone and a temporary inability to relax the muscles. The condition was first reported in 1983, although had been observed in the breed since at least the early 1960's (Herrtage and Palmer, 1983). The onset age is usually between 3 and 7 months and inheritance is consistent with an autosomal recessive mode. The clinical signs are often variable between cases. Episodes can vary in severity and last from a few seconds to several minutes. Episodes often start with an increase in muscle tone, with bunny hopping movements (Herrtage and Palmer, 1983, Rusbridge, 2005) and/or presence of a deer stalker gait (Wright et al., 1986). The back may become arched and the head held close to the ground leading to collapse, either to the side or forwards. Legs may be held out in a rigid, extended fashion, although in some cases the dog may return to its feet within seconds of a collapse. In severe cases forelegs or hindlegs may become protracted until they are positioned over the top of the dog's head as shown in Figure 3.1 (Herrtage

and Palmer, 1983, Shelton, 2004). Dogs appear to remain fully conscious during an episode (Herrtage and Palmer, 1983).

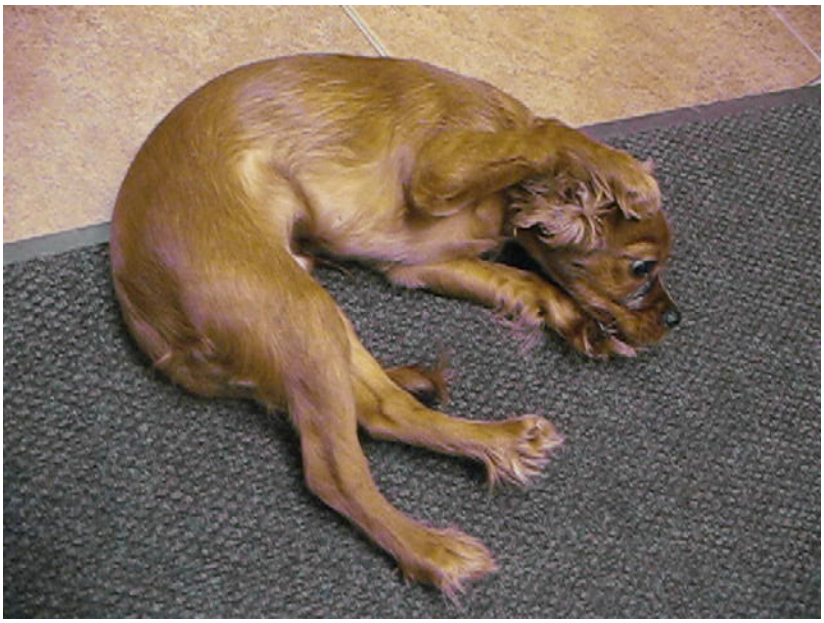


Figure 3.1 Episodic falling in a ten month old Cavalier King Charles Spaniel

Severe muscle hypertonicity in a case of EF. The dog displays an arched back and forelimbs are protracted over the head. Still image from a video provided by Dr Boaz Levitin DVM, DACVIM (Neurology).

Therapeutic agents are often used to treat EF cases, with dogs often responding well to Clonazepam (benzodiazepine) treatment. Reports in the veterinary literature describe severe cases with a high episode frequency becoming almost clinically normal after administration of the drug (Garosi et al., 2002, Shelton and Engvall, 2002).

Episodic falling shares similarities to human disorders, including hyperekplexia, Brody's myopathy and myotonia. Hyperekplexia is a disease of exaggerated startle response and increased muscle stiffness and rigidity, which shows a particularly close resemblance to EF in terms of the positive response to the drug Clonazepam. Clonazepam is thought to alleviate clinical signs by improving neurotransmission in gamma-aminobutyric acid (GABA) pathways (Tijssen et al., 1997). Mutations in several genes have been associated with hyperekplexia in humans (Table 2.2). Brody's myopathy is a disease of exercise induced muscle cramping with the inability to relax muscles (Brody, 1969). Mutations in the *ATP2A1* (ATPase, Ca++ transporting, cardiac muscle, fast twitch 1) gene have been associated with Brody's myopathy (Odermatt et al., 1996). Myotonia is described as a disease with delayed skeletal muscle relaxation after sudden and often exaggerated contraction, and exists in both autosomal recessive (Becker's disease) and dominant forms (Thomsen's disease) (Becker, 1977, Thomsen, 1876). Mutations in the *CLCN1*

(chloride channel, voltage-sensitive 1) gene have been associated with both autosomal recessive and dominant forms of the disease in humans, in the myotonic “fainting” goat, and myotonia in the Miniature Schnauzer and the Australian Cattle Dog (Beck et al., 1996, Finnigan et al., 2007, George et al., 1993, Koch et al., 1992, Rhodes et al., 1999). Characterisation of many diseases which are closely related to EF at the molecular level presents the opportunity to investigate the disease using a candidate gene approach.

3.1.2. Congenital keratoconjunctivitis sicca and ichthyosiform dermatosis

Congenital keratoconjunctivitis sicca and ichthyosiform dermatosis (CKCSID), commonly known as dry eye and curly coat syndrome, was first reported in the scientific literature in 2006 (Barnett, 2006). Clinical signs are recognisable at birth, with affected puppies having a coat that is rough or “crimped” in appearance (Figure 3.2). Affected puppies are often reported to be smaller than unaffected littermates.

Figure 3.2 Two week old puppy with CKCSID

Two week old CKCS puppy displaying clinical signs consistent with CKCSID. Note the rough appearance of the coat. (Hartley et al 2011.)

Clinical signs of keratoconjunctivitis sicca (dry eye) are apparent from eyelid opening at approximately ten days, with a reduced production of aqueous tears which can result in a discharge of tacky mucus and in severe cases ulceration of the cornea. As CKCSID affected dogs progress to adulthood, the skin can become hyperkeratinised and hyperpigmented across the ventral abdominal region. The coat is harsh and frizzy, with scaling and partial alopecia along the dorsum and flanks, which may cause the dog to scratch. Footpads also become hyperkeratinised with abnormal growth of nails and intermittent sloughing, causing lameness (Figure 3.3).

Prognosis for affected individuals is poor and the condition cannot be resolved through treatment. Many owners elect to euthanise affected dogs due to difficulties in disease management, although life expectancy is not obviously affected by the condition (Hartley et al., 2011).

Figure 3.3 Nail and footpad abnormalities in a case of CKCSID

Abnormal growth of a nail and thickening of the footpad. Images provided by Claudia Hartley.

Ichthyosis (thickened, dry and often scaled skin) has been described in other breeds of dog including the Norfolk Terrier (Credille et al., 2005), Jack Russell Terrier (Credille et al., 2009), and Golden Retriever (Grall et al., 2012), with disease-associated mutations identified for all three breeds. No syndromes reported in the human literature describe clinical signs that are entirely comparable with those seen in cases of CKCSID, although individually both keratoconjunctivitis sicca and ichthyosiform dermatosis are widely reported. Keratitis-ichthyosis-deafness (KID) syndrome, caused by mutations in the *GJB2* (gap junction protein, beta 2) gene encoding connexin-26, shows some clinical similarities to CKCSID (Richard et al., 2002, Skinner et al., 1981). A disease of woolly hair, premature tooth loss, nail dystrophy, acral hyperkeratosis and facial abnormalities has also been described in a human kindred, but unlike CKCSID no ocular clinical signs were described (van Steensel et al., 2001).

The clinical signs and progression of CKCSID in the Cavalier have been described in detail by C. Hartley and colleagues, through the diagnosis and follow up of 25 cases (Hartley et al., 2011). DNA samples were also collected and used for a subsequent candidate gene study, in which microsatellite markers were analysed for association between CKCSID and 28 canine orthologues of human disease-associated genes (Hartley et al., 2012). No association was found between CKCSID and any of the genes

investigated thus ruling out any obvious candidate genes; the study therefore progressed to a GWAS, which is described in this chapter.

3.1.3. Aims

The aims of this investigation were to identify the mutations responsible for EF and CKCSID in the CKCS. As both conditions afflict the same breed, this presented the opportunity to map both conditions in parallel using a set of cases for each disease and a single set of clinically unaffected controls. Association loci from the independent GWAS would be followed up using sequencing techniques in order to identify disease-associated mutations, allowing diagnostic assays to be developed.

3.2. Results

3.2.1. EF candidate gene study

A sample cohort of 12 EF cases and 10 controls was selected for investigation by Professor Jacques Penderis (veterinary neurologist). Canine orthologues of genes causing a similar condition to EF in humans were selected as candidate genes (see Table 2.1). For each candidate gene, two flanking microsatellite markers were identified and genotyped across the sample cohort. At least one microsatellite was required to be informative with a minor allele frequency greater than 0.1 to exclude the gene. The genotyping dataset is shown in Table 3.1 (major alleles are highlighted). For microsatellite markers segregating with EF, cases were expected to be homozygous for a single allele and controls either homozygous or heterozygous for any allele. Upon visual inspection marker GLRA1_C4_60.68 showed a pattern suggesting segregation with EF, with 9/12 cases homozygous for allele 200 and only 3/10 controls with a homozygous genotype for allele 200. The gene locus could be excluded however, by the second microsatellite marker (GLRA1_C4_60.64). No markers surrounding the other candidate genes showed a pattern of linkage disequilibrium with EF based on visual inspection. All three microsatellites genotyped around *SLC5A9* were monomorphic, so the gene could not be formally excluded. The lack of variation across the region may be suggestive of a selective sweep in the CKCS breed.

Table 3.1 Genotypes table for EF candidate genes

Gene targets and genomic coordinates are listed on the left. Major alleles are highlighted in blue or grey. Non-determined alleles are marked as n.d.

Microsatellites			Dog ID numbers and genotypes																													
			Cases												Controls																	
Gene	CFA	Mb	1001	1002	1003	1004	1015	1034	1043	1044	1053	1058	1059	1061	1025	1026	1045	1051	1052	1054	1055	1056	1057	1060
CACNL1A3	7	5.29	n.d. n.d.	284 295	295 295	284 298	285 285	285 295	n.d. n.d.	295 295	285 295	295 295	295 295	285 295	285 298	285 295	n.d. n.d.	295 295	295 295	285 295	295 295	295 295	284 295	295 295
CACNL1A3	7	5.34	284 290	284 284	290 290	284 284	272 272	272 290	n.d. n.d.	284 290	272 290	290 290	n.d. n.d.	272 290	272 286	272 290	284 290	284 290	284 290	272 290	290 290	n.d. n.d.	284 290	284 290
ATP2A1	6	21.26	223 225	223 223	227 227	227 227	223 223	223 227	223 223	223 225	223 223	225 227	223 227	223 227	227 227	223 227	225 227	223 225	223 225	223 225	223 225	223 223	223 223	223 225
ATP2A1	6	21.56	n.d. n.d.	263 265	255 265	255 255	263 263	265 265	263 265	263 265	263 265	255 265	255 265	255 263	265 265	265 265	255 271	265 271	263 271	265 271	265 271	265 265	265 265	265 265
CLCN1	16	9.26	208 208	204 212	206 212	204 212	206 212	206 206	208 208	204 204	206 206	208 208	204 208	206 206	204 212	n.d. n.d.	204 206	204 206	204 208	204 208	204 208	204 208	206 212	204 208
CLCN1	16	9.47	n.d. n.d.	99 101	101 101	99 103	101 101	101 101	n.d. n.d.	n.d. n.d.	101 101	112 112	99 112	101 101	99 103	101 101	n.d. n.d.	99 101	99 112	99 112	99 112	99 112	101 101	99 112
GLRA1	4	60.64	162 162	160 160	160 160	160 162	160 162	160 162	162 162	162 162	160 162	146 160	146 162	146 162	156 162	162 162	146 162	146 162	146 156	160 162	162 162	160 162	146 162	146 162
GLRA1	4	60.68	200 200	200 200	200 200	200 200	200 200	200 200	200 200	200 200	200 200	196 200	196 196	196 200	200 202	200 200	196 200	196 200	196 202	200 200	196 200	200 200	196 200	196 200
GLRA3	25	28.14	212 212	208 208	212 212	208 212	208 212	208 212	212 212	208 208	208 208	208 208	208 208	208 212	208 212	208 212	212 212	208 212	208 212	208 212	n.d. n.d.	212 212	208 208	208 212
GLRA3	25	28.25	147 147	144 144	147 147	144 147	n.d. n.d.	144 147	147 147	144 144	144 144	144 144	144 144	144 147	144 147	144 147	147 147	144 147	144 147	144 147	144 147	147 147	144 144	144 147
GLRB	15	57.36	126 128	126 128	128 128	126 128	128 128	128 128	126 128	126 128	128 128	119 128	119 128	119 126	128 128	126 128	128 128	128 128	128 128	115 130	128 128	126 128	128 128	128 128
GLRB	15	57.49	209 209	209 209	209 209	209 209	209 209	209 209	209 209	209 209	209 209	209 209	209 209	209 209	209 209	209 209	209 209	209 209	209 209	209 209	209 209	209 209	209 217	209 209
GPHN	8	43.92	92 92	104 104	92 92	92 104	98 100	92 98	92 92	92 92	100 100	92 98	92 92	92 98	92 98	92 98	92 92	92 98	92 92	92 104	92 104	92 92	100 104	92 92
GPHN	8	43.98	304 304	310 310	304 304	304 310	306 310	304 306	304 304	304 304	310 310	304 306	304 304	304 306	304 306	304 306	304 304	304 306	304 304	304 310	304 310	304 304	310 310	304 304
KCNE3	21	26.87	256 256	248 254	248 254	248 248	248 256	248 248	256 256	256 256	254 256	256 256	256 256	254 256	256 256	254 256	256 256	248 256	254 256	256 256	248 256	256 256	256 256	256 256
KCNE3	21	27.01	279 279	279 281	285 285	279 279	279 279	n.d. n.d.	279 279	279 279	279 285	279 279	279 279	279 285	279 279	n.d. n.d.	279 279	279 279	279 281	279 279	279 279	279 279	279 279	279 279
SCN4A	9	4.81	185 185	174 174	174 178	174 185	185 185	174 185	185 185	174 185	185 185	174 174	185 185	174 185	174 185	174 185	174 185	174 178	174 185	185 185	174 185	n.d. n.d.	174 185	174 185
SLC32A1	24	30.09	255 255	244 252	244 250	244 244	252 255	244 252	244 244	252 252	252 252	244 252	244 255	244 255	244 244	244 252	255 255	n.d. n.d.	n.d. n.d.	244 255	244 244	244 252	252 255	244 252
SLC32A1	24	30.15	n.d. n.d.	298 306	298 306	298 298	301 306	298 306	298 298	306 306	306 306	298 306	298 301	298 301	298 298	298 306	301 301	298 306	301 306	298 301	n.d. n.d.	298 306	301 306	298 306
SLC6A5	21	45.757	182 182	182 182	182 184	178 182	182 182	182 182	182 182	182 182	182 182	182 182	182 182	182 182	182 182	182 182	182 182	182 182	182 182	182 182	182 182	182 182	182 182	182 182
SLC6A5	21	45.76	244 244	244 244	244 244	239 244	n.d. n.d.	244 244	231 244	244 244	244 244	244 244	244 244	244 244	244 244	244 244	244 244	244 244	244 244	244 244	244 244	244 244	244 244	244 244
SLC6A5	21	45.77	n.d. n.d.	176 176	176 179	176 176	175 175	175 176	n.d. n.d.	175 176	175 175	175 175	175 175	176 176	175 175	n.d. n.d.	176 176	175 176	176 176	175 176	175 176	n.d. n.d.	175 175	175 176
SLC6A5	21	45.92	n.d. n.d.	190 190	181 190	186 190	188 190	188 190	n.d. n.d.	n.d. n.d.	188 190	188 190	188 190	190 190	188 190	188 190	n.d. n.d.	188 190	190 190	188 188	188 188	188 190	188 190	190 190
SLC6A9	15	19.01	207 207	207 207	207 207	207 207	n.d. n.d.	207 207	207 207	207 207	207 207	207 207	207 207	207 207	207 207	207 207	207 207	207 207	207 207	207 207	207 207	207 207	207 207	207 207
SLC6A9	15	19.012	241 241	241 241	241 241	241 241	241 241	241 241	241 241	n.d. n.d.	241 241	241 241	241 241	241 241	241 241	241 241	n.d. n.d.	241 241	241 241	241 241	241 241	241 241	241 241	241 241
SLC6A9	15	19.31	299 299	299 299	299 299	299 299	n.d. n.d.	299 299	299 299	299 299	299 299	299 299	299 299	299 299	299 299	299 299	299 299	299 299	299 299	299 299	299 299	299 299	299 299	299 299

3.2.2. Genome-wide association study

3.2.2.1. Illumina CanineHD SNP array genotyping data

The 96 DNA samples sent for processing on the Illumina CanineHD SNP array genotyped successfully achieving call rates of >99%. Raw genotyping data were imported into Genome Studio for viewing and manual processing to increase the overall genotyping call rate and remove poor quality SNPs. Overall call rate was initially improved by re-clustering the SNP calls (Figure 3.4).

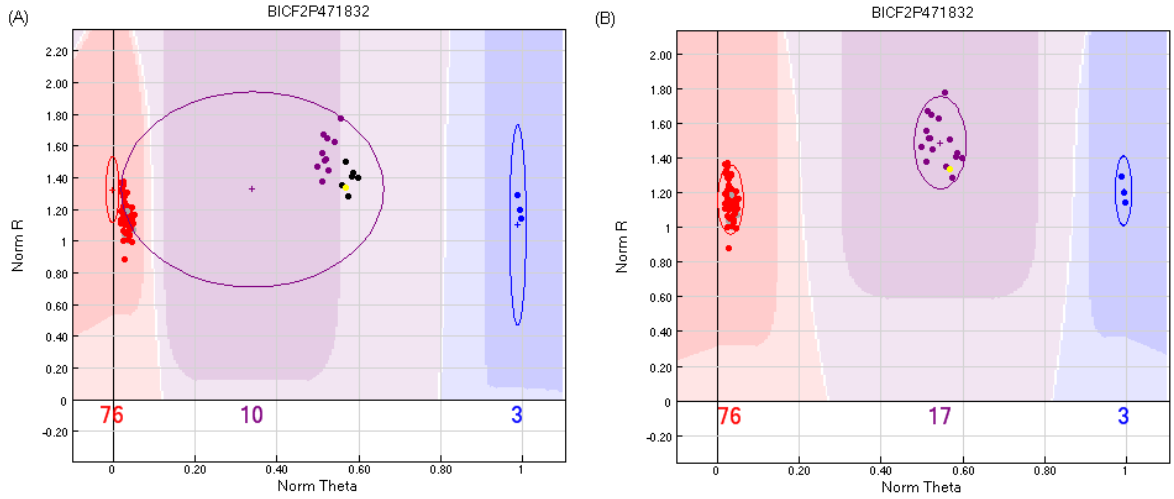


Figure 3.4 Improving call rates using the "cluster all SNPs" command

An example of data improved by the “cluster all SNPs” command in Genome Studio. Each dot on the plots represents a single SNP. The areas shaded in red, purple and blue represent the boundaries for calling alleles AA, AB and BB respectively. (A) Clustering using the predefined cluster file. (B) Clustering after the “cluster all SNPs” command has been executed. The call rate for the SNP has improved from 93% to 100%.

It was noted that SNP calls for four individuals (one Golden Retriever, two IS and one CKCS control) often clustered separately for many otherwise monomorphic SNPs. This caused reclustering errors because of the absence of a heterozygous calls cluster (Figure 3.5A). Call rates were subsequently improved by excluding the four individuals and repeating the SNP re-clustering command. The individuals were then re-included with the new cluster positions in place (Figure 3.5B). Separate clustering was expected for the non-CKCS individuals, but the CKCS control individual appeared to be an outlier and was therefore excluded from further analysis.

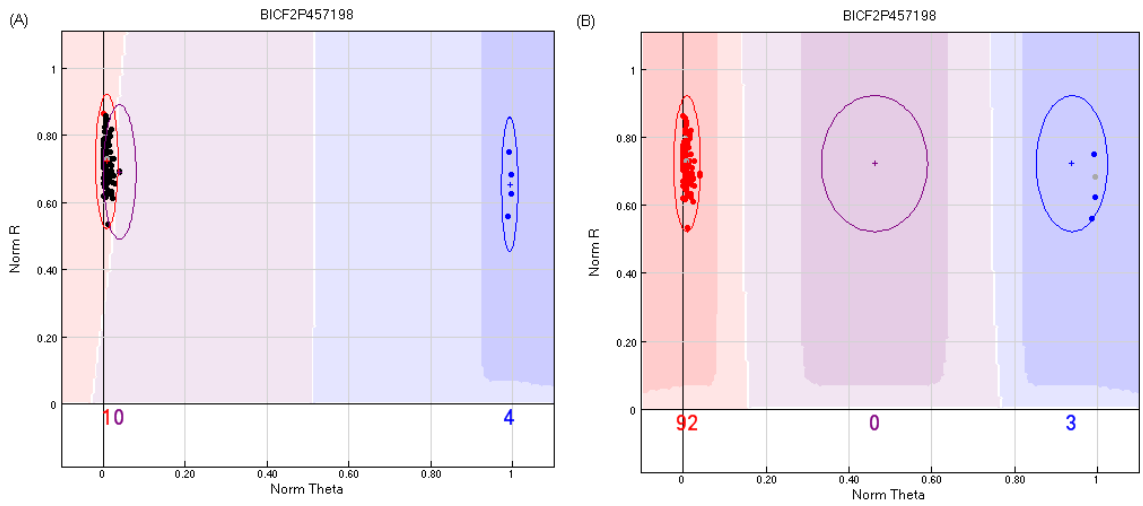


Figure 3.5 SNP clustering problems due to the absence of a heterozygous group

(A) Four individuals often formed a separate homozygous cluster in SNP genotyping data, resulting in calling errors. (B) The clustering problem was resolved by removal of the four individuals before repeating the "cluster all SNPs" command, to improve the average call rate for the entire dataset from 99.48% to 99.84%.

Re-clustered data were exported from Genome Studio into Progeny, for case-control cohort selection.

3.2.2.2. Allelic association analysis

Genotyping data for case-control cohorts were exported from Progeny for allelic association analysis. The sample set for genotyping consisted of 31 EF cases, 19 CKCSID cases, and a common set of 38 controls. SNPs with a minor allele frequency <0.05 and a genotyping call rate <0.95 were excluded from analysis. After filtering 91,427 SNPs remained for EF and 88,384 for CKCSID.

Allelic association analysis was performed using the statistical package PLINK. Strong statistical signals were seen on chromosome 7 for EF (P_{raw} = 1.9 x 10^{-14}) and chromosome 13 for CKCSID (P_{raw} = 1.2×10^{-17}). Allelic association plots for EF and CKCSID are shown in Figures 3.6 and 3.7 respectively. The top 100 SNPs in both studies were located on single chromosomes, producing single distinctive peaks on the association plots.

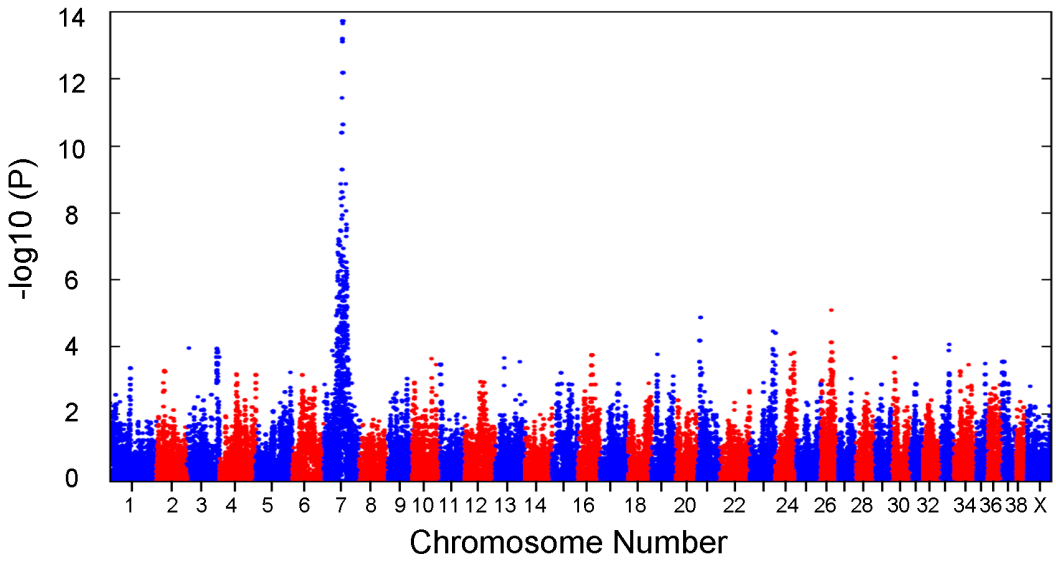


Figure 3.6 Allelic association analysis plot for EF

Allelic association analysis plot for 31 EF cases and 38 controls. Each dot represents a single SNP, with -log10 (p) values on the y-axis plotted against genome position (split into chromosomes) on the x-axis. The strongest statistical signal is on chromosome 7 (P_{raw} = 1.9 x 10^{-14}).

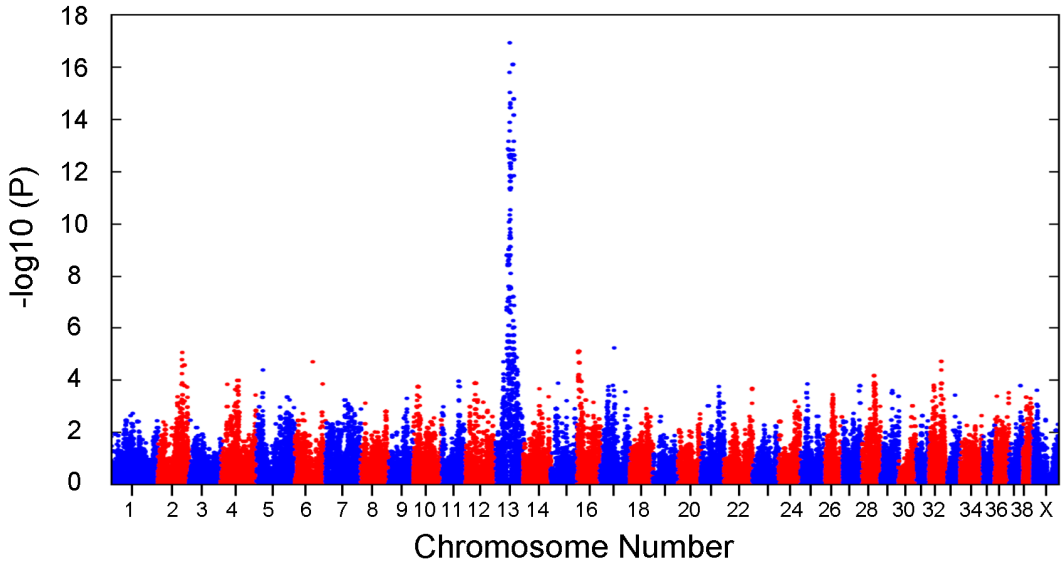


Figure 3.7 Allelic association analysis plot of CKCSID

Allelic association analysis plot for 19 CKCSID cases and 38 controls. Each dot represents a single SNP, with -log10(p) values on the y-axis plotted against genome position (split into chromosomes) on the x-axis. The strongest statistical signal is on chromosome 13 (P_{raw} = 1.2×10^{-17}).

3.2.2.3. Correction for multiple testing

Max(T) permutations analysis (100,000 permutations) was performed using PLINK to correct for multiple testing. For EF and CKCSID single statistical signals remained on chromosomes 7 and 13 respectively, which surpassed genome-wide significance at the 5% level (Figures 3.8 and 3.9). The SNPs displaying the strongest statistical signal for EF and CKCSID after permutations testing both had a P value of 1.0 x 10^{-5} (P_{genome}).

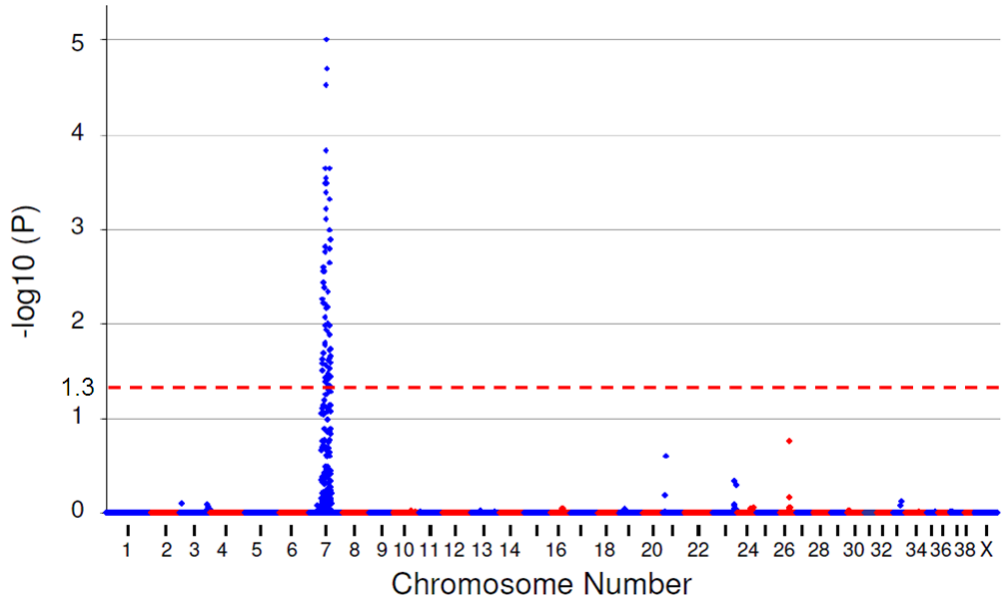


Figure 3.8 EF association analysis including max(T) permutations

Plot of P-values for the EF association study after 100,000 max(T) permutations. Pgenome = 1.0 x 10^{-5}. SNPs above the red dashed line are genome-wide significant at the 5% level.

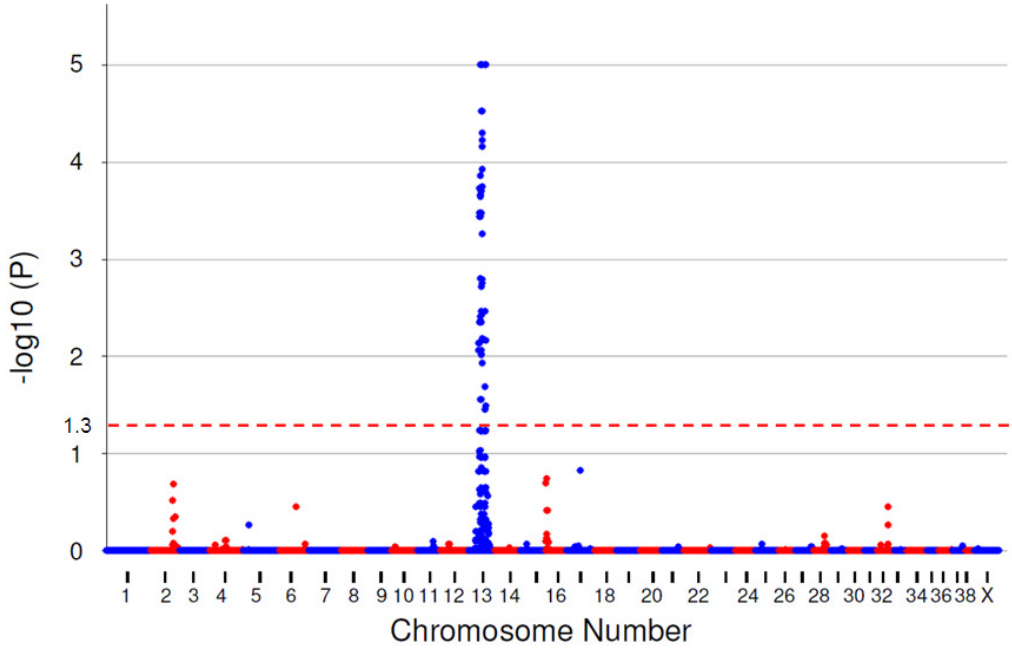


Figure 3.9 CKCSID association analysis including max(T) permutations

Plot of p-values for the CKCSID association study after 100,000 max(T) permutations. Pgenome = 1.0 x 10^{-5}. SNPs above the red dashed line are genome-wide significant at the 5% level.

3.2.3. Population stratification

Genomic inflation values based on the median chi-squared were 1.57 and 1.62 for the EF and CKCSID association analyses respectively, and were suggestive of population stratification in both datasets. Quantile-quantile (QQ) plots of observed versus expected chi-squared values were plotted as a graphical display of genomic inflation (Figure 3.10). The expected pattern for QQ plots in a dataset with no stratification is for the observed chi-squared values to match the expected chi-squared values, and closely track the line y=x. If there is a true signal in the dataset observed chi-squared values will elevate above expected values only for the highest chi-squared values. If genomic inflation is present, the observed versus expected datapoints may be consistently elevated along the line y=x.

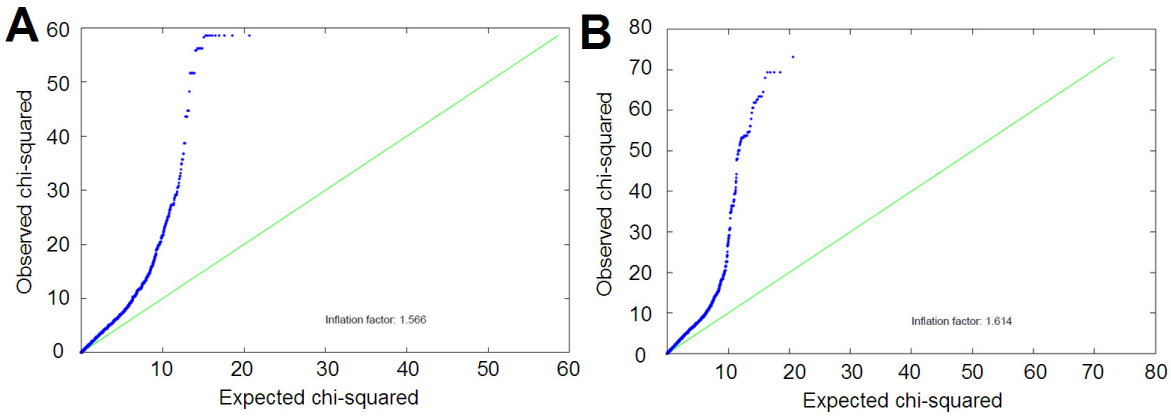


Figure 3.10 QQ plot for the EF and CKCSID SNP genotyping datasets.

QQ plots of observed (y-axis) versus expected (x-axis) chi-squared values for (A) EF and (B) CKCSID. The blue dots represent individual SNPs. The green lines have the equation y=x (ie observed chi-squared = expected chi-squared). The genomic inflation values for EF and CKCSID were 1.57 and 1.62 respectively.

Observed versus expected chi-squared coordinates for both EF and CKCSID datasets, track above the y=x line, and is further evidence of genomic inflation. Multi-dimensional scaling (MDS) plots were generated to assess the relatedness of cases and controls in the EF and CKCSID studies (Figure 3.11). Related individuals cluster closely together on MDS plots. The plots show some separately clustering cases and controls for both EF and CKCSID, which may account for the high genomic inflation values. Clustering of controls away from cases is probably due to a shared set of controls being selected for both EF and CKCSID, rather than tailoring the control set to a particular cohort of cases.

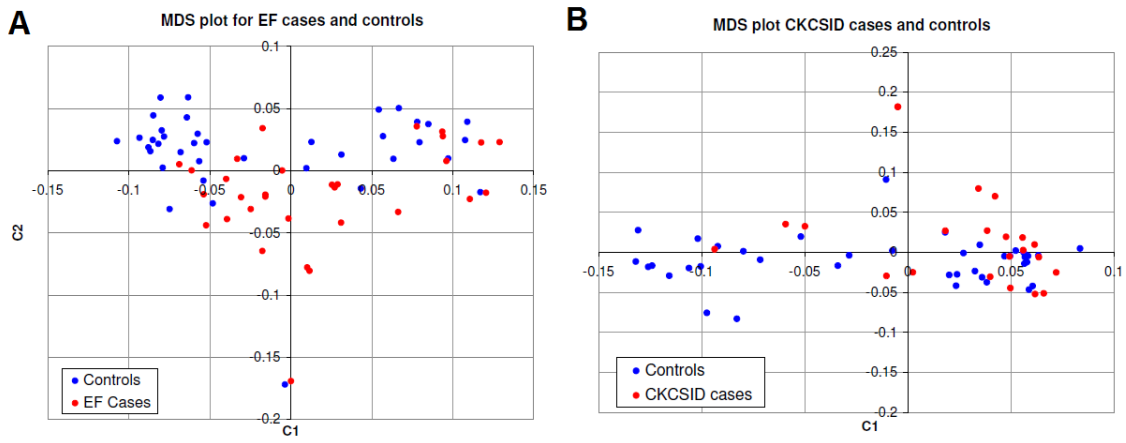


Figure 3.11 MDS plots

MDS plots for (A) EF and (B) CKCSID. Red dots represent cases and blue dots represent controls. Coordinates for the plots were calculated using pairwise identity-by-state (IBS) distance.

3.2.4. Adjusting for genomic inflation

The fast mixed model (FMM) was implemented to adjust for genomic inflation. The two top association signals remained statistically associated at 4.1×10^{-10} and 1.5×10^{-11} for the EF and CKCSID respectively (Figure 3.12).

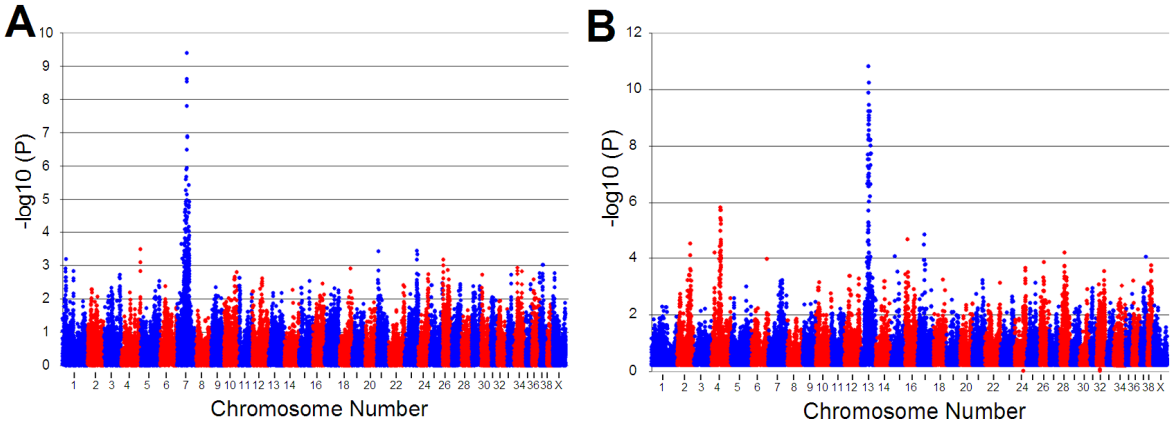


Figure 3.12 FMM corrected allelic association analysis plots

FMM adjusted allelic association plots for (A) EF and (B) CKCSID. Strong statistical signals remained on chromosomes 7 and 13 for EF and CKCSID respectively.

QQ plots are displayed for the FMM adjusted data in Figure 3.13. The genomic inflation value for the EF data after FMM correction was 0.98. Interestingly the genomic inflation value for the CKCSID data after FMM correction remained high at 1.83. The QQ plot for FMM corrected data of observed versus expected chi-squared values does however show the datapoints to more closely follow the line y=x, indicating that the values have been adjusted.

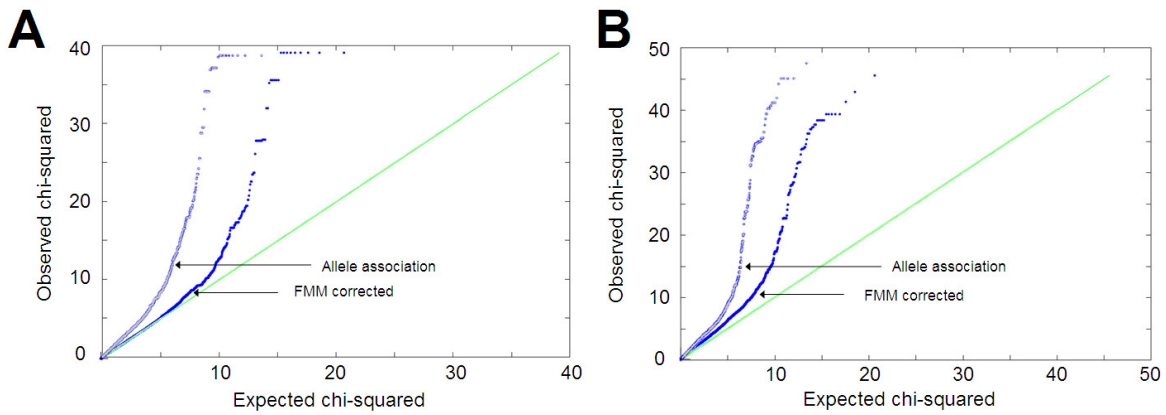


Figure 3.13 QQ plots of FMM adjusted data

QQ plots of observed (y axis) versus expected (x axis) chi-squared values for (A) EF and (B) CKCSID after adjusting for genomic inflation using the fast mixed model. The green lines track the x=y coordinates.

3.2.5. Investigation of strong statistical signals at the chromosome level

Strong statistical signals were first investigated at the chromosome level by plotting raw SNP p-values across chromosome 7 and 13 for EF and CKCSID respectively. For EF the focal point of the strong statistical region was between 40 Mb and 48 Mb on chromosome 7 (Figure 3.14). For CKCSID the statistical signal was spread over a slightly larger region of 32 Mb to 46 Mb on chromosome 13 (Figure 3.15).

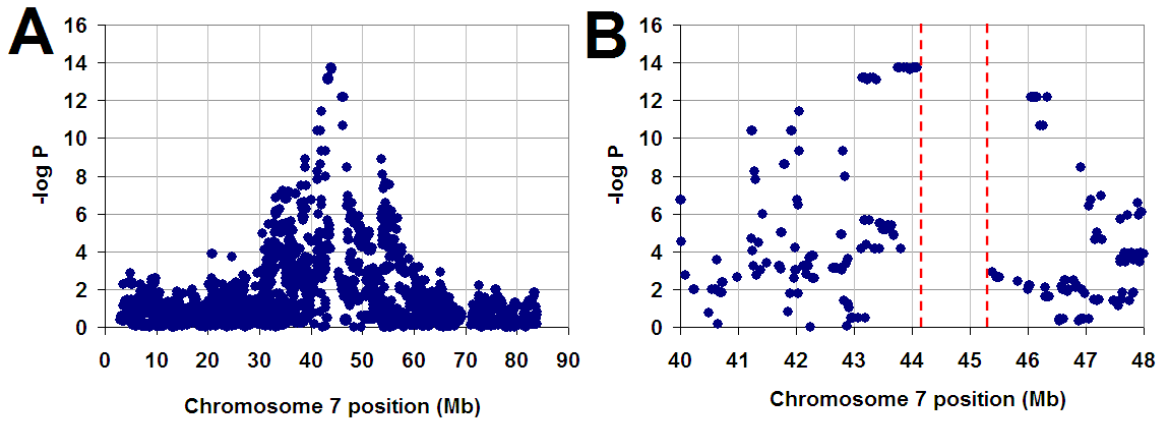


Figure 3.14 Chromosome 7 allelic association plots for the EF study

(A) Allelic association plot for EF across chromosome 7. (B) Focal point of the strong statistical signal. The signal is interrupted by a region of approximately 1 Mb containing no SNPs due to the minor allele frequency filtering parameters defined in the analysis.

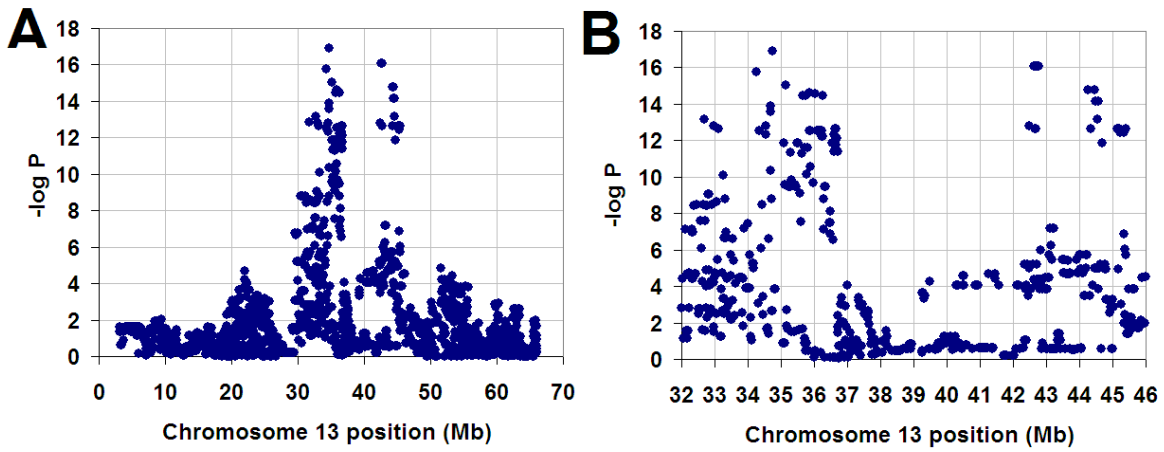
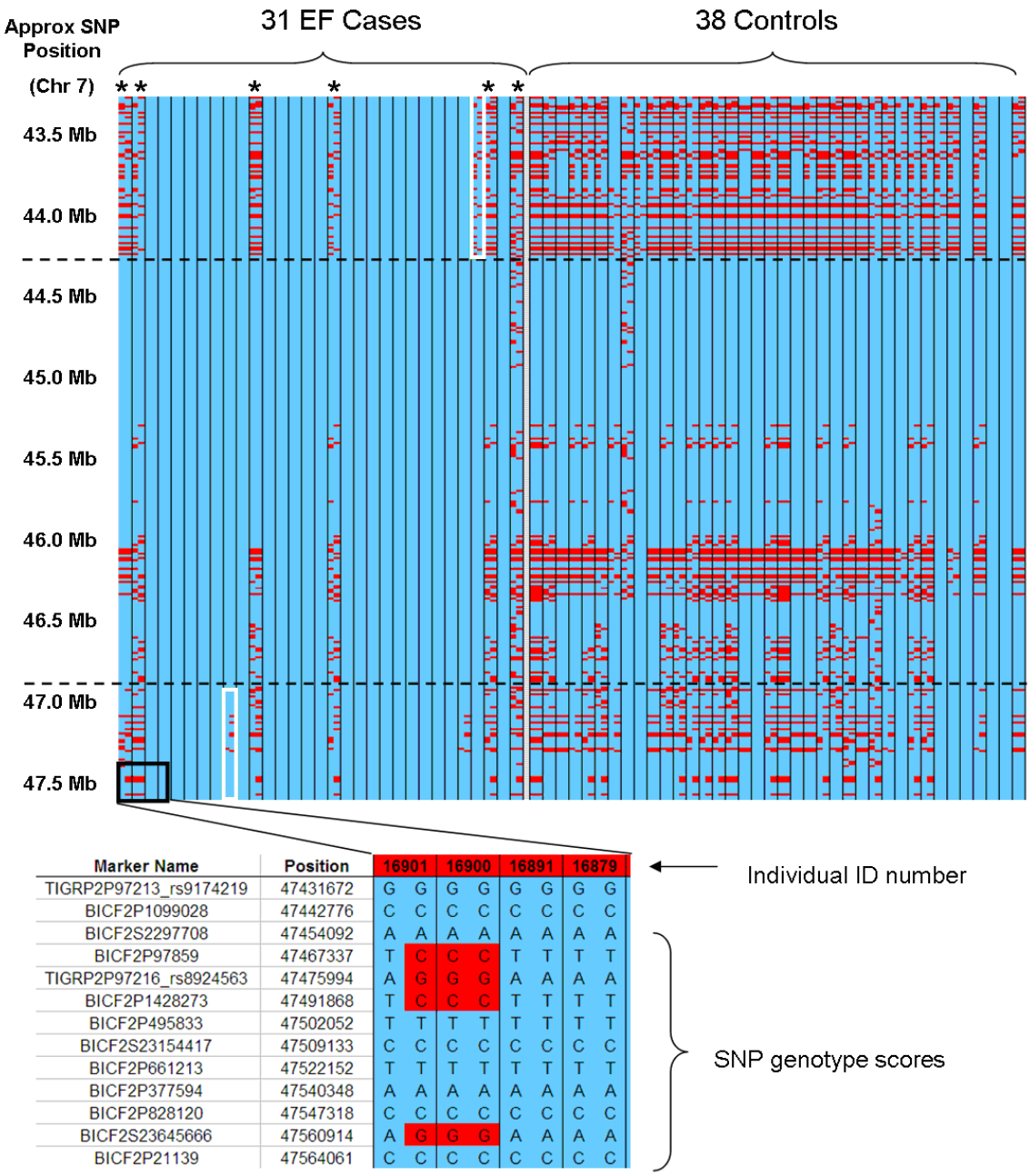


Figure 3.15 Chromosome 13 allelic association plot for the CKCSID study

(A) Allelic association plot for CKCSID across chromosome 13. (B) The focal point of the strong statistical signal.

Regions of strong association were further investigated by visualising raw SNP genotyping data across the disease-associated regions. For EF two individuals defined the shared disease-associated haplotype or "critical region" as CFA7:44,093,554-46,905,272, by a loss of shared homozygosity in cases due to recombination events (Figure 3.16). Six individuals suspected to be affected on the basis of phenotype did not share the disease-associated haplotype, and are marked with an asterisk in Figure 3.16. One of these cases was subsequently reclassified as an epilepsy case in an independent neurological work-up. The other cases could not be resolved because the dogs were deceased or owners could not be contacted. The critical region contained a 1.2 Mb haplotype that was homozygous across 67 of 69 cases and controls, which could be suggestive of a selective sweep in the CKCS breed. Two cases were critical in defining the EF disease-associated region, and it was therefore imperative that the diagnosis of these cases was correct. The case defining the upper boundary (15943) was a typical but severe case of EF, with the owner describing clinical signs during an episode of a arched back, trembling, collapse, stiff paralysed hind legs, extension of the forelimbs above the head and jaw locking. Episodes reportedly lasted from two minutes to up to an hour. A video clip was also provided with the dog presenting an episode at 4 months of age, with clinical signs consistent with EF. The sample defining the lower boundary of the region was 16867 (JP1049). Only a single episode was reported and little further information was available for this case. Owner and veterinarian details were unavailable. Sample 16155 could define the lower boundary at a slightly lower position of 47,048,914 on canine chromosome 7. This case had an initial episode at 6 months of age, presenting with collapse and hyperextension of limbs. The case was treated with and responded to Rivotril. Video footage provided with the case showed clinical signs consistent with EF.

This case therefore provided a more assured definition of the 3' boundary. The EF critical region contained 114 genes and was syntenic with human chromosome 1 (Appendix 6).



Marker Name	Position	16901		16900		16891		16879	
TIGRP2P97213_rs9174219	47431672	G	G	G	G	G	G	G	G
BICF2P1099028	47442776	C	C	C	C	C	C	C	C
BICF2S2297708	47454092	A	A	A	A	A	A	A	A
BICF2P97859	47467337	T	C	C	C	T	T	T	T
TIGRP2P97216_rs8924563	47475994	A	G	G	G	A	A	A	A
BICF2P1428273	47491868	T	C	C	C	T	T	T	T
BICF2P495833	47502052	T	T	T	T	T	T	T	T
BICF2S23154417	47509133	C	C	C	C	C	C	C	C
BICF2P661213	47522152	T	T	T	T	T	T	T	T
BICF2P377594	47540348	A	A	A	A	A	A	A	A
BICF2P828120	47547318	C	C	C	C	C	C	C	C
BICF2S23645666	47560914	A	G	G	G	A	A	A	A
BICF2P21139	47564061	C	C	C	C	C	C	C	C

Figure 3.16 EF critical region raw genotyping data

Raw genotyping data across the EF disease-associated critical region. Each column represents an individual and SNP markers are listed in rows. Cases are listed on the left, controls on the right. Major alleles are coloured in blue, minor alleles in red. Two individuals showed a loss of homozygosity due to recombination events, highlighted by a white border, defining the disease-associated haplotype as CFA7:44,093,554-46,905,272 (boundary marked by black dashed lines). Six cases appear to be apparent outliers that were not homozygous for the disease-associated haplotype (columns marked with an asterisk)

For CKCSID recombination events in two individuals defined the disease-associated haplotype as CFA13:39,648,169-42,481,707, which was shared between all cases (Figure 3.17). Diagnosis of cases was checked and confirmed in consultation with Claudia Hartley (veterinary ophthalmologist). The CKCSID critical region contained 85 genes, and was syntenic to regions of human chromosomes 4, 8 and 15 (Appendix 6).

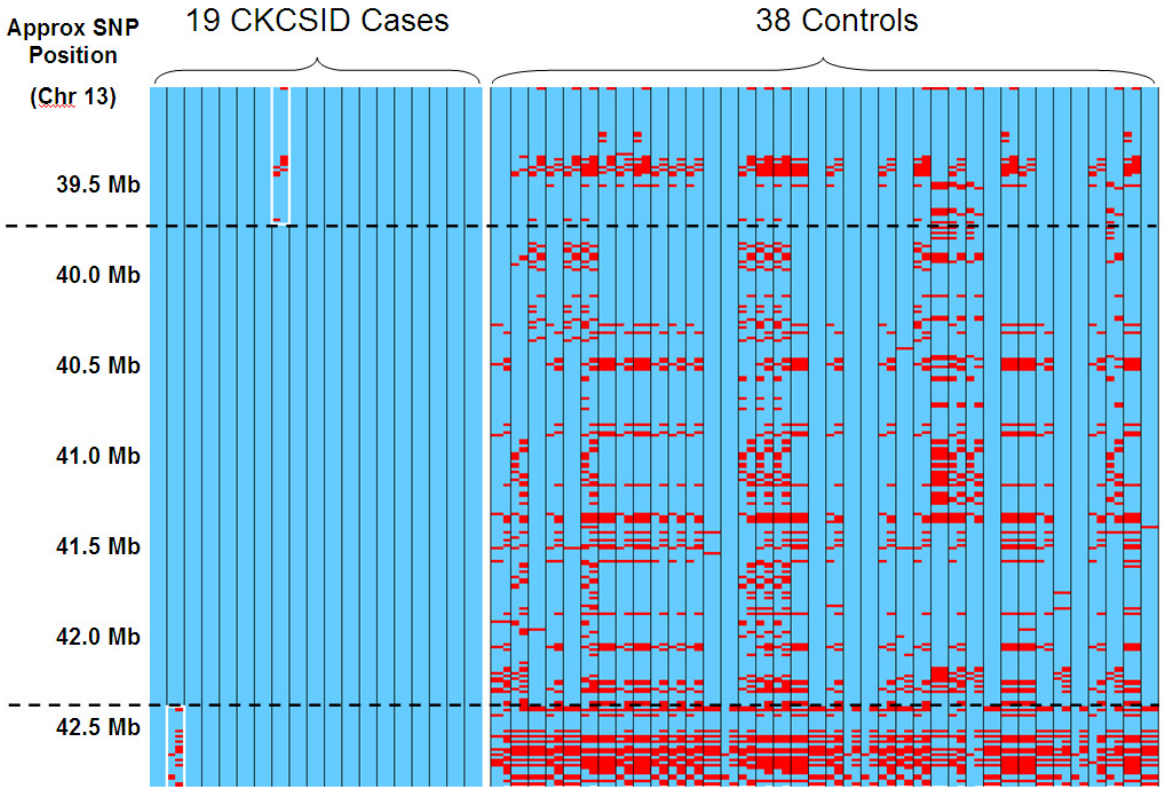


Figure 3.17 CKCSID disease-associated critical haplotype

Raw genotyping data across the CKCSID disease-associated critical region. Each column represents an individual and SNP markers are listed in rows. Cases are listed on the left, controls on the right. Major alleles are coloured in blue, minor alleles in red. Two individuals showed a loss of homozygosity due to recombination events, highlighted by a white border, defining the disease-associated haplotype as CFA13:39,648,169-42,481,707 (boundaries marked by black dashed lines).

Using the Ensembl genome browser both regions were interrogated for genes that could be potential candidates for the respective disorders. The EF critical region contained no genes that were associated with similar conditions in humans. The gene *SLURP1* was identified in the CKCSID critical region. Mutations in *SLURP1* have been associated with Mal de Meleda, an autosomal recessive skin disorder in humans, with clinical signs of transgressive palmoplantar keratoderma, keratotic skin lesions, perioral erythema, brachydactyly and nail abnormalities (Fischer et al., 2001), and was considered a good candidate gene for CKCSID that had not previously been investigated.

3.2.6. *SLURP1* sequencing

Because of the similarities between Mal de Meleda and CKCSID, the gene was exon re-sequenced in two CKCSID cases and two control individuals. No polymorphisms were identified and the gene was ruled out as being potentially causal.

3.2.7. SureSelect target enrichment and massively parallel sequencing

3.2.7.1. Probe design

Probes (RNA baits) were designed for the SureSelect solution based target enrichment system using the online tool e-array (https://earray.chem.agilent.com/earray/). Design parameters were set to capture both EF and CKCSID critical regions with 2x bait tiling (ie each base was covered by at least two RNA baits where possible) and repeat masking was applied (ie no probes were placed across known repeat regions). A total of 57,676 baits were designed across a combined region of 5,788,900 bp; 29,429 baits for the EF region and 28,247 baits for the CKCSID region, achieving a base coverage of 3,749,814 (64.8%). Bait locations were uploaded to the UCSC genome browser to assess bait coverage across exonic regions. Coverage of exons appeared to be nearly complete, although isolated examples of exons with no bait coverage were occasionally seen, often due to exons being in close proximity to repetitive elements (Figure 3.18).

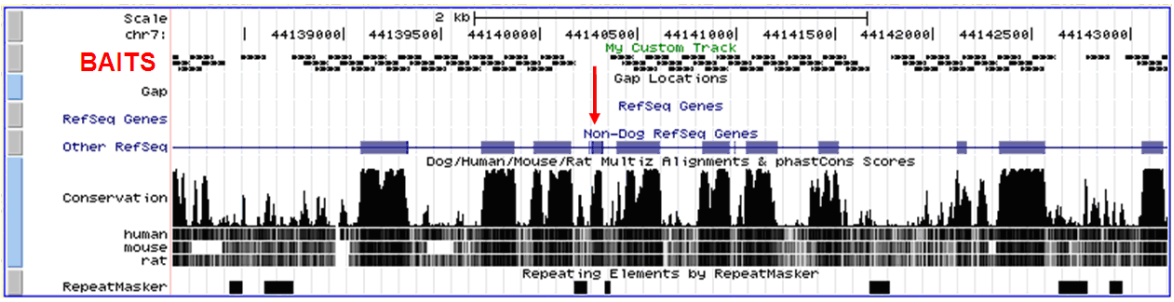


Figure 3.18 Visualisation of baits on the UCSC genome browser

Positioning of baits across a section of the EF region and displayed by the UCSC genome browser. An exon with zero bait coverage is indicated by the red arrow.

3.2.7.2. Investigation of fragmentation methods

A key stage of library preparation for next generation sequencing is fragmentation of the nucleic acid. At the time of investigation the recommended method of fragmentation for Illumina sequencing was nebulisation. To help fulfil the aim of completing bench work in-house, three methods of DNA fragmentation were tested.

3.2.7.2.1 Sonication

The results of sonication for 10, 20 and 30 cycles of 20 seconds on / 20 seconds off at full power using a cup horn sonicator are shown in Figure 3.19. Fragmentation of DNA with a size range of between 100 bp and 1,000 bp was achieved.

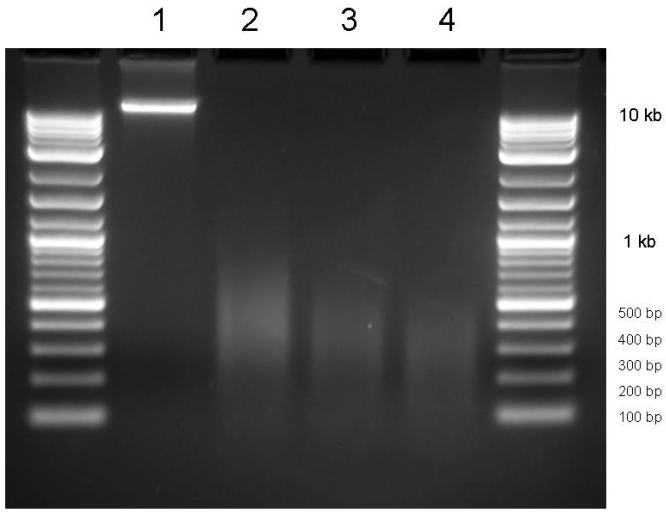


Figure 3.19 DNA sonication results

Results of sonication of high molecular weight genomic DNA (1.5% agarose gel). Lane 1 contains 200 ng non-treated high molecular weight genomic DNA. Lanes 2, 3 and 4 contain 200 ng genomic DNA sonicated for 10, 20 and 30 cycles of sonication respectively (20 seconds on / 20 seconds off at full power).

3.2.7.2.2 Nebulisation

Nebulisation using a pressure of 24 psi and 50% glycerol buffer produced fragments in the 150–1000 bp range (Figure 3.20A). On agarose the fragments appeared most abundant in the 400–600 bp range. Only 350 µl of the 750 µl starting material could be recovered after nebulisation, indicating that half the material is lost during the shearing process, through vaporisation.

3.2.7.2.3 Double stranded DNA Fragmentase

Genomic DNA was digested for 30 minutes using NEB dsDNA Fragmentase. The process yielded fragments of 100-1,000 bp. The highest concentration of fragments was in the 100-600 bp size range (Figure 3.20B).

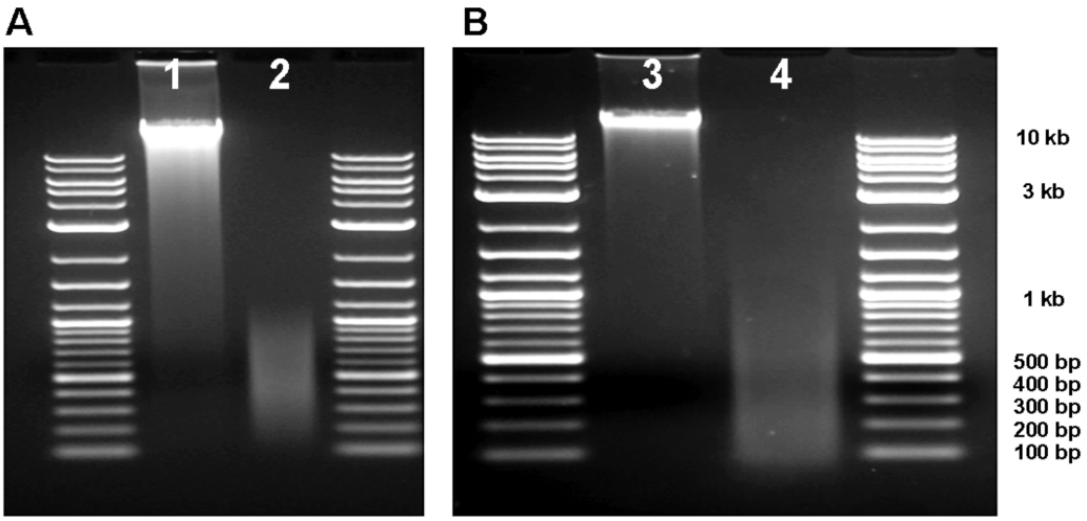


Figure 3.20 Nebulised and fragmentase treated genomic DNA

(A) Genomic DNA treated by nebulisation. Lane 1 - 200 ng high molecular weight genomic DNA. Lane 2 - 200 ng nebulised genomic DNA. (B) Genomic DNA before (lane 3) and after (lane 4) treatment with NEB dsDNA Fragmentase.

3.2.7.3. Trial library preparation and clone sequencing

Trial libraries were prepared to test in-house library preparation using the NEBnext kit. To reduce costs, primer and adapter sequences for use in the library preparation were synthesised by an oligonucleotide manufacturer, rather than being purchased from Illumina. Trial library preparation would also test the effectiveness of these custom made oligonucleotides. Three libraries were made in the trial. Libraries 1 and 2 were made using nebulised DNA. Library 3 was created from DNA treated with dsDNA Fragmentase. Libraries were agarose gel size selected after adapter ligation and amplified with Phusion polymerase using the indexing method. Library fragments were checked for correct adapter sequences by molecular cloning. Results are shown in Table 3.2.

Overall 52% of sequenced cloned fragments had perfect adapter sequences at both ends. The percentage of correctly adapted fragments was higher for Fragmentase treated DNA (77%) compared with nebulised DNA (42%). Fragmentase treatment was chosen for use in SureSelect experiments based on cloning results and the high proportion of fragments produced in the desired 100-300 bp range.

Table 3.2 Summary of the clone sequencing results.

LIBRARY No	CLONE No	PCR	SEQUENCE	ADAPTERS OK?	INSERT SIZE
1	1	3 products	NO	-	-
1	2	OK	YES	YES	319
1	3	OK	YES	NO	250
1	4	OK	YES	NO	113
1	5	OK	YES	YES	273
1	6	OK	NO	-	-
1	7	OK	NO	-	-
1	8	OK	YES	NO	201
2	1	OK	YES	NO	212
2	2	OK	YES	YES	171
2	3	OK	YES	YES	225
2	4	Faint	NO		-
2	5	OK	YES	NO	236
2	6	OK	YES	NO	251
2	7	OK	YES	YES	225
2	8	OK	YES	NO	201
3	1	No product	NO	-	-
3	2	OK	YES	YES	286
3	3	OK	YES	YES	275
3	4	OK	YES	YES	275
3	5	Two products	NO	-	
3	6	OK	YES	NO	248
3	7	OK	YES	YES	249
3	8	OK	YES	NO	258
3	9	Unexpected size	NO	-	-
3	10	OK	YES	YES	207
3	11	OK	YES	YES	274
3	12	OK	YES	NO	280
3	13	OK	YES	YES	255
3	14	OK	YES	YES	293
3	15	OK	YES	YES	322
3	16	OK	YES	YES	265

3.2.7.4. Sample selection for target enrichment

Two EF cases, two CKCSID cases and three controls were selected for potential use in the first attempt at target enrichment (Table 3.3). Cases were selected that were homozygous for the defined disease-associated haplotypes with clinical diagnosis confirmed by Jacques Penderis or Claudia Hartley. Controls were selected based on haplotype structure across the two disease-associated regions. A control with chromosome 7 and 13 haplotypes identical to disease-associated regions, but clinically unaffected with respect to EF and CKCSID was selected in an attempt to reduce the number of potential causal variants.

Table 3.3 Individuals selected for target enrichment

WTH - wild-type haplotype; DAH - disease-associated haplotype. Note control sample 16874, although clinically normal, is homozygous for the chromosome 7 and chromosome 13 disease-associated haplotypes.

Sample ID	Clinical status	CKCSID region	EF region
6823	Control	Heterozygous	Homozygous WTH
6975	**CKCSID case**	Homozygous DAH	Homozygous WTH
9916	**CKCSID case**	Homozygous DAH	Heterozygous
15943	**EF case**	Homozygous WTH	Homozygous DAH
16823	**EF case**	Homozygous WTH	Homozygous DAH
16874	Control	Homozygous DAH	Homozygous DAH
16878	Control	Homozygous WTH	Heterozygous

3.2.7.5. Library preparation including pre-capture amplification

Libraries were fragmented to the 100-600 bp range using dsDNA Fragmentase, and prepared using the NEBnext kit. Fragments in the 200-300 bp range were selected after adapter ligation. Nine cycles of pre-capture amplification were required to increase the concentration of the libraries to 147 ng/µl for the hybridisation stage, which was slightly higher than the 4-6 cycles recommended by the SureSelect protocol. The increased number of cycles was due to a change in methodologies recommended by SureSelect. In the SureSelect Illumina single-end protocol v2.0 (December 2009) the recommended fragmentation method was nebulisation and included a size selection stage. In the SureSelect Illumina paired-end protocol for multiplexed sequencing v1.0 (May 2010) the recommended fragmentation methodology had been changed to Covaris shearing with no size selection, allowing a reduction in the number of PCR cycles required. As no Covaris service was available at the time of the experiment and to keep the library preparation in-house a size selection stage was adopted. Results of the pre-capture amplification are shown in Figure 3.21.

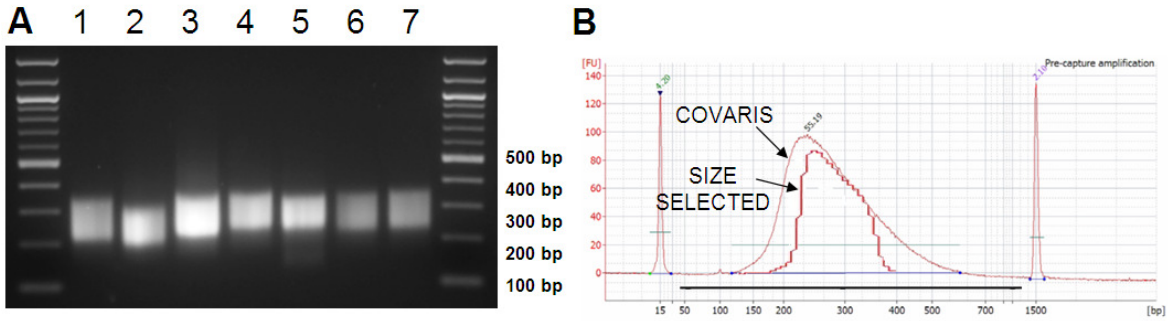


Figure 3.21 Precapture libraries

(A) Precapture libraries (147 ng) on a 2% agarose gel. Lanes 1 to 7 correspond to samples 6823, 6975, 9916, 15943, 16823, 16874 and 16878 respectively. (B) Pre-capture library 9916 was subjected to analysis on a Bioanalyser DNA 12000 chip. The size range of fragments falls inside of those seen for library preparation after Covaris shearing.

3.2.7.6. SureSelect hybridisation and post-capture amplification

Five libraries were taken forward to the enrichment stage consisting of two EF cases, two CKCSID cases and a single control which was homozygous for both the chromosome 7 and chromosome 13 disease-associated haplotypes. The hybridisation stage was followed by 13 cycles of post capture amplification to produce the final sequencing libraries, which were quantified using KAPA library quantification qPCR results in conjunction with results of accurate library sizing on the Bioanalyser. Libraries were pooled in equal amounts to 10 nM. Blunt end cloning of pooled library fragments was carried out to check end sequences and to estimate capture efficiency before libraries were sent for sequencing. A total of 23 colonies were selected for PCR and sequencing.

Results indicated a capture efficiency of 78% and an average insert size of 194 bp (Table 3.4).

Table 3.4 Summary of blunt end cloning results to estimate capture efficiency

Clone sequences not hitting target regions after BLAST searching are highlighted in yellow.

Clone No.	Index No	Chromosome	Position (Mb)	Insert size (bp)	Ends ok?
1	2	13	41.86	221	Y
2	2	13	41.85	160	Y
3	2	7	44.85	134	Y
4	4	7	45.59	231	Y
5	4	13	42.12	214	Y
6	4	7	46.08	176	Y
7	4	1	68.79	256	Y
8	4	7	45.97	148	Y
9	4	13	40.27	172	N
10	5	13	41.49	210	Y
11	5	7	45.09	166	Y
12	5	13	42.10	196	Y
13	5	7	83.59	163	N
14	5	7	44.96	221	Y
15	5	7	45.86	190	N
16	5	1	52.85	244	Y
17	6	7	46.27	189	Y
18	6	7	44.85	158	N
19	6	16	51.72	255	N
20	6	7	44.53	140	Y
21	6	13	40.27	172	Y
22	7	13	42.07	228	Y
23	7	26	16.22	213	Y

3.2.8. Illumina raw sequencing results

Illumina sequencing on the GAIIx generated a 3.47 Gb dataset consisting of 68 million reads of 51 bp in length. As expected from cloning experiments the mean insert size was ~200 bp and read share was evenly distributed amongst the chosen indices (Figure 3.22).

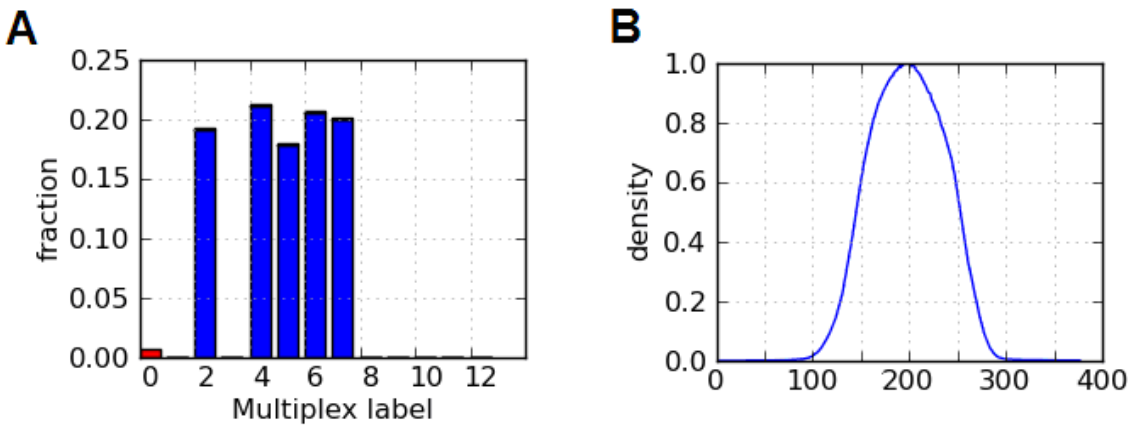


Figure 3.22 Summary histograms of CKCS Illumina sequencing

(A) Proportion of reads allocated to each index. Index 2, 4, 5, 6 and 7 were used in library preparations of samples 6975, 9916, 15943, 16823 and 16874 respectively. Label 0 represents reads that could not be demultiplexed from the raw data. (B) Insert size histogram.

3.2.9. Development of a data analysis pipeline

Illumina sequencing data for the CKCS was available as raw unaligned FASTQ files only. An example of a single read in FASTQ format is shown below.

```
@SEQUENCE_ID
TACGATCGATGCTAGCATGCATGCTTGACTGGGGACTGATCGTAGCTAGGATCGTGCAGT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

The @ symbol which acts as a marker is followed by an identifier tag on the first line. The second line details the read sequence. The third line has a + symbol marker and the fourth line shows the per base quality score for the sequence in line two.

The two key requirements for analysis of massively parallel sequencing data are the ability to align reads to a reference genome and to call variants. A huge number of freeware programs are available online for handling and manipulating data generated by next generation sequencing platforms. Most are operated from the Linux command prompt and long strings of complex commands are often required for their operation. Additionally each program has different input requirements, and files often have to be manipulated by associated programs before they can be used to perform the desired task. This means that a long chain of complex commands have to be executed individually to achieve the desired results. To simplify the process a Perl script called “NGS analysis”, which could be run from the Linux command line, was written to sequentially process the string of required commands, after the user had been prompted to enter simplified details for the analysis. The initial user interface is shown in Figure 3.23.

3.2.9.1. Features of the NGS analysis pipeline

The NGS pipeline can handle paired-end or single-end Illumina data in FASTQ format and single end 454 data in ssq format. Reads are aligned to a reference genome using the program BWA. The best practice is to align data to a whole genome, but data can be aligned to an individual chromosome to reduce computational requirements. Datasets can also be reduced for easier handling. For instance if a target enrichment is performed across two genomic regions, alignment files can be created for each region separately to reduce file size and subsequent processing time. The pipeline will create SNP and indel calls, and there is an option to run the Ensembl Variant Effects Predictor to annotate calls with genomic information, such as consequence for variants positioned within exons.

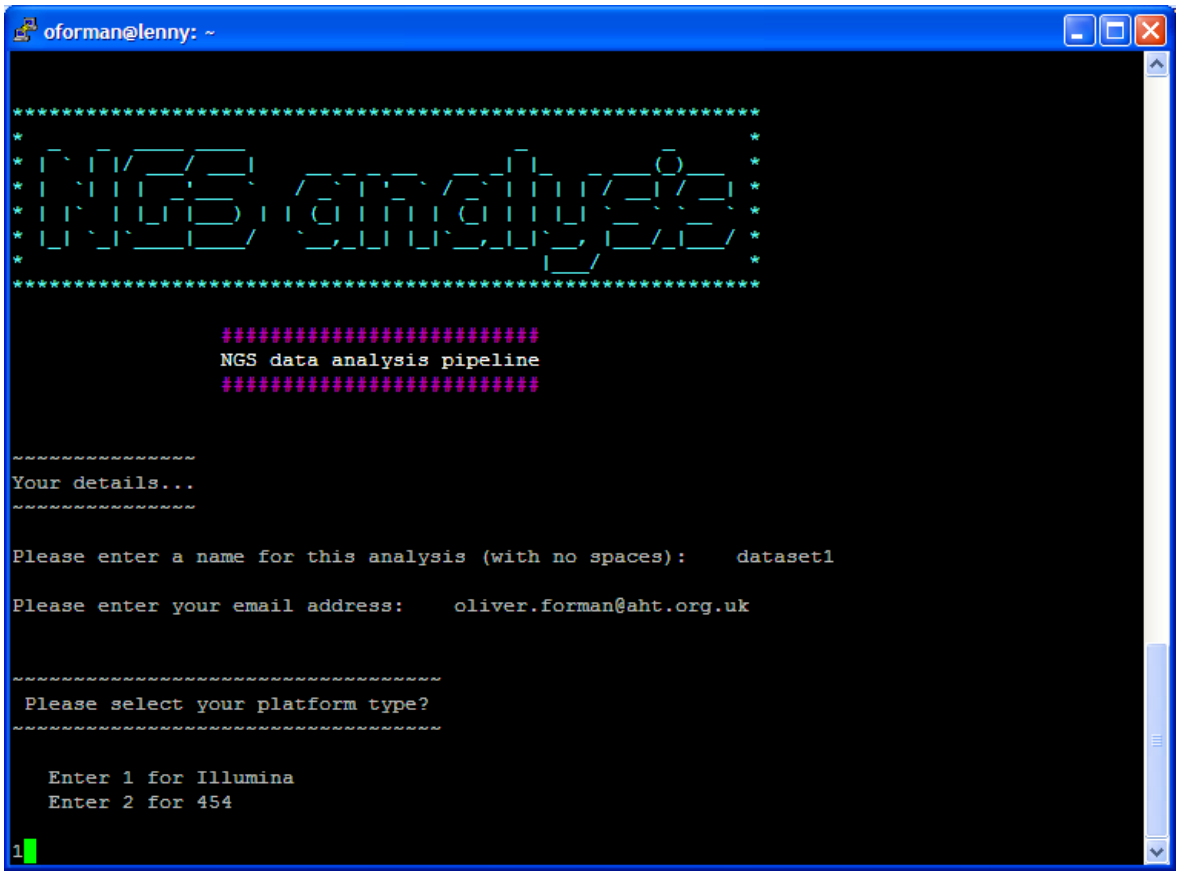


Figure 3.23 NGS analysis Perl script user interface

Screenshot of the user interface displayed at the start of the NGS analysis Perl script. The script was written to enable users to enter file names and analysis parameters, before automatic and sequential implementation of commands in the script to generate results files.

Other modules in the pipeline include removal of PCR duplicates, an option to run the structural variant analysis program Pindel, an option to run an early next generation sequencing analysis package called Maq and tools to produce alignment summary, GC bias and insert size histograms. Key files are transferred to a new results folder that is created at the end of the pipeline. A log is created during the running of the pipeline that details the analysis settings selected. If an error appears in the log file the script will automatically terminate and send the user an email notification. Email notification is also given to users on completion of the analysis, which may take several hours depending on the size of the dataset. A summary of results files is listed in Appendix 7. This list is also presented in a readme.txt file created at the end of the analysis pipeline. A workflow of options available to users of the NGS analysis pipeline is shown in Appendix 8 and a video demonstrating use of the NGS pipeline is available on the supplementary CD.

3.2.10. Sequence data analysis

The paired sequencing data files for each sample were passed through the sequence analysis pipeline. A summary table of the processed data is shown in Table 3.5.

Table 3.5 Summary statistics for sequencing data

Individual ID	No. of reads (Million)	Dataset size (Mb)	Target enrichment efficiency (%)	Percent bases achieving 10x Coverage	PCR duplicates (%)
6975	13.1	673	84.8	77.9	14.9
9916	14.6	746	85.0	79.2	9.5
15943	12.3	627	85.1	78.7	11.2
16823	14.2	723	85.7	79.0	13.8
16874	13.7	701	85.7	80.2	18.5

SNP calls across all samples were combined into a single file format using The NGS SNP Handler, an Excel macro tool written by Dr Mike Boursnell. Across the five samples a total of 13,301 SNPs were identified. The NGS SNP Handler was adapted to accept indel files after completion of the CKCS investigation. Re-analysis of the data showed that 1,149 indels were identified across the five datasets.

Compiled SNP calls from the pipeline were initially filtered by expected segregation pattern (ie homozygous in cases and either heterozygous or homozygous wild-type in controls). The positions of these candidate SNPs were used to create a Browser Extensible Data (BED) file for uploading to the Integrative Genome Viewer (IGV). The five indel files that were created by the NGS analysis pipeline were in BED format and could be loaded directly into IGV. Ensembl gene predictions were uploaded to IGV to annotate the target regions which were then scanned visually for candidate variants occurring in exonic regions. This method of analysis revealed no variants for the EF region that could be considered as potentially causal. In the CKCSID region one indel was identified in an exonic region of *FAM83H* (family with sequence similarity 83, member H) which fully segregated with the disorder. To verify the indel, the sequence alignment files (.BAM files) were uploaded to IGV and the sequence reads were viewed across the candidate locus (Figure 3.24). The two CKCSID cases were homozygous for the indel at the locus and the two EF cases and the control were wild-type homozygous.

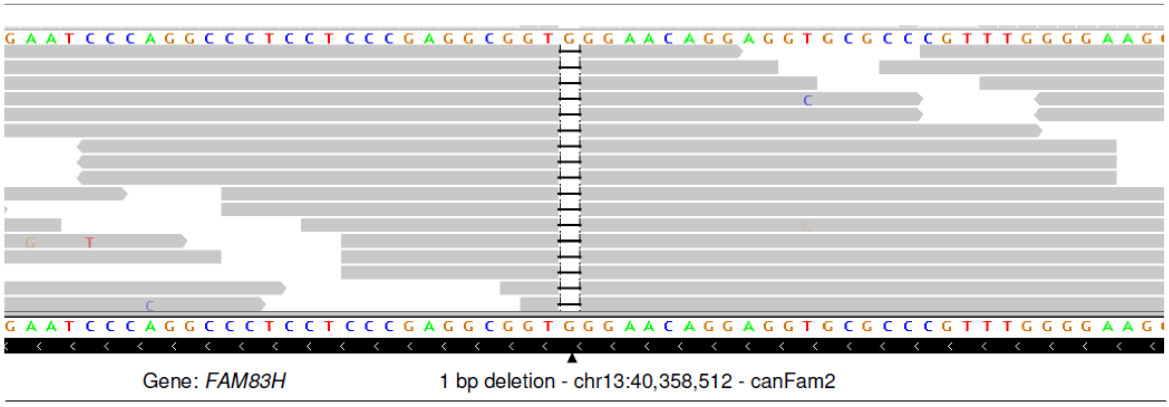


Figure 3.24 Reads from a CKCSID case aligned across the *FAM83H* candidate locus

View in IGV of reads aligned across the *FAM83H* candidate locus for a single CKCSID case, enabling the indel call to be validated. Grey bars represent individual reads. The indel is represented by a solid black horizontal line in reads.

As no potential causal SNP or indel calls had been identified for EF, the five read alignment files were loaded into IGV and the EF target region manually scanned for other potential causal variants. Three large deletions of ~6 kb, ~10 kb and ~16 kb were identified, although only the latter was situated across a coding region. The genomic positions of the three deletions within the disease-associated interval and a schematic overview of normal gene structure are shown in Appendix 6. The 16 kb deletion spanned the first three exons of the gene brevican (*BCAN*) and was potentially a full gene knock-out (Figure 3.25).

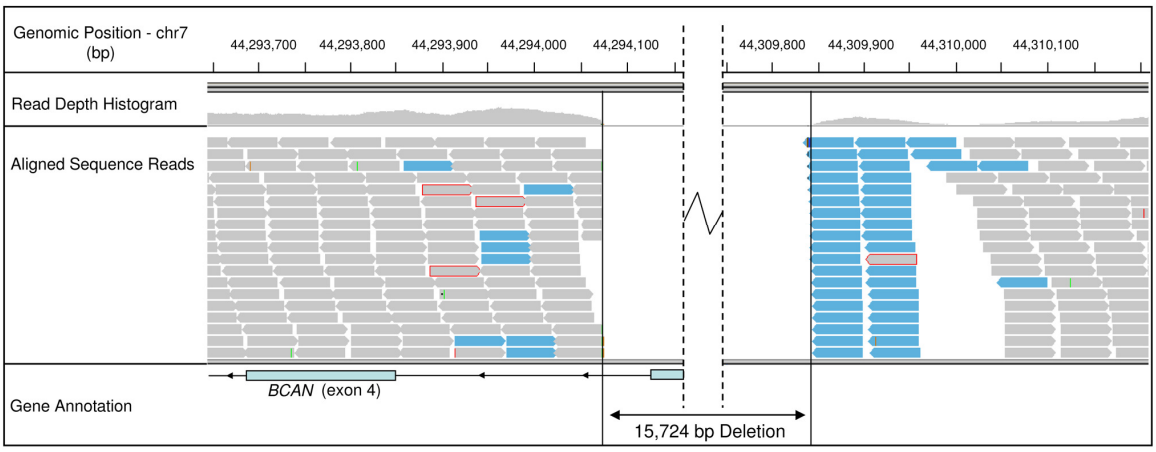


Figure 3.25 The 16 kb brevican deletion

View in IGV across the brevican deletion region. Grey bars represent individual reads. Grey bars with a red perimeter indicate reads that have not been mapped in a pair. Blue bars show read pairs with a greater than expected insert size based on selected fragment size, indicative of a large deletion.

The two EF cases were homozygous for the deletion, one CKCSID case (ID 9916) was heterozygous, and the other CKCSID case was homozygous wild-type (ID 6975). The

clinically unaffected control however was homozygous for the deletion, which was not consistent with the expected segregation pattern. In the absence of any further strong candidate mutations, the deletion was considered to be a strong potential causal variant. In order to find other potentially causal variants with a similar segregation pattern to the 16 kb deletion a second scan across the chromosome 7 target region was performed, omitting the control sample from consideration. One additional variant was identified which was predicted to cause a coding change. This was a SNP in the gene *DENND4B* (DENN/MADD domain containing 4B) which caused an arginine to a histidine amino acid substitution, which followed the same segregation pattern as the *BCAN* deletion.

3.2.11. Investigation of candidate variants

Candidate variants were assessed by genotyping the GWAS sample cohort. As the *DENND4B* mutation was identified in the Golden Retriever included as a genotyping array control and the two Italian Spinoni included for the cerebellar ataxia project (described in Chapter 4), it was considered to be a common polymorphism. The *DENND4B* SNP also showed a weaker statistical signal across a 30 EF case and 38 control cohort than the *BCAN* deletion and was ruled out as potentially causal (*DENND4B* P_{raw} = 6.81 x 10^{-11}; *BCAN* P_{raw} = 5.36 x 10^{-13}). The number of cases was reduced from 31 to 30 as a *DENND4B* genotyping result was not obtained for one EF case, and the number of successful genotypes needed to be the same for both *BCAN* and *DENND4B* to make the comparison valid. Over the sample cohort of 31 EF cases and 38 controls the P_{raw} value for the *BCAN* genotyping dataset was identical to the top statistical signal from the EF GWAS. Interestingly the *BCAN* deletion is positioned inside the region that was suggestive of a selective sweep because of the high levels of homozygosity seen in the SNP genotyping data. The *FAM83H* mutation fully segregated with CKCSID status in the cohort of 19 cases and 38 controls, producing a P_{raw} value of 2.98 x 10^{-22}, which exceeded the top statistical signal from the CKCSID GWAS of $P_{raw} = 1.2\times10^{-17}$.

To further validate the associations between the *BCAN* and *FAM83H* deletions and EF and CKCSID respectively, a panel of 308 CKCS was genotyped for both variants. This panel included the 31 EF cases, 19 CKCSID cases and 38 controls used in the GWAS analyses, and an additional 17 EF cases, 5 CKCSID cases and 198 controls. Results are shown in Table 3.6.

Table 3.6 *BCAN* and *FAM83H* genotyping results across an extended CKCS cohort

	***BCAN* Genotype**		
	(-/-)	**(-/wt)**	**(wt/wt)**
EF cases	39	3	6
EF controls	17	62	181
	***FAM83H* Genotype**		
	(-/-)	**(-/wt)**	**(wt/wt)**
CKCSID cases	24	0	0
CKCSID controls	0	38	246

In addition a panel of 341 dogs from 34 other breeds (with at least 2 dogs per breed) were assayed for both the *FAM83H* and *BCAN* mutations. All 341 dogs were homozygous wild-type for both polymorphisms. From the panel of 308 CKCS genotyped for the *BCAN* or *FAM83H* mutations, individuals that were not clinically affected, unrelated at the parent level and not related to cases at the parent level were used to estimate the mutation frequencies in the UK. From these 122 individuals, the allele frequency of both variants was estimated to be 0.08.

3.2.12. Validating mutation consequence

To validate the consequence of the two mutations, brain (cerebellum) and buccal epithelia cDNA sequencing confirmed the exon boundaries of the *BCAN* and *FAM83H* genes respectively (Genbank accession numbers JN968466–JN968467). The sequencing reads in the *BCAN* deletion region were exported and *de novo* assembled to define the exact deletion breakpoints (Figure 3.26). Assembly revealed the deletion to be 15,724 bp with a small insertion of 5 bp spanning the deletion breakpoints. The *FAM83H* single base deletion is in exon 5, and is predicted to truncate the peptide from 1,151 to 582 amino acids, with 257 aberrant amino acids at the C terminal.

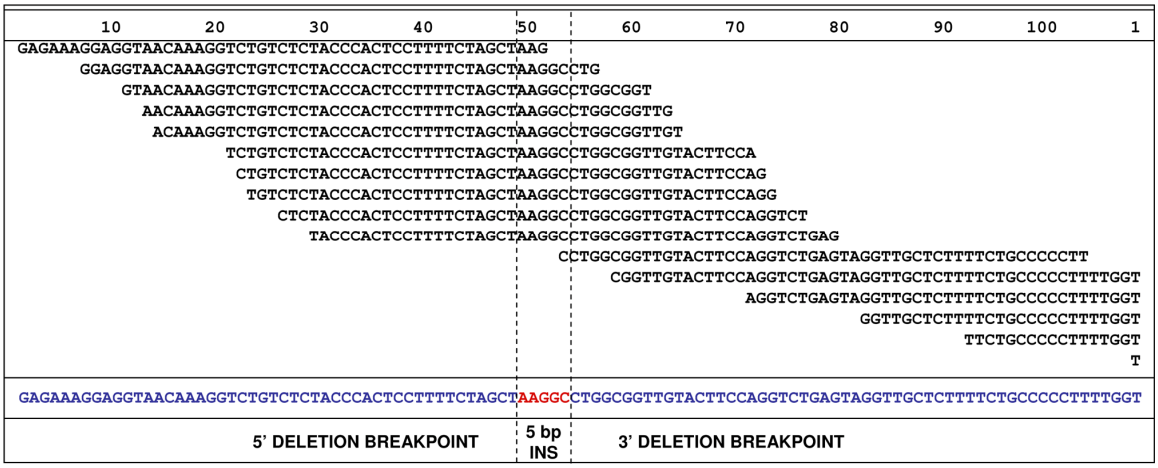


Figure 3.26 *De novo* assembly of reads across the *BCAN* deletion

Sequence reads mapping to the deletion breakpoints were exported and *de novo* assembled to define the deletion breakpoints.

3.2.13. Expression analysis

Quantitative reverse transcription PCR (qRT-PCR) was used to assess *FAM83H* and *BCAN* expression levels in canine skin and brain (cerebellum) tissues, using *ACTB* (beta actin) as a control gene. *BCAN*, *FAM83H* and *ACTB* reaction efficiencies were estimated at 97.5%, 95.7% and 94.3% respectively, with standard curve r^2 values all > 0.99. *BCAN* expression was confirmed in the brain, but was not detected in the skin. A similar level of *FAM83H* expression was detected in both skin and brain (Figure 3.27). *FAM83H* expression was also detected in footpad and buccal epithelia by RT-PCR.

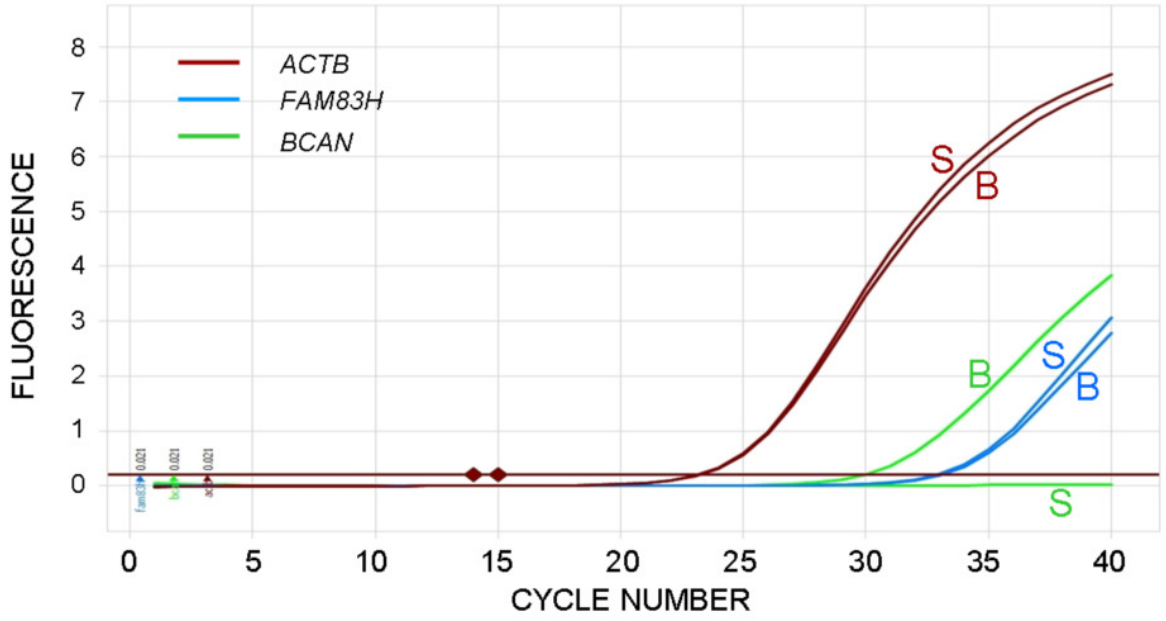
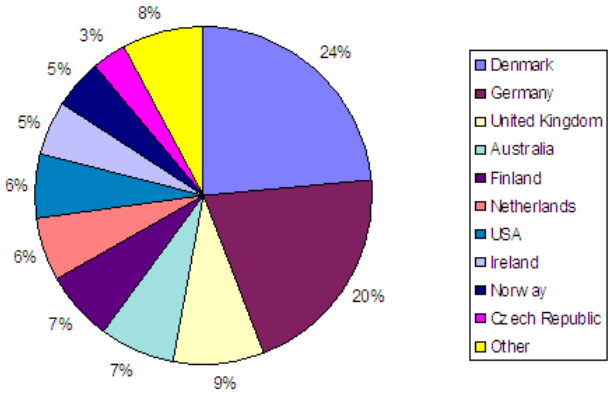


Figure 3.27 qRT-PCR assessment of *FAM83H* and *BCAN* levels

Limited expression analysis of *FAM83H* and *BCAN* using qRT-PCR, with *ACTB* as a control gene. (S = Skin; B = Brain).

3.2.14. Diagnostic DNA testing

A diagnostic DNA test to assay for both the *FAM83H* and *BCAN* mutations simultaneously was launched in April 2011, offered by the AHT Genetics Services department (www.ahtdnatesting.co.uk/). In the first year of testing, 2,457 samples were tested for CKCSID and 2,469 for EF (some owners requested individual disease testing). Dogs with two copies of the disease-associated allele were defined as affected, with one copy of the disease-associated allele as carriers, and with no copies of the disease-associated allele as normal. Four affected and 196 carrier individuals were identified for CKCSID and 96 affected and 539 carriers were identified for EF. Samples originated from 26 countries worldwide. A summary of testing is shown in Figure 3.28.



	CKCSID	EF
AFFECTED	4	96
CARRIER	196	539
NORMAL	2257	1834
Total	2457	2469

Figure 3.28 First year results of EF and CKCSID DNA testing

Ten countries submitting the most samples for EF and CKCSID testing.

3.3. Comments and conclusions

3.3.1. EF candidate gene study

Although no EF associated genes were identified, the investigation successfully ruled out many genes previously associated with hypertonicity disorders. When many strong candidate genes are available, a candidate gene study is often a logical first step to take before undertaking an expensive GWAS approach. The absence of an association with a candidate gene, may suggest that the disorder is caused by a previously unassociated gene.

3.3.2. Parallel mapping of EF and CKCSID by GWAS

The fact that EF and CKCSID were CKCS specific presented the opportunity to map both conditions in parallel using a single set of controls. This approach helped to reduce cost and decreased the overall timeframe for the studies as all DNA samples could be sent for processing on the CanineHD SNP array in a single batch. The disadvantage of the approach was that controls could not be tailored to the individual case sets. For this reason a largely unrelated set of controls was chosen, which may have contributed to the high genomic inflation values seen for both studies, indicating population stratification. Association analysis results for both studies however revealed strong statistical signals on single chromosomes, which remained after corrective analysis, indicating that the results were not largely affected by the genomic inflation, and confirmed the involvement of single, large effect, high penetrance genes, suggesting autosomal recessive inheritance. The same level of genomic inflation in the study of complex disease may have had a more significant impact on the results, where association signals on several chromosomes can be seen, and confidence in the results is crucial.

On analysis of the raw genotyping data there were clear homozygous disease-associated haplotypes for EF and for CKCSID, defined by single recombination events in two individuals, resulting in a loss of shared homozygosity. Amongst the dogs originally defined as EF cases, based on their clinical signs, there were six outliers that were not homozygous for the *BCAN* deletion, highlighting that diagnosis of the condition is not simple and that other conditions, such as epilepsy, exist in the breed which may show a similar clinical presentation.

3.3.3. Target enrichment and massively parallel sequencing

In the absence of any strong candidate genes in the EF critical region and after exclusion of *SLURP1,* a candidate gene in the CKCSID critical region, it was decided that the fastest

and most cost effective method of exploring the two gene rich critical regions would be to use a newly available solution based target enrichment system (SureSelect) and a massively parallel sequencing approach (Illumina/Solexa). The solution based target enrichment system could be performed in-house, negating the need to outsource the work.

The first stage of library preparation is the DNA fragmentation. Out of the three methods that were trialled an enzymatic method was chosen. This method appeared to give the most fragments in the desired range after processing, which is critically important especially as loss of fragments at the size selection stage may lead to additional PCR cycles being required. Any duplicates in sequencing data as a result of PCR are omitted during analysis, reducing the dataset size and eventual read depth. Enzymatic fragmentation also allowed several DNA samples to be processed in parallel, which was more time effective and ensured all samples were subjected to identical treatment.

The entire library preparation was performed in-house. Accurate sizing of libraries on an Agilent Bioanalyser was outsourced due to unavailability of resources. Average library fragment size could have been assessed by agarose gel electrophoresis, but more precise sizing can be calculated using the Bioanalyser, which is important for determining final library concentration, especially when pooling libraries for multiplexed sequencing.

Data were analysed in-house using a Perl script to run a pipeline to manipulate sequence files, align reads to the genome and make variant calls. The sequencing experiment was successful and a high level of target enrichment was achieved, with low levels of PCR duplicates. Bait coverage of target regions was limited to 65% because of the high levels of repetitive elements, particularly SINEs (short interspersed nuclear elements), present in the dog genome. The number of bases achieving at least 10x coverage was 79% however, and near complete exonic coverage of the critical region was achieved.

By browsing through read alignments to target regions in IGV, which was annotated with filtered SNP and indel positions, deletions in *BCAN* and *FAM83H* were identified as strong candidate mutations for EF and CKCSID respectively. The dataset was the first large-scale target enriched sequencing project to be processed by the NGS analysis pipeline, and highlighted areas of the analysis tool that could be improved. As both candidate mutations were identified by manual browsing, measures were subsequently made to make data processing more automated. Firstly the variant effects predictor module was added to the NGS analysis pipeline for annotation of SNP and indel calls with gene

information. Running indel calls through the predictor would have flagged the 1 bp *FAM83H* deletion as a frameshift mutation, warranting further investigation. Additionally a copy number variation (CNV) analysis Perl script was written to compare the number of aligned reads between two samples across overlapping genomic windows, and would have detected the 16 kb *BCAN* deletion without manual browsing of the sequence alignments.

3.3.4. Candidate mutations and phenotype concordance

EF has a variable phenotype in the CKCS and 17 out of 56 dogs that were homozygous for the *BCAN* deletion were reportedly not affected by EF, which suggests that the disorder may be influenced by variation in environmental stimuli and potential variants in modifier genes. As EF is an exercise-induced condition, differences in levels of activity among affected dogs may account for some of the phenotypic variation. One dog in the study that was homozygous for the *BCAN* mutation but did not display clinical signs consistent with EF was reported by its owner to be “docile and unexcitable”, suggesting, for this dog at least, insufficient environmental stimuli were provided to trigger the condition. In addition, nine out of 48 EF cases were not homozygous for the *BCAN* mutation. An extensive neurological assessment would be required to mitigate against misclassification of these cases, but this is typically not possible for the majority of canine patients due to expense or lack of owner consent. The CKCS breed is also affected by idiopathic epilepsy, and the EF episodes may often be difficult to distinguish from epileptic seizures, as a definitive diagnostic test is not available for either condition (Rusbridge, 2005). It is therefore possible that for some cases epilepsy or other neurological conditions have wrongly been diagnosed as EF, although it is formally possible that there may be a second, genetically distinct form of EF in the CKCS. The EF disease-associated region was not completely resequenced due to the repetitive nature of around 35% of the sequence, and although unsequenced regions were largely non-coding, potential causal mutations could be situated within these regions and therefore would not have been identified in the current study.

Results for *FAM83H* genotyping were fully concordant with CKCSID disease status across the extended sample cohort, showing that the CKCSID cases can be precisely diagnosed and that the condition has a simple autosomal recessive mode of inheritance, with no modifiers.

No mutant *BCAN* or *FAM83H* alleles were detected among 341 dogs from 34 other breeds. This suggests that the two mutations are limited to the CKCS breed, although only a small selection of dogs was tested from just a subset of all dog breed populations.

Additional dogs would need to be screened to formally conclude that the mutation is not present in any other breeds.

3.3.5. *BCAN*

The EF-associated gene *BCAN* encodes brevican, which is one of the central nervous system specific members of the hyaluronan-binding chondroitin sulphate proteoglycan family (Yamada et al., 1994). Brevican is important in the organisation of the nodes of Ranvier in myelinated large diameter axons (Bekku et al., 2009) and disruption of this region results in a delay in axonal conduction (Bekku et al., 2010). Interestingly the gene *HAPLN2* (hyaluronan and proteoglycan link protein 2) is tandemly arranged upstream of *BCAN*. *HAPLN2* encodes Bral1, a brain specific hyaluronan and protoglycan link protein and is co-localised with brevican and versican V2 to form complexes at the nodes of Ranvier (Bekku et al., 2010). The *BCAN* deletion moves the 3′ UTR of *HAPLN2* to within 2 kb of exon 4 of *BCAN*. Expression analysis would be required to fully establish whether the ~16 kb deletion causes a complete knock-out of the *BCAN* gene and to investigate any potential effects on *HAPLN2* expression.

Mutations in *BCAN* have not previously been associated with a disease phenotype and brevican-deficient mice are viable, fertile, physiologically normal, display normal behaviour and have a normal life expectancy (Brakebusch et al., 2002). However, the absence of any apparent abnormalities in brevican-deficient mice may relate to an absence of episode triggers within the environment the mice were maintained in. In dogs, episodes are induced by exercise or excitement and it is highly likely that mice will not exercise at a sufficiently high intensity within their routine laboratory environment.

EF is a condition that becomes self-limiting and can self-rectify in some cases, with some dogs becoming clinically normal after a period of months to years of being clinically affected. It is interesting to speculate that this might be due to compensatory effects of other chondroitin sulphate proteoglycans in the brain, in particular versican V2 (Bekku et al., 2009) taking over the role of brevican, although the effect could also be due to modified owner and/or dog behaviour in response to the episodes, such as a change in exercise levels or the avoidance of trigger events once these have been identified.

An identical mutation in *BCAN* has recently been associated with EF in the CKCS, by an independent research group (Gill et al., 2011). This independent investigation helps to further validate the association of *BCAN* with EF.

3.3.6. *FAM83H*

Several mutations in the CKCSID-associated gene *FAM83H* have been associated with autosomal-dominant hypocalcification amelogenesis imperfecta (ADHAI) in humans, which is a disease of faulty tooth enamel formation (Kim et al., 2008). To date the mutations associated with ADHAI have all been found in exon 5 of *FAM83H* and are either nonsense or frameshift mutations leading to a premature stop codon after a sequence of aberrant amino acids. Further to this, mutations in the 5′ region of exon 5 appear to result in a generalised phenotype, affecting all teeth, compared to mutations occurring in the 3′ region, which appear to give a localised phenotype, with just a subset of teeth being affected (Urzua et al., 2011). The canine mutation is at a position within the gene that would predict a more generalised phenotype. Anecdotal evidence suggests that CKCSID cases do show clinical signs of tooth disease, although this is a post-hoc observation and would require further investigation to determine the exact nature of the dental problems.

The CKCSID phenotype suggests that *FAM83H* has an important role in skin development and regulation, in addition to enamel formation, at least in the dog. Limited expression analysis has revealed that *FAM83H* is expressed in canine skin, and also in the brain (cerebellum), footpad and buccal epithelia, in concordance with previous reports that *FAM83H* may be ubiquitously expressed (Kim et al., 2008). Species-specific differences in gene expression and function have not currently been investigated and no significant skin or nail phenotypes have been associated with ADHAI in human patients. In humans all *FAM83H* mutations reported to date have been dominant, and no human patients with homozygous *FAM83H* mutations have been reported. In contrast the canine mutation is recessive and heterozygous dogs do not have a discernable phenotype, so it is interesting to speculate that the gene is playing a different role in enamel formation between the two species and that human patients may present additional phenotypes, similar to CKCSID, if a deleterious homozygous *FAM83H* mutation was identified.

3.3.7. Summary

Mutations in *BCAN* and *FAM83H* have been identified which strongly associate with EF and CKCSID respectively, using an efficient parallel mapping and simultaneous sequencing approach. Neither of the two genes had been previously associated with similar disease phenotypes in other species. The discovery of these mutations may suggest potential novel biological functions for *FAM83H* and *BCAN*, although formal proof of this would require further functional data to confirm the causality of the two mutations with respect to their associated disease phenotypes. The study illustrates how two disease phenotypes in a single dog breed can be investigated using a very modest

sample set to successfully identify disease-associated mutations, using resources available in-house where possible to maximise cost effectiveness and efficiency.

Chapter

4. Spinocerebellar ataxia in the Italian Spinone

4.1. Background

4.1.1. The Italian Spinone

The Italian Spinone (IS) is a hunting and retrieving dog, first imported into the UK in 1981. The IS was granted UK Kennel Club status in 1984, and has since gradually increased in popularity, with 441 dog registrations in 2012.

4.1.2. Spinocerebellar ataxia in the Italian Spinone

Spinocerebellar ataxia (SCA) is caused by degeneration of nerve cells in the cerebellum, brainstem and spinocerebellar tract. SCA in the IS is a progressive neurological disease characterised by hypermetria (high stepping gait or overreaching gait), particularly evident in the hind limbs (Figure 4.1A), truncal ataxia (loss of coordination), and impaired balance (Figure 4.1B). Clinical signs start to appear at three months of age and progress to a degree of dysfunction that leads to euthanasia of affected dogs at an average of one year of age. The disease shows an autosomal recessive mode of inheritance. The first and only scientific paper on SCA in the IS was published in 1996 as a brief communication giving a clinical overview of the condition (Wheeler and Rusbridge, 1996). Cases of SCA are infrequent, but have been reported in a number of countries including Italy, the UK, USA, and Germany.

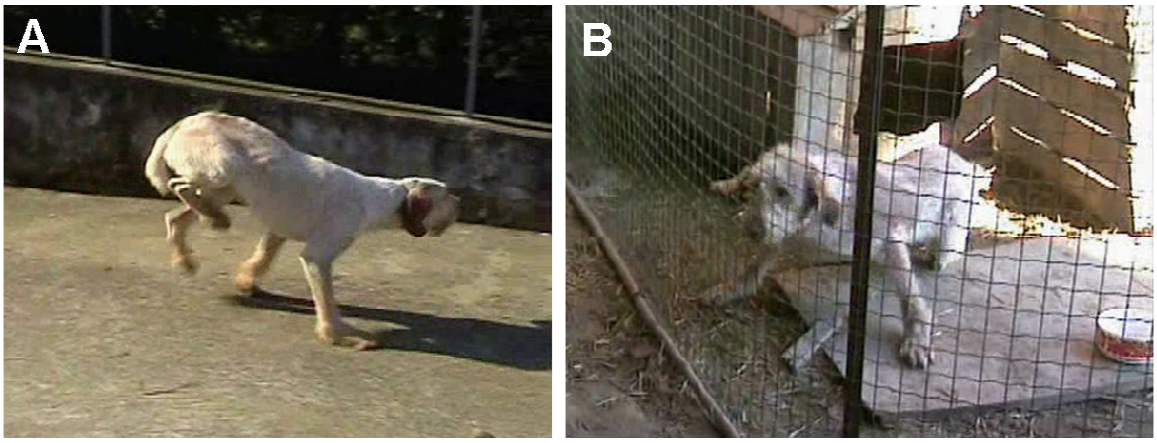


Figure 4.1 Clinical signs of spinocerebellar ataxia in the Italian Spinone

(A) Hypermetria affects all four limbs but is more obvious in the hind limbs. (B) The forelimbs adopt a wide-based stance to compensate for the impaired balance. Photos from videos provided by Dr Luisa de Risio.

4.1.3. Ataxic disorders in other breeds

Canine hereditary ataxia is widely reported in the veterinary literature and has been described in several breeds including the Kerry Blue Terrier (Darke and Kelly, 1976), the

Gordon Setter (Steinberg et al., 1981), Border Collie (Clark et al., 1982), the Portuguese Podengo (van Tongern et al., 2000), the Bullmastiff (Carmichael et al., 1983), the Rottweiler (Boersma et al., 1995), the Scottish Terrier (van der Merwe and Lane, 2001), the Papillion (Nibe et al., 2007), the Brittany Spaniel (Higgins et al., 1998), the American Staffordshire Terrier (Hanzlicek et al., 2003), the English Bulldog (Gandini et al., 2005), the Coton de Tulear (Coates et al., 2002), and the Rhodesian Ridgeback (Chieffo et al., 1994). Clinical signs are highly variable between breeds, with an age of onset that can range from neonatal, which is the case for the Coton de Tulear, through to adult onset which is the case for the American Staffordshire Terrier. Clinical signs are frequently progressive and dogs in the later stages of the disease are often euthanised on welfare grounds, as quality of life diminishes. There are no effective treatments for hereditary ataxia, although clinical signs can be stabilised in some breeds without further progression.

Ataxic disorders have also been described in the Parson Russell Terrier (Wessmann et al., 2004) and the Beagle (Kent et al., 2000), and are described in Chapters 5 and 6 of this thesis.

4.1.4. Spinocerebellar ataxia in humans

A number of distinct forms of SCA have been reported in humans, classified as SCA1 through to SCA28 for autosomal dominant forms and SCAR1 through to SCAR12 for autosomal recessive forms, based on the chronological order in which they were first characterised. Similar to canine ataxias, human ataxias are a highly heterogeneous group of disorders, which vary in age of onset, severity of clinical signs and disease progression. Early symptoms in ataxic patients may include a difficulty balancing and controlling movement. Patients often have trouble walking steadily giving them a drunken appearance. Speech and vision may be affected and patients may experience difficulty swallowing. As with canine SCA, clinical signs can be progressive, although can stabilise for some forms, with variable effects on life expectancy. Hereditary forms of ataxia are thought to affect approximately 1 in 25,000 people in the United Kingdom, with Friedreich's ataxia accounting for half of all cases, although this is not a spinocerebellar form of the disease (National Health Service 2013).

4.1.5. Mutations associated with human spinocerebellar ataxia

Mutations in 27 genes have been associated with SCA in humans, with repeat expansions, especially polyglutamate expansions, being a particularly common mutation type. Polyglutamate repeat expansions in six different genes account for six forms of human SCA (SCA1, 2, 3, 6, 7 and 17) (Higgins et al., 1996, Koide et al., 1999, Lindblad et

al., 1996, Orr et al., 1993, Riess et al., 1997, Sanpei et al., 1996). Expanded alleles become pathogenic above a particular copy number threshold, with the number of repeats often linked to the severity of clinical signs. The mechanism by which polyglutamate expansions cause ataxia is not fully understood, but it is thought that expansion may lead to protein agglutination and cellular dysfunction. Non-coding mutations have also been associated with SCA. A repeat expansion of CTG in the 3' untranslated region (UTR) of the *ATXN8OS* (ATXN8 opposite strand (non-protein coding)) gene is associated with SCA8 (Koob et al., 1999), a CAG repeat in the 5' UTR of the *PPP2R2B* (protein phosphatase 2, regulatory subunit B, beta) gene is associated with SCA12 (Holmes et al., 1999), and an intronic pentanucleotide ATTCT repeat expansion in the *ATXN10* (ataxin 10) gene is associated with SCA10 (Matsuura et al., 2000). Heterozygous gene deletions and missense mutations have also been associated with human SCA (Tonelli et al., 2006, van de Leemput et al., 2007).

4.1.6. Diagnosis of hereditary spinocerebellar ataxia

In addition to inherited forms of SCA, other common causes of ataxia in human patients include head trauma, brain tumours, cerebrovascular accidents (strokes), viral infections and multiple sclerosis. The diagnosis of a hereditary form of ataxia is therefore made by a combination of observation of the clinical signs consistent with the disorder and exclusion of other possible underlying causes. Magnetic resonance imaging of the brain may detect morphological changes associated with degeneration of the cerebellum, and post-mortem histopathological examination may confirm cerebellar degeneration. In human patients, known causal mutations may be screened in an attempt to confirm the diagnosis and to categorise the ataxia type.

4.1.7. Aims

The aims of the study were to identify the mutation responsible for SCA in the IS and to design a diagnostic test to help reduce the frequency of the disease through a controlled breeding programme.

4.2. Results

4.2.1. Genome-wide homozygosity mapping

Six cases and six controls (obligate carriers) were used to perform genome-wide homozygosity mapping using 300 microsatellite markers (as the project commenced prior to the availability of the first canine high-density SNP genotyping array). Because of the small number of cases and controls, results were analysed visually and by comparing allele frequency between the case and control groups.

On visual inspection microsatellite markers C20.374 and Ren124F16 showed a pattern suggestive of linkage with SCA (Table 4.1). Genotypes for the two markers were all homozygous for the same allele in cases and either heterozygous or homozygous in controls. The pattern across the two markers was suggestive of a shared homozygous haplotype across all cases, with controls (obligate carriers) having one copy of the shared haplotype, consistent with the segregation pattern expected for an autosomal recessive trait.

Table 4.1 Markers initially suggestive of linkage to SCA in the IS

Alleles of markers C20.374 and Ren124F16 gave results suggestive of linkage to SCA in the IS. Cases were homozygous for a shared haplotype (A) and controls (obligate carriers) were heterozygous for the same shared haplotype (B). Homozygous genotypes are highlighted in blue.

(A)

Marker	Position	Cases ID number					
	(Chromosome:Mb)	5357	5397	5404	6422	6477	6685
C20.374	20:15.60	191/191	191/191	191/191	191/191	191/191	191/191
Ren124F16	20:21.93	233/233	233/233	233/233	233/233	233/233	233/233

(B)

Markers	Position	Control ID number					
	(Chromosome:Mb)	5297	5298	5405	5407	5436	6478
C20.374	20:15.60	187/191	191/191	182/191	182/191	191/191	191/191
Ren124F16	20:21.93	233/238	238/240	233/233	233/233	233/238	233/238

4.2.2. Linkage analysis

Using C20.374 and Ren124F16 and additional microsatellite markers in the surrounding region, linkage analysis was performed by genotyping 60 individuals of an extended pedigree comprising a total of 13 SCA cases and 47 controls. Two point linkage analysis was performed using MLINK, with marker C20.374 giving a maximal LOD score of 4.41 (theta = 0), which provided significant evidence of linkage to chromosome 20 (Figure 4.2).

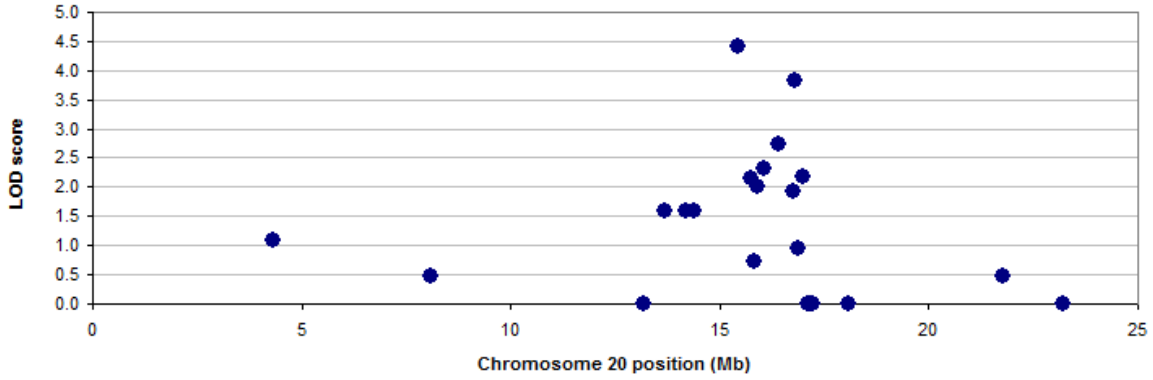


Figure 4.2 Plot of chromosome 20 LOD scores

LOD scores for markers on chromosome 20 genotyped across an extended IS pedigree calculated with a theta value of zero. The maximal LOD score was 4.41 at 15.60 Mb.

4.2.3. Fine mapping

Upon confirmation of linkage to chromosome 20, fine mapping was performed using additional microsatellite and SNP markers to define the disease-associated haplotype. The 13 cases were genotyped for a total of 40 microsatellites and 15 SNPs, to define the boundaries of the disease-associated haplotype as chr20:15.60-17.14 based on the genomic coordinates of the CanFam2 genome assembly, a region of 1.54 Mb. Genotypes of the two boundary defining individuals are shown in Table 4.2.

Table 4.2 Boundary defining genotypes for the SCA disease-associated region

Genotypes of two cases defined the boundaries of the disease-associated haplotype. Heterozygous genotypes are shown in blue. All 13 cases were homozygous for the disease-associated interval defined as chr20-15.60-17.14. Non-determined alleles are marked as n.d.

Marker	Position (canfam2)	4950		5404	
CFA20_13.17	13.32	247	245	247	247
C20_13.53	13.68	204	212	204	204
CFA20_13.70	13.85	240	240	240	240
CFA20_14.18	14.33	208	208	208	208
CFA20_14.27	14.42	279	279	279	279
CFA20_14.34	14.49	203	207	203	203
CFA20_14.38	14.53	134	134	134	134
C20_14.44	14.59	230	230	230	230
CFA20_15.03	15.18	162	157	162	162
CFA20_15.13	15.28	131	141	131	131
CFA20_15.27	15.42	261	267	261	261
CFA20_15.37	15.52	239	233	239	239
CFA20_15.39	15.54	167	169	167	167
C20.374	15.60	191	187	191	191
CFA20_15.59	15.74	285	285	285	285
CFA20_15.60	15.75	204	204	204	204
CFA20_15.74	15.89	177	177	177	177
CFA20_15.82	15.97	n.d.	n.d.	183	183
CFA20_15.90	16.05	n.d.	n.d.	328	328
CFA20_16.06	16.21	246	246	246	246
CFA20_16.42	16.57	n.d.	n.d.	164	164
C20_16.76	16.91	224	224	224	224
CFA20_16.77	16.92	155	155	155	155
CFA20_16.80	16.95	201	201	201	201
CFA20_16.82	16.97	n.d.	n.d.	299	299
CFA20_16.86	17.01	n.d.	n.d.	315	315
CFA20_16.89	17.04	n.d.	n.d.	289	289
SNP9	17.14	n.d.	n.d.	C	G
SNP8	17.15	n.d.	n.d.	G	A
CFA20_17.01	17.16	n.d.	n.d.	320	320
SNP5	17.19	n.d.	n.d.	T	C
SNP3	17.21	n.d.	n.d.	T	C
SNP2	17.22	n.d.	n.d.	A	G
SNP2	17.22	n.d.	n.d.	C	T
SNP1	17.23	n.d.	n.d.	T	A
C20_17.09	17.24	n.d.	n.d.	245	237
CFA20_17.16	17.31	n.d.	n.d.	132	134
CFA20_17.24	17.39	297	297	297	303
CFA20_18.07	18.22	129	129	129	135

4.2.4. Gene sequencing

The disease-associated region contained the five genes *BHLHE40* (basic helix-loop-helix family, member e40), *ITPR1* (inositol 1,4,5-trisphosphate receptor, type 1), *SUMF1* (sulfatase modifying factor 1), *SETMAR* (SET domain and mariner transposase fusion gene) and *LRRN1* (leucine rich repeat neuronal 1). The exons and 500 base pairs upstream and downstream of each gene were Sanger sequenced in two cases, two obligate carriers and a clinically normal Miniature Long Haired Dachshund as an additional control. The sequencing data generated were screened for correctly segregating polymorphisms (ie homozygous non wild-type in cases, heterozygous in obligate carriers and homozygous wild-type in the control).

4.2.4.1. *BHLHE40*

Seven SNPs were identified within *BHLHE40*, including four in the 5' UTR, one in intron 3, one in exon 4 and one in exon 5. None of the variants fully segregated with disease status and were therefore not potentially causal.

4.2.4.2. *ITPR1*

The *ITPR1* gene is a strong candidate for the cause of SCA in the IS as mutations in this gene have been associated with SCA15 in humans (Hara et al., 2008). Through Sanger sequencing a total of 61 polymorphisms were identified in the gene. Twelve exonic SNPs were identified, including one non-synonymous base change causing a glutamine to glutamic acid substitution. None of the SNPs across the gene segregated appropriately with disease status and were therefore not potentially causal.

One polymorphic microsatellite was located in a highly conserved region of the 5' UTR of *ITPR1* (Figure 4.3). Alleles of the variant segregated consistently with disease status for the five samples initially sequenced. Sequencing of additional obligate carriers revealed individuals that were homozygous for the disease-associated allele, indicating its presence on a normal haplotype and ruling out the microsatellite as the causal mutation.

```
HSA  1    GTAACCATGTGGATGTGCTGCTGAAGCGTTTCCTCAAGCTCGCTGGGGTGGGAGGAGAGG  60
          ||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||
CFA  1    GTAACCATGTGGATGTTCTGCTGAAGCGTTTCCTCAAGCTCGCTGGGGTGGGAGGAGAGG  60

HSA  61   AGGAGGAGGAGGTGGTGGTGGAGGAGGAGGCAGGGGGTG----------GAGAGAGAG  108
          ||||||  ||   |||||| ||||||||||||||| |||||          |||||||||
CFA  61   AGGAGGTGG---TGGTGGAGGAGGAGGAGGCAGAGGGTGGAGAGAGAGAGAGAGAGAG  117

HSA  109  AAAGCGCACGCCGAGAGGAGGTGTGGGTGTTCCGCTTCCATCCTAACGGAACGAGCTCCC  168
          ||||||||||  |||||||||||||||||||| |||||||||||||||||||||||||||
CFA  118  AAAGCGCACGCACAGAGGAGGTGTGGGTGTTCCGTTTCCATCCTAACGGAACGAGCTCCC  177
```

Figure 4.3 A conserved section of the 5' UTR of *ITPR1* containing a microsatellite sequence

Human (HSA) to dog (CFA) alignment of a conserved section of the 5' UTR of *ITPR1* containing a short microsatellite sequence (highlighted in blue). An allele of the microsatellite initially segregated with disease status across the five sequenced individuals. The variant was ruled out as causal by additional sequencing.

4.2.4.3. *SUMF1*

Eleven sequence variations were found including two synonymous exonic, six intronic and three 5' UTR SNPs. An intronic polyT variation was also seen. None of the alleles segregated consistently with disease status.

4.2.4.4. *SETMAR*

In *SETMAR* two SNPs and a mononucleotide repeat variation were seen, all in the 5' UTR and none segregating consistently with disease status.

4.2.4.5. *LRRN1*

One SNP, two microsatellite repeats and a homopolymer repeat sequence variant were found in the 5' UTR, which were ruled out as potentially causal.

4.2.4.6. Other predicted genes in the disease-associated region

Two other genes were predicted by Ensembl to be in the disease-associated region (*SAP18-like* and *XP_533756.1*). Supporting evidence for the *SAP18-like* gene was weak, as the human orthologue is found on chromosome 13 rather than the syntenic region on chromosome 3 (Figure 4.4). The gene was however sequenced for completeness. Sixteen sequence polymorphisms were found, but no alleles fully segregated with disease status. *XP_533756.1* was a predicted gene in Ensembl, but was later reclassified as a pseudogene (*CWC15-like*). The pseudogene was sequenced for completeness but no polymorphisms were identified.

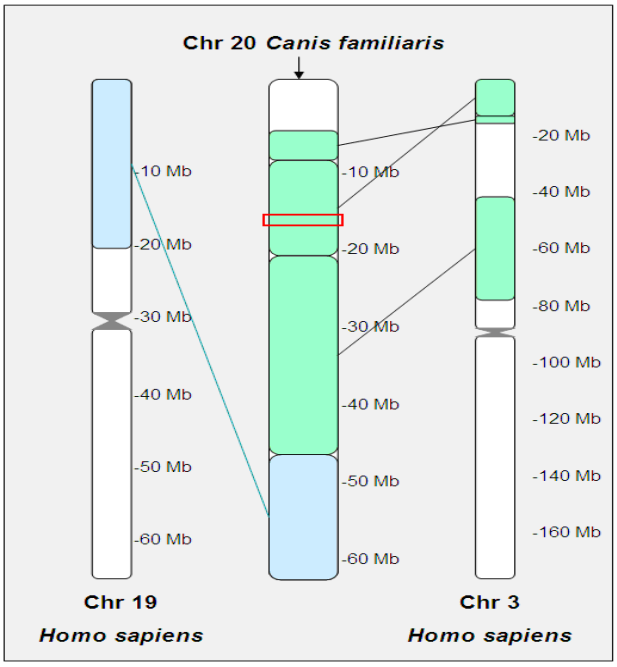


Figure 4.4 Human syntenic regions of canine chromosome 20

Canine chromosome 20 is syntenic to regions of human chromosomes 3 and 19. The SCA disease-associated region, indicated by the red marker, is syntenic to a region on human chromosome 3.

4.2.5. RNA sequencing

4.2.5.1. *ITPR1* mRNA sequencing

The transcript sequence of the top candidate gene, *ITPR1*, was confirmed by sequencing mRNA from canine cerebellum tissue. Comparison of the human *ITPR1* gene and the canine Ensembl prediction suggested *ITPR1* to have an additional untranscribed upstream exon. BLAST searching of the additional human exon against the dog genome revealed a region on canine chromosome 20 showing 91% sequence similarity in the expected position upstream of the first exon of the Ensembl prediction. The exon was also confirmed by mRNA sequencing.

Figure 4.5 provides an overview of the key features of the *ITPR1* transcript expressed in the cerebellum. The translational start point is located in the third exon, as the gene has two untranslated exons. An isoform exists with skipping of exon 12. Additionally there are splice variants for exon 41, producing two versions of the exon with a 3 base pair difference, equating to loss or gain of a glutamine residue. The stop codon is located in exon 62.

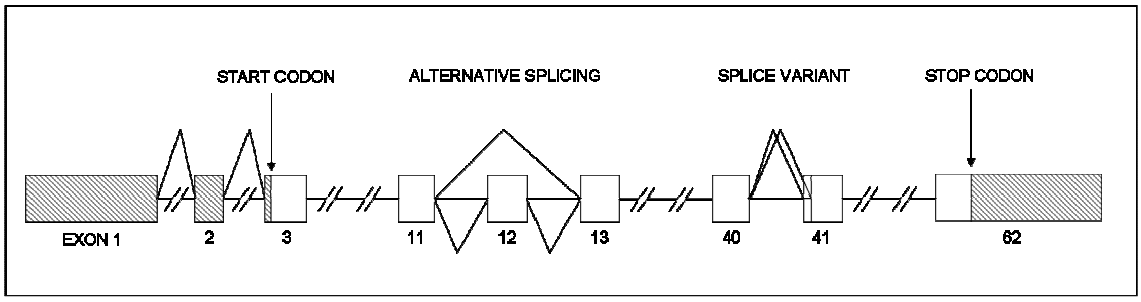


Figure 4.5 Canine *ITPR1* transcript structure

Features of the canine cerebellum *ITPR1* transcript. Untranslated regions are shaded in grey.

4.2.5.2. Non-quantitative assessment of critical haplotype gene expression.

Expression of other genes across the region were assessed in a non-quantitative approach using reverse transcription PCR (RT-PCR). Primers were designed to span across the entire genes, from first to last exons to confirm expression and transcript length. All genes showed some expression in the cerebellum apart from *SUMF1*. PCR products of expressed genes were subsequently sequenced, confirming Ensembl gene predictions and intron exon boundaries for *BHLHE40*, *SETMAR*, *SAP18*-like and *CWC15-like*. An additional untranslated upstream exon was identified for the *LRRN1* transcript. PCR products are shown in Figure 4.6.

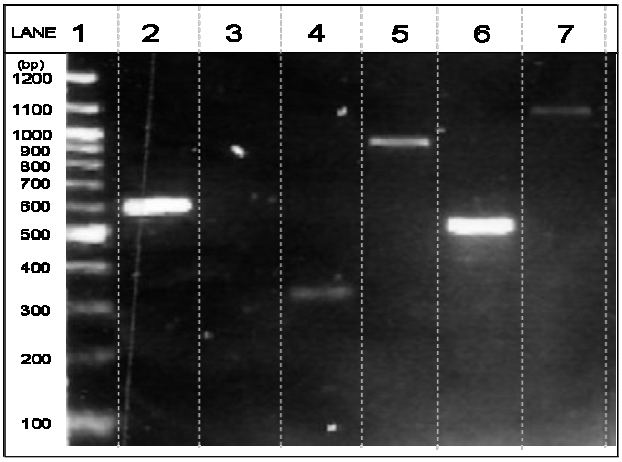


Figure 4.6 Non-quantitative assessment of critical haplotype gene expression by RT-PCR

PCRs were designed across *BHLHE40*, *SUMF1*, *SAP18-like*, *SETMAR*, *CWC15-like*, and *LRRN1* to assess gene expression, with products shown in lanes 2 to 7 respectively.

4.2.5.3. Using mRNA-seq to assess gene expression

The availability of mRNA-seq data from experiments detailed in Chapter 6 provided the opportunity to scan the SCA disease-associated region and confirm gene expression in the cerebellum. The data provided good evidence that all genes listed were expressed to some extent across the region, including *SUMF1*, for which PCR results suggested no expression in the cerebellum. A summary of genes expressed in the disease-associated region is shown in Table 4.3.

Table 4.3 Expressed genes across the SCA disease-associated region

Expression levels of the genes within the disease-associated interval based on data from genome-wide mRNA-seq experiments

Gene	Position (canfam2)	mRNA-seq approximate read depth
BHLHE40	15,643,983-15,651,394	30
ITPR1	15,748,745-16,044,359	120
SUMF1	16,095,419-16,192,508	50
SAP18-like	16,159,039-16,159,395	30
SETMAR	16,213,239-16,231,854	10
CWC15-like	16,345,948-16,346,727	100
LRRN1	16,618,166-16,653,827	15
Unknown transcript	16,849,315-16,854,497	5

Although there were sequence alignments for the genes *SAP18*-like and *CWC15*-like, mapping quality of the reads to the reference sequence were indicated as zero. A mapping quality value of zero indicates that the alignment is not unique and that the read sequence is present elsewhere in the genome. This gives an indication that *SAP18-like* and *CWC15*-like may be pseudogenes. Displays of mRNA-seq data across genes in the region is shown in Appendix 10.

4.2.6. Illumina sequencing of the *ITPR1* gene

4.2.6.1. Experiment design

As the causal mutation had not been identified by exon resequencing of all the genes in the region, the causal mutation was assumed to be non-coding or in an uncharacterised gene. Given the association of *ITPR1* with SCA15 and SCA16 in humans and also ataxia in mice (Iwaki et al., 2008, van de Leemput et al., 2007), it was decided to sequence the entire gene including introns and 10 kb of upstream and downstream sequence using newly available massively parallel sequencing technology. As this was the first massively parallel sequencing experiment to be undertaken at the AHT it was decided to use a single SCA case for template generation and use the experiment as a pilot study. Polymorphisms would be considered as potentially causal if they were in conserved regions of the genome.

4.2.6.2. Template generation

The work commenced in March 2008 prior to the launch of commercial sequence capture methodology, so long range PCR was chosen as a method of template generation. In total 60 PCRs were required to span the *ITPR1* gene and 10 kb of upstream and downstream sequence. PCR products were resolved by agarose gel electrophoresis (Appendix 11). The total combined PCR product length was 376,356 bp, giving a contig length of 339,597 bp with no gaps (chr20:15,738,385-16,077,981).

4.2.6.3. Illumina sequencing and analysis

Library preparation and sequencing were performed by Fasteris Life Science, Switzerland. Sequencing was performed on the Illumina GAII, generating 1,467,675 reads of 33 bp in length. Reads were aligned to CanFam2 using the Linux based software package Maq. On alignment 88.3% of reads mapped to chromosome 20, giving an average read depth of 124, with 98.9% of bases achieving at least 10x coverage. *De novo* assembly was performed using EDENA software at Fasteris Life Science. EDENA software assembled 70.7% of the reads into 291 contigs when the minimum read overlap was set to 25. The total sequence assembled by EDENA was 350,854, with 256,084 mapped by BLAST to the target region of chromosome 20, equating to 75.4% base coverage.

Average read depths were calculated across 20 bp windows in Maq, to look for any spikes in read depth that could indicate small copy number variation. Datapoints were plotted in Excel and an example is shown in Figure 4.7. Although read depth across the target region was too variable to consider small copy number variations, a doubling of read depth could clearly be seen in positions of overlap between PCR products.

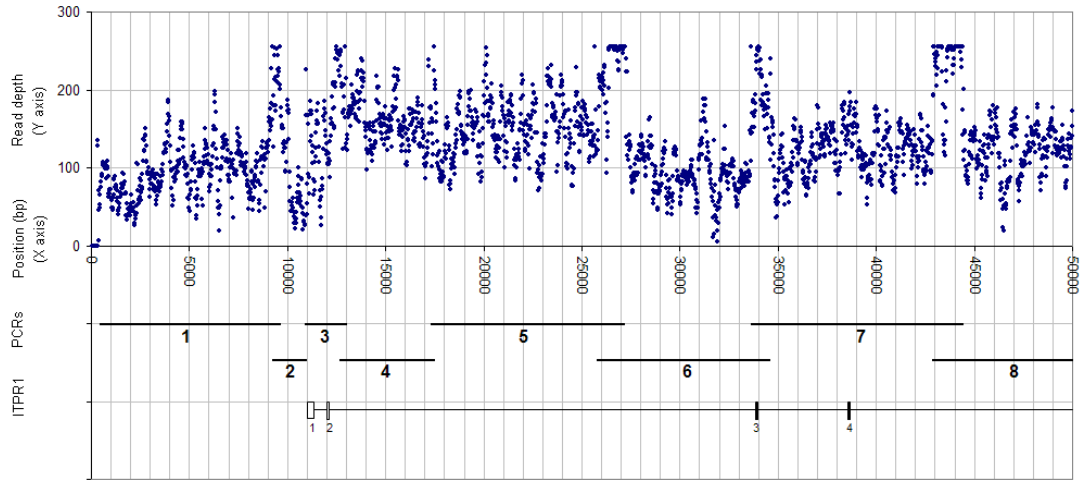


Figure 4.7 Read depth over 20 bp windows across the *ITPR1* target region

No obvious copy number variations were noted across the *ITPR1* gene, although a doubling of read depth was noted in the region of overlap between PCR products.

The consensus sequence of reads aligned to the canine reference and *de novo* assembled contigs, were aligned together using the Staden Gap software package. Capillary sequence reads across the gene were also aligned to assess the accuracy of the sequencing technology.

A total of 1,018 variants were called by Maq across the target region, including 615 SNPs and 469 indels, although the number of indels is likely to be an overestimation due to calling inaccuracies in early massively parallel sequencing analysis tools. Six short interspersed nuclear elements (SINEs) were identified by visually browsing contig alignments to the consensus sequence in Staden Gap. The positions of all variants were assessed for sequence conservation by viewing in the UCSC genome browser. Eight SNPs in semi-conserved regions and all identified SINEs, detailed in Table 4.4, were investigated further by Sanger sequencing five additional samples including one case, two obligate carriers and two predicted wild-type individuals based on haplotype analysis. None of the polymorphisms segregated with disease status.

Table 4.4 Variants in conserved regions of *ITPR1* for further investigation

Variants across the *ITPR1* gene were assessed for sequence conservation across species, and sequenced across additional cases and controls. None of the variants fully segregated with disease status.

Genomic region	Feature
15747434-15747873	SNP
15807374-15807822	SNP
15821684-15822358	SNP
15883508-15883914	SNP
15888739-15889265	SINE insertion
15901797-15902333	Insertion
15919562-15920003	2 bp deletion
15930447-15931004	SNP
15952316-15952708	SNP
15955174-15955544	SINE insertion
15990820-15991263	SINE insertion
16005179-16005716	9 bp deletion
16010817-16011268	SINE deletion
16011285-16011680	SINE insertion
16015255-16015700	SINE deletion

4.2.6.4. Investigation of a possible repeat expansion

As the sequenced region was manually browsed a large intronic tetranucleotide repeat sequence was identified between exon 5 and 6 of *ITPR1* (47 perfect repeats, CanFam2 chr20:15,941,265-15,941,537). Given the association of repeat expansions with human SCA, the repeat was assessed by genotyping across two cases, two obligate carriers and two wild-type individuals. The cases were both homozygous for a 454 bp amplicon, obligate carriers were heterozygous (454,484 and 454,447) and the two wild-type controls were homozygous 484 and heterozygous 473,476. As the disease-associated allele was not expanded, the variant could be ruled out as causal of SCA.

4.2.7. Copy number investigation using the CanineHD SNP array

Two DNA samples were submitted for genome-wide SNP analysis as a means of detecting potential copy number variations across the disease-associated region. The genotyped individuals were an obligate carrier and the 3' boundary defining case of the disease-associated interval (ID 5404 in Table 4.2). Based on marker BICF2P939857 being heterozygous in the case, the 3' boundary could be refined as chr20:17,116,778.

In Genome Studio, individual SNP B allele frequencies and log R ratio values were analysed to assess the disease-associated region for possible copy number variations. B allele frequencies of 1 and 0 signify homozygous genotypes and frequencies of 0.5 signify

heterozygous genotypes. The point of recombination in the genotyped case can be clearly defined by B allele frequency, due to loss of homozygosity (Figure 4.8 A). Log R ratios show the relative signal intensity of each genotyped SNP. Increased or decreased log R ratios may indicate duplications or deletions respectively. Log R ratios for the SNPs across the region stayed close to 0 for both the case (5404) and obligate carrier (2275) and suggested that there is no copy number variation across the region (Figure 4.8 B).

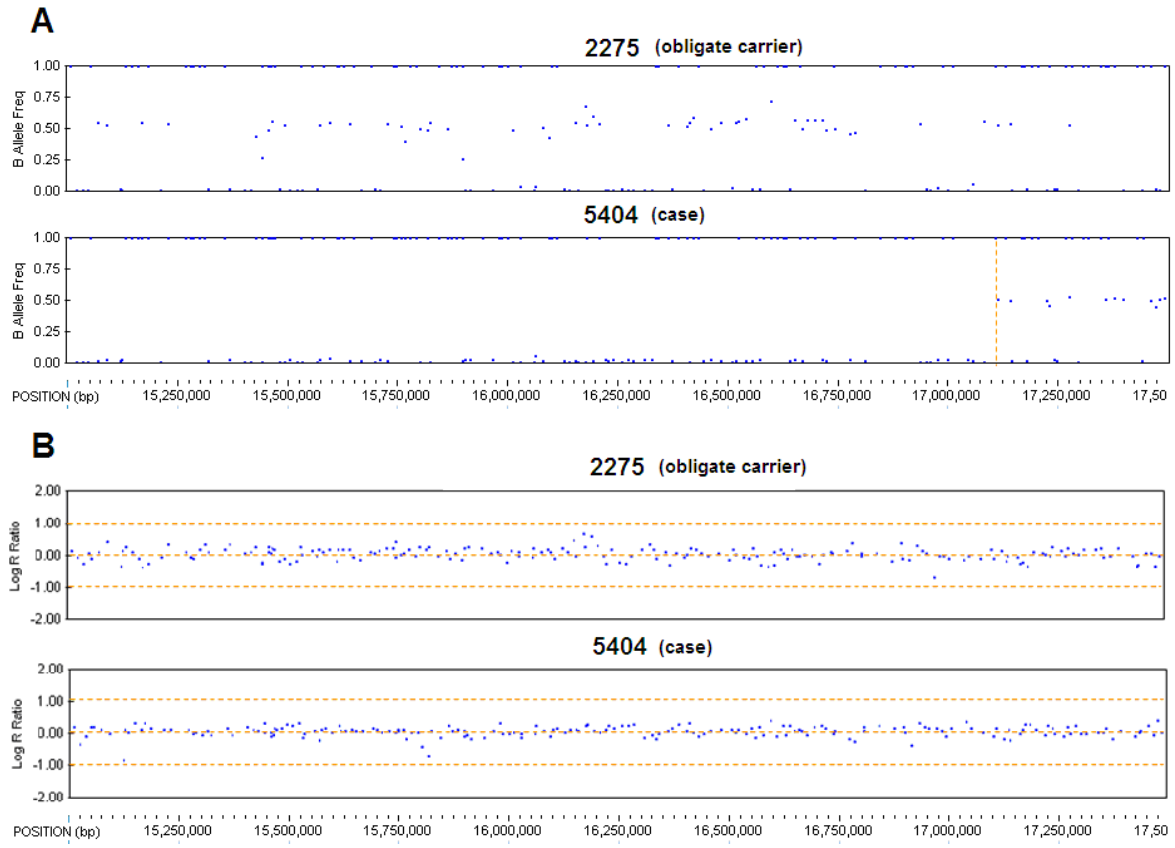


Figure 4.8 Assessing copy number by log R ratio and B allele frequency

Individual SNP B allele frequencies (A) and log R ratios (B) were used to assess potential copy number variation across the disease-associated region. Individual SNPs are marked as blue dots. A shift in log R ratio from zero to one would indicate a duplication. For the obligate carrier, heterozygous SNPs located within a heterozygous duplication would have a B allele frequency of 0.33 or 0.66.

4.2.8. Illumina sequencing of the disease-associated interval

4.2.8.1. Experiment design

With the availability of target enrichment methodology, a second resequencing experiment was performed across the entire disease-associated interval. As the locus for SCA in the Parson Russell Terrier (PRT) had also been mapped (Chapter 5), the two regions could be captured in parallel using a single combined probe set. A summary of probe design is shown in Table 4.5. Probes achieved a base coverage of 2.21 Mb (56.7%) after repeat masking.

Table 4.5 Summary of target intervals for target enrichment of SCA regions

Breed	Chromosome	Position	Total baits
Parson Russell Terrier	18	53,533,360-55,860,486	20535
Italian Spinone	20	15,601,140-17,116,778	12685

4.2.8.2. Sample selection

Two IS SCA cases and three controls (obligate carriers) were chosen for sequencing. Obligate carriers that had different wild-type haplotypes across the associated region were chosen, to aide exclusion of candidate causal variants. Five PRT samples were chosen for parallel sequencing (Chapter 5). These samples acted as additional controls for the IS SCA study.

4.2.8.3. Library preparation

Library preparation was adapted to enable buccal DNA to be used. Buccal swab DNA is often highly sheared, so an alternative to enzymatic fragmentation was needed (enzymatic fragmentation would have resulted in too many unusable small fragments). As the recommended method of fragmentation in the SureSelect protocol was Covaris shearing, and a Covaris shearing service had become available since the CKCS targeted sequencing experiment, this method was chosen for fragmentation.

4.2.8.4. Illumina sequencing

Illumina sequencing was performed at the Wellcome Trust Centre for Human Genetics, University of Oxford. Paired-end sequencing with 51 bp read lengths was performed on a single lane of a HiSeq 2000 generating a 19.64 Gb dataset. A summary of the dataset is shown in Table 4.6.

Table 4.6 Summary of the SCA targeted sequencing dataset

Individual ID	No. of reads (Millions)	Dataset size (Gb)	Target enrichment efficiency (%)	Bases achieving 10x Coverage (%)	PCR duplicates (%)
2275	36.5	1.86	77.9	78.4	38.8
5404	41.2	2.10	78.6	79.6	40.9
5405	30.9	1.58	77.6	78.8	35.2
6479	35.3	1.80	78.3	78.9	35.6
6685	35.9	1.83	78.3	79.0	35.4
12015	38.4	1.96	77.0	79.2	35.5
12205	34.9	1.78	77.5	79.6	43.7
12898	37.9	1.93	76.0	79.1	36.5
12947	33.8	1.72	76.6	78.7	32.1
18500	33.6	1.71	76.6	78.7	37.9

4.2.8.5. Data analysis

4.2.8.5.1 SNP and indel analysis

Data were processed in Linux using the NGS pipeline (described in Chapter 3). SNP and indel calls were aligned across all individuals using the SNP Handler Excel macro developed by Dr Mike Boursnell at the AHT. Compiled SNPs were annotated with gene consequence predictions and assigned a segregation score based on how well SNPs segregated with disease status. In total 4,495 SNPs and indels were identified across the SCA disease-associated region. Fifty-two variants segregated appropriately with disease status ie homozygous in the two cases and heterozygous in the three obligate carriers. Variants for which the disease-associated allele was the same as the reference allele were excluded, leaving eight variants. Data from the PRTs were used for further filtering. As five of the disease-associated SNPs were present in some or all of the PRT individuals this left two SNPs and a 1 bp insertion to consider as potentially causal for SCA in the IS.

The two SNPs were intergenic and in non-conserved regions of the dog genome. The 1 bp insertion however was within a region of high conservation and worthy of further investigation (Figure 4.9). The insertion was located approximately 2 kb downstream of the 3' UTR of *BHLHE40*. The syntenic region of the human genome has been shown to be a major transcription factor binding site through ChIP-seq experiments indicating that the sequence could be part of an important regulatory element. The indel was investigated by genotyping in 149 Italian Spinoni previously genotyped for the two microsatellite markers used for linkage-based diagnostic testing (see section 4.2.9). The results from the 1 bp insertion and microsatellite markers were concordant apart from three dogs, in which the 1 bp insertion was heterozygous, suggestive of a carrier status, and the genotypes of the two microsatellites were suggestive of a normal status. All three

individuals shared a rare non-disease-associated microsatellite allele for one of the diagnostic markers, suggesting the insertion, in addition to being disease-associated, was also present on a rare normal haplotype, indicating that it was therefore not causal of SCA.

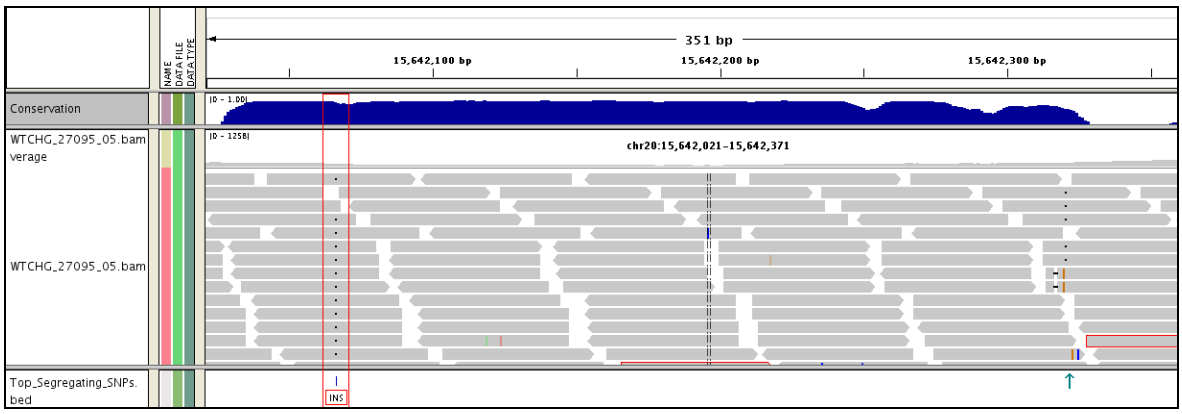


Figure 4.9 Insertion in a conserved region of the SCA associated interval

A 1 bp insertion which segregated with disease status across the individuals selected for target enrichment was identified in a conserved part of the disease-associated region (highlighted in red). An additional 1 bp insertion was also identified 300 bp downstream, indicated by the green arrow. This insertion did not fully segregate with disease status and was therefore excluded.

4.2.8.5.2 Copy number analysis

A high resolution scan for copy number variations was performed by comparing read count between cases and controls over a sliding window of 10 kb. As SNPs on the CanineHD array are approximately every 14 kb an approach using read counting increased resolution and ensured the region was evenly investigated. Results are displayed in Figure 4.10. The ratio of case read count to control read count remained very close to 1 across the captured region, suggesting no copy number variation. Analysis across a 5 kb window also suggested no copy number variations, although read depth ratio did show more deviation from 1:1. On visual analysis, this was due to very low read count over some 5 kb windows, rather than genuine differences in copy number between the case and the control. The inability to scan for copy number over smaller sliding windows was a consequence of the repeat masked SureSelect bait design, resulting in lack of coverage across some regions. Approaches producing more even coverage, such as whole genome sequencing would allow for higher resolution analyses for copy number, but would also be dependent on read depth as well as overall base coverage.

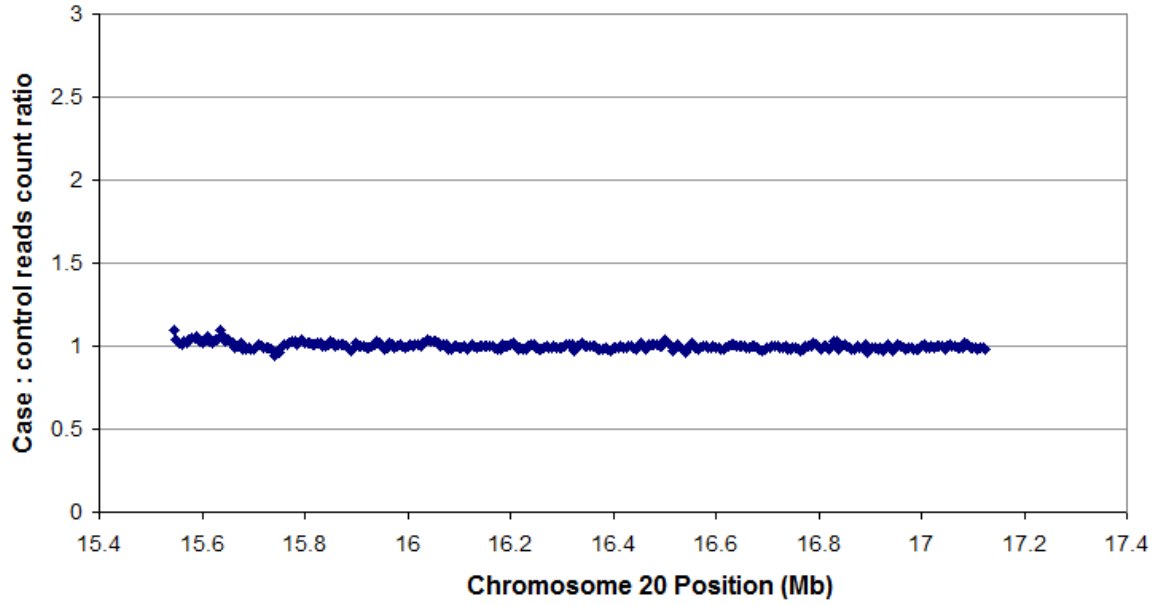


Figure 4.10 Case versus control read count comparison

Comparison of case and control read counts across overlapping 10 kb windows. Ratios of 1:1 suggest no copy number variation between the case and control.

4.2.8.5.3 Other variants in the target enriched sequencing data

The sequencing data were visually scanned in IGV to search for other variants that could be considered causal for SCA in the IS. SINEs detected during the *ITPR1* resequencing experiment were confirmed manually by visualising data in IGV. Additional variants not picked up by automated variant calling programs were also seen and are summarised in Table 4.7.

Table 4.7 Additional variants across the SCA critical region

Genomic region	Feature	Correct segregation pattern
chr20:15,773,266-15,773,280	14 bp deletion	No
chr20:15,882,747-15,883,097	CTT repeat expansion in SINE	Yes
chr20:15,914,340-15,914,514	SINE deletion	No
chr20:16,104,369-16,104,411	16 bp insertion	No
chr20:16,125,448-16,125,698	SINE deletion	No
chr20:16,133,144-16,139,501	LINE deletion	No

4.2.8.5.4 Triplet repeat expansion identification

The only additional variant to segregate with disease status was a GAA triplet repeat expansion. On visual scanning of the sequence read alignments, a 30 bp region showing variable coverage between cases and controls was identified (Figure 4.11). In this region read depth dropped to zero for cases but remained in all three controls, although the low coverage in the controls meant that the profile of the coverage histogram appeared

comparable for all individuals. The region that was unsequenced in the cases was located in intron 35 of *ITPR1* and consisted of a poly(T) tract which was 5' of a SINE element and 3' of a short trinucleotide repeat sequence $TTC_{(8)}$. Just upstream of the region many singleton reads (reads lacking an aligned mate) were present. Further investigation revealed the sequence of all the unaligned mates to be $GAA_{(17)}$ suggesting expansion of the trinucleotide repeat sequence in the cases was potentially responsible for the region's failure to sequence in the DNA from these dogs.

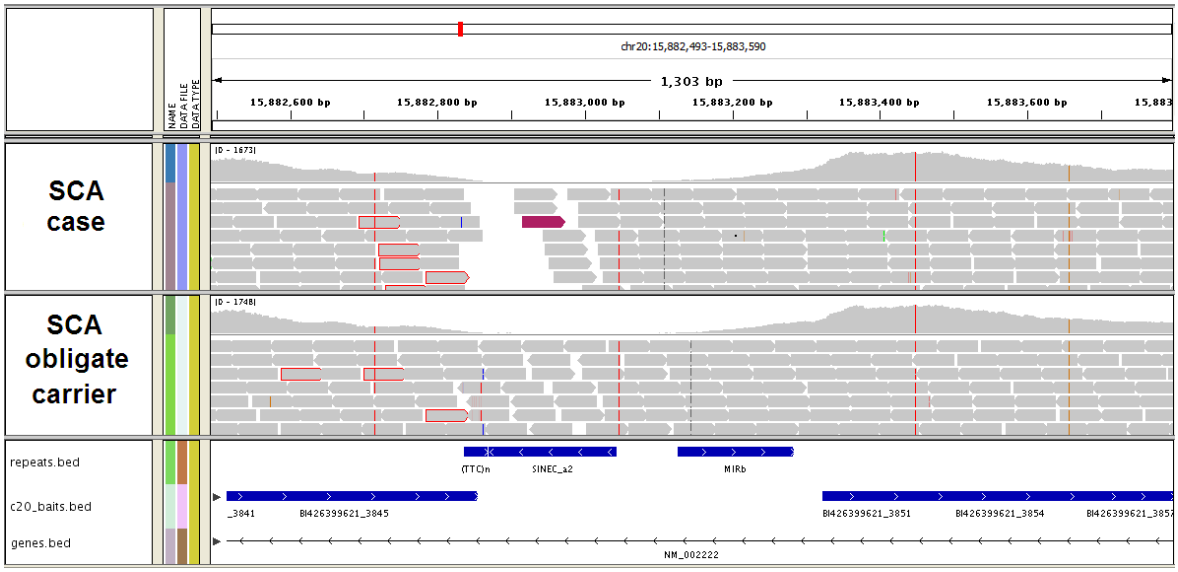


Figure 4.11 GAA repeat expansion in *ITPR1* intron

A GAA repeat expansion was identified in the target region, flanking a SINE. In the figure the position of capture baits, and repeat elements are highlighted by the blue bars. Grey reads with a red perimeter indicate reads with unaligned mates.

The GAA repeat expansion was further investigated by PCR to genotype all individuals of known status, ie 13 cases and 16 obligate carriers, to assess whether repeat expansion could be causal. All 16 obligate carriers were heterozygous for the repeat expansion. Seven of the thirteen cases produced PCR results (results could not be obtained from DNA extracted from FFPE tissue). Each of the seven cases had two expanded alleles of different lengths, suggesting the repeat sequence is highly mutatable. Example results are shown in Figure 4.12. Amplification of the GAA repeat expansion required the use of a long range enzyme mix consisting of a blend of *Taq* polymerase and a proofreading enzyme, to enable visualisation of carriers. Results for homozygous repeat expanded alleles and wild-type alleles could be produced with a *Taq* polymerase, but for heterozygotes amplification of wild-type alleles outcompeted expanded alleles, making it impossible to accurately distinguish wild-type homozygous and heterozygous individuals. The reaction using the long range enzyme blend was highly sensitive and lacking

robustness making results difficult to repeat and the assay inappropriate for diagnostic testing purposes.

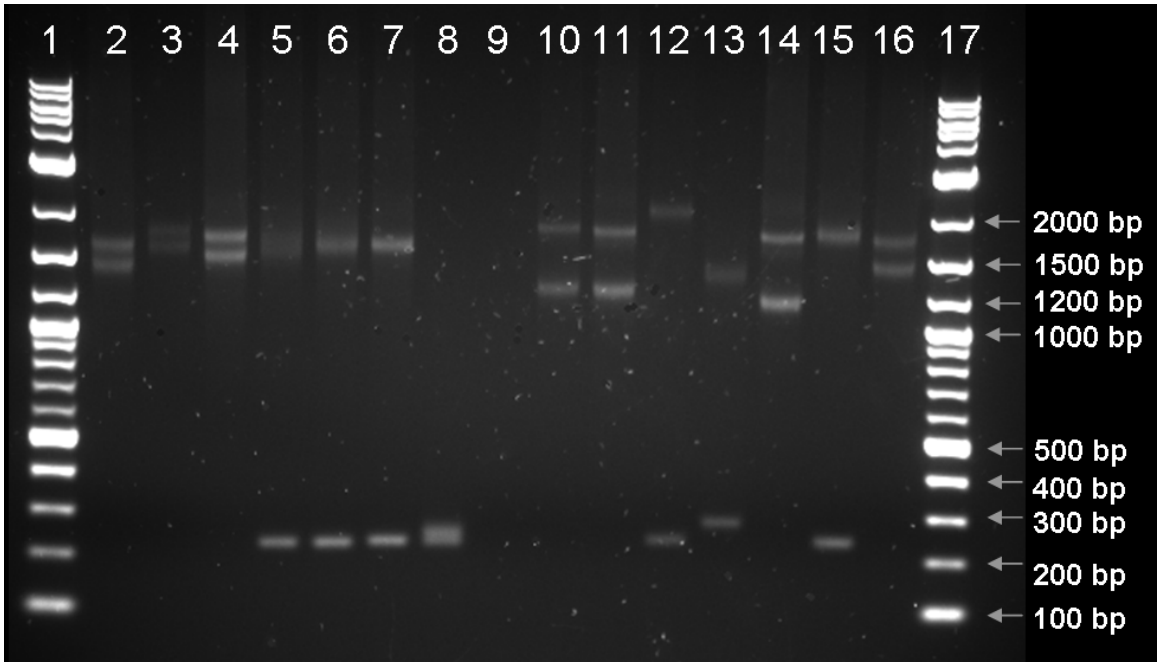


Figure 4.12 PCR analysis of the GAA repeat polymorphism

The expected wild-type amplicon was 238 bp. PCR products from cases are in lanes 2, 3 ,4, 10, 11, 14, and 16, from obligate carriers in lanes 5, 6, 7, 12, 13, and 15 and from a homozygous wild-type individual in lane 8. Lane 9 contains a no-template control. All cases have two expanded alleles of varying lengths. Results suggest expanded alleles ranging from approximately GAA_{300} – GAA_{650}.

Repeat expansion was confirmed by Sanger sequencing the longer alleles. A single allele of a case containing approximately 300 repeat subunits was gel extracted and used as a template for Sanger sequencing by capillary electrophoresis. Results confirmed a pure TTC.GAA repeat expansion for the allele, although sequencing diminished after approximately 60 GAA subunits. The sequencing trace is shown in Figure 4.13.

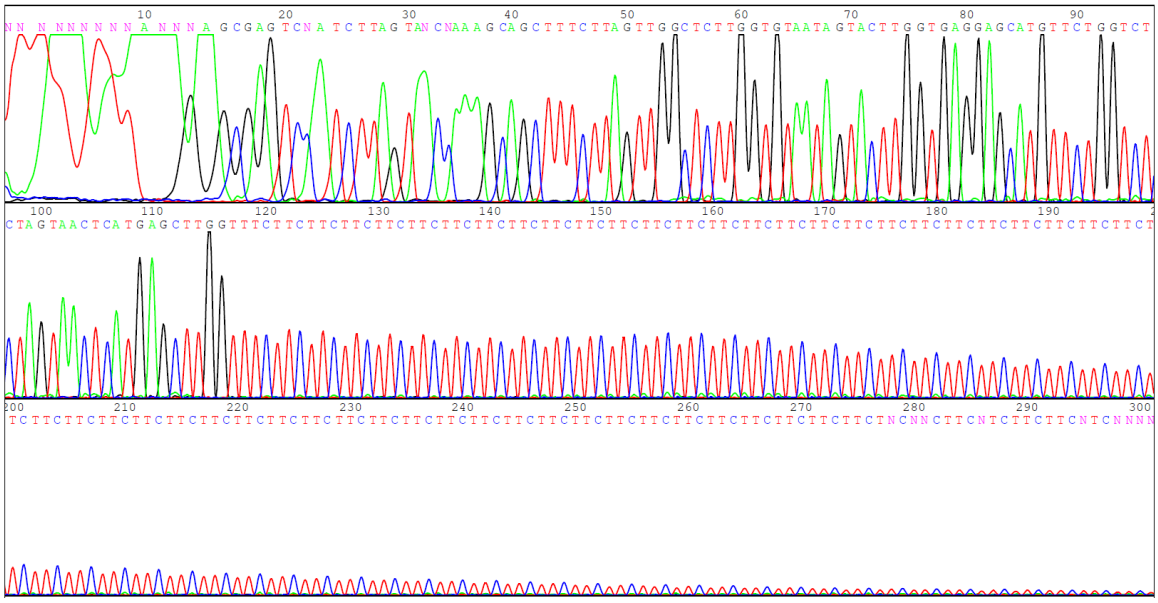


Figure 4.13 Sanger sequencing trace confirming GAA.TTC repeat expansion

Sanger sequencing of an expanded amplicon produced from a SCA case, confirmed a pure GAA repeat expansion, although a complete trace spanning the entire length of the GAA repeat could not be produced.

4.2.9. Allele length distribution and intergenerational repeat stability

To estimate allele lengths a standard curve was drawn by plotting the migration distances of DNA ladder bands on the agarose gel (Figure 4.12) against the known fragment sizes. To allow calculation of allele sizes from migration distances an order four polynomial regression was performed in Excel (Appendix 12). Using regression equations and based on the expected amplicon size of 238 bp for the reference sequence containing eight GAA repeats, an estimate of repeat copy number for unknown alleles could be calculated. Based on this calculation and the available data, the repeat copy number range for expanded alleles could be estimated as 318-651 and for normal alleles as 7-22. Familial relationships between individuals allowed intergenerational stability to be assessed. A range from an increase of 62 repeat subunits to a decrease of 109 repeat subunits per generation was observed. Repeat copy number also fluctuated to a smaller extent between siblings (Table 4.8).

Table 4.8 Repeat copy number and generational changes

Repeat copy number estimates based on migration distances on agarose gel. Allele sizes were calculated using polynomial regression (order 4) equations (Appendix 12). Allele copy numbers were calculated based on the reference amplicon size of 238 bp containing eight GAA repeat units.

Family	Gel Lane	ID	Relationship	Allele 1 copy number	Allele 2 copy number	Minimum generational expansion/contraction
1	7	5436	Father	10	495	n/a
	2	5357	Daughter	414	495	0
2	6	5407	Father	7	480	n/a
	5	5405	Mother	7	480	n/a
	3	5397	Daughter	466	542	+62/-14
	4	5404	Son	453	510	+30/-27
3	13	6479	Father	22	402	n/a
	12	6478	Mother	10	651	n/a
	10	6422	Son	357	576	-45/-75
	11	6477	Daughter	347	559	-55/-92
	14	6685	Son	318	542	-84/-109
4	15	8636	Father	10	542	n/a
	16	8637	Son	426	526	-16

4.2.10. DNA testing

A linkage-based test for SCA using microsatellite markers was launched in 2008. The diagnostic potential of seven polymorphic microsatellites within the disease-associated haplotype was considered by genotyping 147 individuals, including 13 cases and 16 obligate carriers. For all the microsatellite markers one or more of the obligate carriers was homozygous for the disease-associated allele, indicating that none of the markers tested were in complete linkage disequilibrium with SCA. However, using a combination of two microsatellites to produce two-marker haplotypes a pattern of complete linkage disequilibrium could be achieved with the individuals of known status ie none of the obligate carriers were homozygous for both disease-associated allele 204 and 155 of the two chosen microsatellites respectively (Table 4.9). Of the remaining 118 clinically normal IS individuals of unknown genetic status, 63 were predicted to be genetically clear (ie homozygous wild-type) and 55 were predicted to be carriers (ie heterozygous for the SCA mutation). In the run up to the test an additional 147 clinically normal IS were genotyped using the two-marker test, with 15 predicted to be carriers and 132 predicted to be clear. Out of all 294 individuals tested, only the known affected individuals produced a haplotype pattern suggesting a clinically affected status ie no wild-type associated 204/155

haplotypes were detected, indicating the high diagnostic value of using a linkage test with two markers.

Table 4.9 Microsatellite diagnostic test results for case individuals and obligate carriers

Disease status of all cases and obligate carriers could be predicted based on a two microsatellite marker haplotype (assuming successful PCR amplification).

Sample No.	Status	Chr20:15.75		Chr20:16.92	
4489	Affected	-	-	-	-
4490	Affected	204	204	155	155
4491	Affected	204	204	155	155
4914	Affected	-	-	-	-
4950	Affected	204	204	155	155
4951	Affected	204	204	155	155
5357	Affected	204	204	155	155
5397	Affected	204	204	155	155
5404	Affected	204	204	155	155
6422	Affected	204	204	155	155
6477	Affected	204	204	155	155
6685	Affected	204	204	155	155
8637	Affected	204	204	155	155
2224	Obligate carrier	200	204	155	155
2228	Obligate carrier	200	204	155	157
2236	Obligate carrier	196	204	155	157
2247	Obligate carrier	200	204	155	155
2248	Obligate carrier	196	204	151	155
2255	Obligate carrier	-	-	151	155
2275	Obligate carrier	196	204	155	158
2524	Obligate carrier	196	204	155	157
5297	Obligate carrier	196	204	151	155
5298	Obligate carrier	200	204	155	158
5405	Obligate carrier	196	204	155	166
5407	Obligate carrier	196	204	155	166
5436	Obligate carrier	200	204	155	158
6478	Obligate carrier	200	204	155	158
6479	Obligate carrier	204	204	155	158
8636	Obligate carrier	200	204	155	157

Since the test was launched in March 2008, the genetic services department at the AHT have tested DNA from 497 Italian Spinoni using the microsatellite linkage test. Over this period there has been a gradual decrease in the demand for DNA testing (Table 4.10), with no samples submitted for testing in 2012. This is probably because once a genetically clear line is established no further testing is required. Since the availability of the diagnostic test there have been no reported cases of SCA in the IS, suggesting that

diagnostic testing in the breed has led to the effective elimination of the mutation from the breed.

Table 4.10 Diagnostic test statistics for the SCA test

Year	Number tested	Number of carriers	Carrier percentage
2008	223	15	6.7
2009	128	5	3.9
2010	98	13	13.3
2011	48	1	2.1
2012	0	0	-

4.3. Comments and conclusions

4.3.1. Homozygosity mapping approach

The investigation of SCA in the IS shows the successful use of a homozygosity mapping approach to identify a disease-associated locus. Although homozygosity mapping approaches have largely been superseded by the use of GWAS approaches, this study demonstrates how a result can be achieved using the homozygosity method in a very modest sample cohort of six cases and six controls. Homozygosity mapping using microsatellite markers is a cost effective approach, but the technique is relatively labour intensive in terms of experimental work and data analysis, and because far fewer markers are used than in high-density SNP arrays there is a possibility that a disease-associated region may be missed.

4.3.2. Sequencing of genes in the disease-associated region

After confirmation of association to canine chromosome 20 by linkage analysis, the exons of all five known coding genes in the associated interval were resequenced. Although several polymorphisms were identified, all could be ruled out as causal as they did not segregate consistently with disease status. This indicated that the causal mutation was potentially in a previously unidentified gene or was a non-coding mutation. A range of mutation types have been associated with SCA in humans, including genomic deletions, point mutations and both intronic and exonic repeat expansions. Glutamine repeat expansions are associated with human SCA1, 2, 3, 6, 7 and 17. Intronic repeat expansions account for SCA10 and Friedreich ataxia, and 5' and 3' UTR repeat expansions are associated with SCA8 and SCA12 respectively, highlighting the potential for a non-coding mutation to be the cause of SCA in the IS.

4.3.3. Resequencing of the *ITPR1* gene

The *ITPR1* gene has been associated with ataxia in humans (SCA15 and SCA16) and mice making it a strong candidate for the cause of SCA in the IS. Given that many mutations associated with cerebellar ataxia are non-coding, the *ITPR1* gene was resequenced entirely including intronic regions and 10 kb of upstream and downstream sequence. Although exonic resequencing of *ITPR1* had previously ruled out coding mutations, functionally important promoters and enhancers of *ITPR1* have been described in the literature, which if altered could potentially lead to disease progression (Deelman et al., 1998, Kirkwood et al., 1997, Zhou et al., 2005). There is also evidence that the 3' region of the *ITPR1* transcript is functionally important and may have a role in trafficking of the mRNA and defining the location of translation (Bannai et al., 2004). As the *ITPR1*

gene spans a 320 kb region, it would not have been practical to sequence the entire gene using standard Sanger sequencing techniques. Instead *ITPR1* sequencing presented an opportunity to trial newly available "next generation sequencing" technology, using PCR to generate a template as solution based enrichment methods were not commercially available at the time. Although in the early stages, massively parallel sequencing approaches were rapidly advancing, and promised to play an important part in the future of genomics. With this in mind the project was an exciting opportunity to learn more about the technology and test the in-house data handling capabilities available at the AHT. A template from a single case was successfully used to generate sequencing data and in-house data handling was deemed feasible. Although interesting polymorphisms in non-coding regions were followed-up, none segregated fully with disease status when genotyped in further cases and controls. A long intronic tetranucleotide repeat was identified in the sequence which was investigated for expansion by genotyping cases and obligate carriers but was also ruled out.

4.3.4. Copy number investigation

Copy number investigation was carried out by genotyping one case and one obligate carrier on the CanineHD SNP array. The disease-associated region was assessed for variations in copy number by assessing B allele frequency and log R ratios. As no deviations in log R ratios were seen, copy number variation was ruled out as a potential cause of SCA, although as the CanineHD array has a density of one SNP approximately every 14 kb, small increases or decreases in copy number could potentially be missed. This was the case for the ~16 kb deletion in *BCAN* which was located between SNPs on the CanineHD SNP array, and therefore not identified at the GWAS stage when visualising the SNPs across the disease-associated interval.

4.3.5. Massively parallel sequencing of the entire disease-associated interval

As mutations in the *ITPR1* gene had been ruled out by almost entire sequencing of the gene from the 5' UTR through to the 3' UTR, albeit using a single case, the next logical step was to sequence the entire critical region in a number of cases and controls using target enrichment and massively parallel sequencing. Although coding mutations had been ruled out by exonic sequencing of all genes in the region, the sequencing experiment would serve as a check for coding mutations that could potentially have been missed due to inaccurate sequencing through primer site polymorphisms or interpretation errors. No additional coding polymorphisms were identified, and although a large number of non-coding SNPs and indels were identified through the sequencing analysis pipeline, all apart from three could be ruled out either due to an incorrect segregation pattern across cases and controls or the case-associated allele matching the reference allele.

Being able to exclude all but three potential variants remaining after filtering highlights the importance of selecting controls which have different haplotype patterns across the associated region, and where possible including control haplotypes with a high degree of similarity to the disease-associated haplotype to allow the maximum number of variants to be filtered out. All SNPs and indels were eventually ruled out either because the polymorphism was present in the set of Parson Russell Terriers which were simultaneously sequenced, or by genotyping across a larger sample cohort. Visual rescanning of the sequencing data, although very time-consuming and tedious, led to the eventual identification of a GAA triplet repeat expansion located in intron 35 of *ITPR1,* which segregated fully with disease status. Cases had extended alleles in the range of approximately $GAA_{(318)}$ to $GAA_{(651)}$ in comparison to a wild-type range of $GAA_{(7)}$ to $GAA_{(22)}$. Although cases shared a common disease-associated haplotype, all had two expanded repeat alleles of varying length, implying that repeat length is highly unstable and prone to generational expansion or reduction.

4.3.6. Comparison between target enriched sequencing attempts

The failure to identify the repeat expansion during the first targeted resequencing attempt highlights a number of key differences between the two sequencing experiments. The long PCR capturing the expanded region had an expected amplicon size of 10.6 kb. The increase in size of approximately 1 kb for the case amplicon was not sufficient to be clearly resolved on a 1% agarose gel.

For the sequencing experiment that used PCR as an enrichment method, reads were aligned all the way through the expanded region, and therefore a complete consensus sequence was produced highlighting no differences to the reference sequence. Because only single-end sequencing was performed, there were no un-aligned read mates to indicate potential insertions in the region. Although manual browsing of the sequence read alignments was performed, a basic alignment viewing program called Maqview was used, which is vastly inferior to IGV which was first launched in November 2008.

Several factors facilitated the identification of the repeat expansion in the second massively parallel sequencing experiment using a probe-based target enrichment methodology. As mentioned previously the second massively parallel sequencing experiment employed a paired-end sequencing strategy, and used IGV to visualise sequence read alignments, which has the capability to flag singleton reads. Although this is an improved strategy, many singleton reads were also flagged in controls because obligate heterozygote carriers were used, so with the cohort of individuals chosen

presence of singleton reads alone may not have been enough to identify the causal mutation. However, choosing homozygous wild-type individuals as controls may also have resulted in some singleton reads as wild-type alleles also varied in length, and sequencing reads from some of the longer alleles may not have aligned to the reference genome. This was in fact the case for some of the PRT individuals which were sequenced alongside the Italian Spinoni, after target enrichment with the same probeset. Bait positioning and the absence of capture baits across the repetitive region containing the SINE and GAA triplet repeat were the main factors facilitating mutation identification. Because the mutation locus and surrounding region contained repetitive elements, a 500 bp region containing the repeat expansion was excluded from bait design. In controls sufficient flanking sequence was captured by the baits to result in sequence coverage all the way through the repetitive region. For cases however, because the repeat expansion increased the size of the masked repetitive region no DNA fragments were captured that were of sufficient size to span across the GAA repeat. This resulted in a region of zero coverage in cases 3' of the GAA repeat sequence, as no probes 3' of the GAA repeat were located close enough to capture this region. The gap in coverage in cases, but not controls provided a discernable difference that led to eventual mutation identification.

4.3.7. Support for an intronic GAA triplet expansion as the cause of SCA

4.3.7.1. GAA repeat expansion is the cause of Friedreich ataxia

Although located in different genes, the GAA triplet repeat expansion discovered in the IS is similar to the repeat expansion known to cause autosomal recessive Friedreich ataxia in humans (Campuzano et al., 1996). Both the Friedreich ataxia and SCA repeat expansions stem from a short GAA repeat on the edge of a SINE. For Friedreich ataxia the intronic GAA repeat expansion is located in intron 1 of the frataxin (*FXN*) gene, and results in reduced expression of the *FXN* transcript, potentially due to the formation of a triplex DNA structure and "sticky DNA" (Sakamoto et al., 1999). Triplet repeat copy number in Friedreich ataxia patients has been reported to range between 66 and 1,700 subunits, thus the GAA repeat expansion observed in canine *ITPR1* falls within this range (Durr et al., 1996, Epplen et al., 1997). Like the repeat expansion event causing SCA there is evidence that Friedreich ataxia expanded alleles have a common founder (Cossee et al., 1997). Furthermore, large normal alleles also exist, which have arisen from a single founding chromosome. These large normal alleles in frataxin represent a pre-mutation reservoir and catastrophic single generation normal range to pathogenic repeat range changes have been observed (Cossee et al., 1997). This evidence suggests that even if expanded alleles can be successfully eradicated from the IS gene pool, expansion

of a large normal allele is possible, which could cause a resurgence of the disease in a breed with an isolated breeding population.

4.3.7.2. The *ITPR1* gene is associated with SCA in humans and mice

The inositol 1,4,5-trisphosphate receptor, type 1 (*ITPR1*) gene encodes an intracellular inositol triphosphate (IP3) gated calcium release channel. The products of the *ITPR1* gene form tetrameric proteins which are bound to the membrane of the endoplasmic reticulum. A cascade of cellular reactions are involved in the activation of *ITPR1*. Stimulation of cell surface bound G protein coupled receptors or receptor tyrosine kinases, result in a cascade of reactions and activation of phospholipase C to produce the secondary messenger IP3. On binding of IP3 to the IP3 receptor, the ligand-gated ion channel is opened, causing an influx of calcium ions into the cell cytoplasm from intracellular stores and activation of many calcium dependent cellular processes (Banerjee and Hasan, 2005). Cellular processes driven by calcium signalling include fertilisation, neuromodulation, cell growth and sensory perception (Berridge, 1993).

The *ITPR1* gene is predominantly expressed in the brain and nervous system and shows particularly high expression levels in the Purkinje cells of the cerebellum (Furuichi et al., 1993, Matsumoto et al., 1996). Heterozygous deletion of the *ITPR1* gene has been shown to cause SCA15 (also designated SCA16) in humans (van de Leemput et al., 2007). SCA15 is an autosomal dominant pure cerebellar ataxia with a reported onset age ranging from 10 to 66 years (Miyoshi et al., 2001, Storey et al., 2001). Using Epstein-Barr virus immortalised lymphocytes *ITPR1* protein levels were shown to be considerably lower in cell lines derived from affected individuals and haploinsufficiency had been suggested as the cause of the disease. Mice that are *ITPR1* null display a severe phenotype, although most die in utero. Surviving mice show normal behaviour at birth with signs of ataxia apparent at day nine. Tonic or tonic/clonic seizures develop at day 20-23 and mice die by the weaning period (Matsumoto et al., 1996). A similar but less severe phenotype is displayed by the opt mouse, which has an in-frame deletion of exons 43 and 44 of *ITPR1* (Street et al., 1997). For SCA in the IS, the ataxia is less severe than the mouse, perhaps due to production of a limited amount of *ITPR1* protein. Unlike for human SCA15 cases, ataxia has not been observed in heterozygous dogs, perhaps because more *ITPR1* transcript is produced than in haploinsufficient human cases. Alternative explanations include insufficient canine longevity to allow disease progression or failure of owners to notice phenotypic changes which may be more subtle in quadrupeds.

4.3.8. DNA testing

The linkage-based test for SCA in the IS was launched in April 2008. Since the launch of the test there have been no cases of SCA reported to the AHT, suggesting that the frequency of expanded alleles in the breed population is now very low. Although a linkage-based test has been successful for this disorder there are potential risks associated with running a linkage-based test. Firstly there is a risk of a recombination event occurring between the causal mutation and the marker used, leading to a loss of linkage disequilibrium and potentially causing false negative results. Secondly it is possible for false positive results to occur because of the disease-associated allele/alleles occurring on wild-type haplotypes. Thirdly, because the SCA test used microsatellite markers, there is potential for new alleles to form causing the disease-association to be lost. Although the causal mutation for SCA has now been identified it is likely that the strategy of using the linkage-based test will remain because an assay for the expanded allele, especially for identification of heterozygous individuals is not sufficiently robust to run in a diagnostic testing laboratory. Refinement of the linkage test may be possible to select markers which are very close in location to the causal mutation to reduce the risk of recombination events disrupting the linkage disequilibrium, or switching to SNP markers which are much less likely to mutate in comparison to microsatellites.

4.3.9. Summary

Using a small cohort of six cases and six controls a homozygosity mapping technique was used to identify a locus of the canine genome associated with SCA in the IS. Coding mutations were ruled out as a possible cause by exon resequencing through capillary based methodology. A GAA triplet repeat expansion in intron 35 of *ITPR1* was identified and strongly associated with SCA through probe-based target enrichment followed by massively parallel sequencing. An initial massively parallel sequencing experiment using PCR target enrichment failed to lead to identification of the repeat expansion, highlighting how choice of methodology can be crucial to mutation identification. The GAA repeat expansion is the first repeat expansion associated with canine ataxia, and is only the second form of ataxia, after Freidreich ataxia, to be associated with a GAA repeat expansion.

Chapter

5. Spinocerebellar ataxia in the Parson Russell Terrier

5.1. Background

5.1.1. The Parson Russell Terrier

The Parson Russell Terrier (PRT) was originally developed in the 19th century from foxing terriers, with the foundation of the breed type accredited to the Reverend John Russell. In the early part of the 20th century the Parson Jack Russell Terrier club was formed with an affiliation to the Fox Terrier Club, although it was not until 1990 that full Kennel Club recognition was granted, acknowledging the PRT as a separate breed to the shorter legged Jack Russell Terrier (JRT). The PRT is a moderately popular breed in the UK with approximately 500 registrations per year. The breed is also popular across Europe, America and Australia.

5.1.2. Spinocerebellar ataxia in the PRT

Spinocerebellar ataxia, often referred to as late onset ataxia (LOA) by breeders to distinguish the disorder from a reported juvenile form seen in the JRT, is a disease of progressive incoordination of gait, including loss of balance and hypermetria as the condition progresses. Clinical signs usually become notable to owners between 6 and 12 months of age although onset of clinical signs between 4 and 24 months has been reported. Upon onset owners may initially notice a slight weaving gait pattern of the hind legs, causing the dog to sway slightly when walking. As the disease progresses incoordination becomes more notable and a hypermetric (high stepping) gait may be observed with occasional hopping. On further progression clinical signs become more severe making ambulation difficult, with owners often electing to euthanise affected dogs on welfare grounds by five years of age.

5.1.3. Reports of ataxia in the PRT and related breeds

Cases of ataxia in breeds closely related to the PRT have been reported in the veterinary literature since the 1950s. In 1957 and 1962 two case reports describing a progressive ataxia and hypermetria affecting Fox Terriers were published (Bjorck, 1957, Bjorck, 1962). The cases described had an age of onset of approximately four months, consistent with LOA in the PRT, with test matings suggesting an autosomal recessive mode of inheritance. Pathological changes to the spinal cord were noted including bilateral demyelination of the cervical, thoracic and lumbar regions. Lesions were also seen in the dorsal spinocerebellar tracts. In a more recent study three new cases of ataxia in the Smooth-Haired Fox Terrier (SHFT) were investigated and five old cases retrospectively reviewed including an affected dog born in 1957 as part of the test mating experiments (Rohdin et al., 2010). Onset for the three new cases was four months of age, consistent

with previous observations. Full neurological examinations were performed on the three new cases and ataxic gait observed with hypermetria in all four legs. Swaying of the trunk was observed as the dogs walked and frequent falling was noted. For all three dogs a severely decreased menace response was recorded and postural reactions were impaired in all four limbs. A fine head tremor, and tremor of the muscles was also seen which worsened with excitement. On port-mortem examination no gross pathological changes of the nervous system were observed. Histopathological investigation of the spinal cord revealed oedema of the myelin sheaths, axonal swelling and macrophage infiltration across the circumference of the spinal cord, especially in the fibres of the pyramidal, spinocerebellar and ventral reticulospinal tracts. Analysis of the parietal cortex, basal ganglia, hippocampus and cerebellum revealed no morphological changes (Rohdin et al., 2010).

Two cases of ataxia were first described affecting short legged Jack Russell Fox Terriers (ie JRT) in 1973 (Hartley and Palmer, 1973). Notably the clinical signs were similar but apparent age of onset and progression was different for the two cases. This variability is more in line with observations in the PRT, rather than the more consistent onset age reported in the SHFT. Of the two cases, the clinical signs described for the later onset case are more comparable to those of LOA in the PRT. For this case the clinical signs were first apparent at six months of age with swaying of the hind limbs initially observed. Incoordination of the hind limbs with occasional collapse was also observed. Significant progression of clinical signs was observed on re-examination after fifteen months, with increased incoordination, unpredictable hindlimb movements, overprotraction of forelimbs and rotational intentional head tremor. For the earlier onset case, an abnormal gait was suspected when the puppy became ambulatory. After six months clinical signs had become severe, with the dog displaying a “dancing” gait. Additional clinical signs included exaggerated abduction, depressed hindlimb hopping reflex, and excessive forelimb protraction. Histopathological examination of nervous system tissue showed significant lesions of the central auditory pathway, with superior olivary nuclei displaying a marked loss of myelinated nerve fibres and moderate diffuse gliosis (increased glial cell numbers, predominantly comprising astrocytes, in damaged regions). Similar changes were also seen for the cochlear nuclei, but to a lesser extent. Extensive degeneration was also seen in nerve fibres connecting the superior olivary and cochlear nuclei and the trapezoid body and lateral lemniscus, with many swollen axons. Wallerian-like degeneration (retrograde and anterograde axonal degeneration after focal nerve damage) was observed in the brain and spinal cord. A second report into hereditary ataxia in the JRT and PRT was published in 2004 investigating 35 JRT and PRTs (Wessmann et al., 2004). The aims of

the study were to investigate abnormal brain stem auditory-evoked potentials (BAEPs) and to calculate mode of inheritance of hereditary ataxia. Affected individuals had an average onset age of 4.4 months with a range of 2 to 9 months of age and on average were euthanised at the owner's request at 16 months. In addition to clinical signs of ataxia previously described, this report described generalised seizures in some dogs (13/35) as evidenced by muscle twitching over the head or limbs, exercise weakness (11/35), respiratory distress (7/35) and behaviour changes (7/35). This muscle twitching could have represented myokemia rather than generalised seizures, particularly in the absence of involvement of the autonomic nervous system and the more recent reports of myokemia in the JRT (Vanhaesebrouck et al., 2012). Reduced menace responses were observed in one dog. Abnormal BAEP patterns were observed in four of the eight individuals investigated. Authors discuss that this may be consistent with the observed axonal damage associated with the disease, especially damage to the trapezoid body and lateral lemniscus. Histopathological investigation revealed changes across the entire nervous system, with bilateral myelopathy and other degenerative changes in the brain, brain stem and spinal cord consistent with previously reported cases (Hartley and Palmer, 1973). Complex segregation analysis across three families containing 115 individuals suggested a hereditary cause for the ataxia with a polygenic model most likely, although a major gene effect with additional polygenic factors was also not excluded. In the discussion Wessmann and colleagues highlighted that although the gait disturbance in the JRT and PRT appeared identical to the SHFT, different modes of inheritance have been suggested and histopathological investigations indicated brain involvement for the PRT and JRT, but not the SHFT, suggesting a different disease process. However, Rohdin and colleagues proposed that the different genetic modelling systems used in predicting inheritance patterns for the JRT and SHFT may be the primary reason for the different conclusions and suggested that clinical signs and histopathological changes were similar, implying the same disease process for the two breeds.

5.1.4. Summary

Hereditary ataxia in the PRT and related breeds has been described since the 1950s. At the AHT a cohort of owner reported cases were collected with a view to performing a genetic investigation to improve understanding of the disease, elucidate the mode of inheritance and identify the causal mutation(s). Identification of the genetic cause would enable the development of a diagnostic test to allow breeders to identify genetic carriers or risk factor carriers, depending on the determined mode of inheritance. Identification of genetic risk factors would allow investigation into ataxia in related breeds and potentially determine whether the same genetic defects underlie ataxia in the PRT, JRT and SHFT.

5.2. Results

5.2.1. Genome-wide association study

A genome-wide association study was performed with 16 PRT LOA cases and 16 controls. Where possible first-degree relatives were used as controls (ie parents or siblings of cases). Failing this half-siblings were selected as controls or, for isolated cases where DNA from relatives could not be obtained, unrelated controls were used. A strategy of using close relatives for controls was used with the aim of limiting levels of genomic inflation such as the high levels seen in the Cavalier King Charles Spaniel studies described in Chapter 3.

DNA samples were genotyped on the Illumina CanineHD array. All 32 samples genotyped successfully, achieving a genotyping call rate of >99.8% after reclustering all SNPs in the Genome Studio software package. No additional QC or manual editing of SNPs was performed before exporting for statistical analysis using PLINK.

5.2.2. Allelic association analysis

Allelic association analysis was performed using the statistical package PLINK which was implemented at the Linux command prompt. After exclusion of SNPs with a minor allele frequency <0.05 and genotyping success rate <0.95, 126,225 SNPs remained. The genomic inflation factor based on the median chi-squared value was 0.81 indicating no stratification between the case and control populations. Basic allele associated analysis on the filtered SNPs revealed a strong statistical signal on chromosome 18 (P_{raw} = 7.04 x 10^{-9}) (Figure 5.1 A). To correct for multiple testing, allelic association analysis was performed using 100,000 MaxT permutations in PLINK, which resulted in a single signal on chromosome 18 of genome-wide significance (P_{genome}=1.06 x 10^{-3}) (Figure 5.1 B). Results are highly indicative of a simple autosomal recessive mode of inheritance for the disorder.

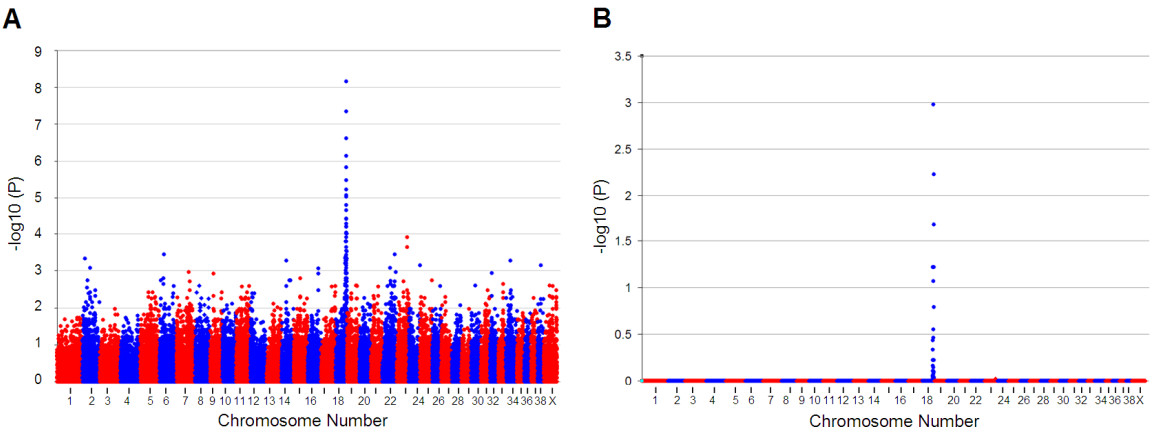


Figure 5.1 Allelic association analysis for LOA in the PRT

Allelic association analysis on 16 cases and 16 controls. Each dot represents a single SNP, with -log10 (p) values on the y-axis plotted against genome position (split into chromosomes) on the x-axis. (A) Raw unadjusted –log10(p) values with a strong statistical signal indicated on chromosome 18 (P_{raw} = 7.04 x 10^{-9}). (B) Plot of -log10(p) values after 100,000 maxT permutations analysis to correct for multiple testing, showing a single peak reaching genome-wide significance on chromosome 18 (P_{genome}=1.06 x 10^{-3}).

Identity-by-state (IBS) statistics were calculated using PLINK and values plotted in Excel to produce a multidimensional scaling (MDS) plot showing relatedness of individuals used in the study. As expected from the genomic inflation value of 0.81 individuals were evenly spaced, with no separate clustering of cases and controls (Figure 5.2 A). A quantile-quantile plot was constructed, which showed the observed probability values to closely track the expected values, further confirming the low genomic inflation value (Figure 5.2 B).

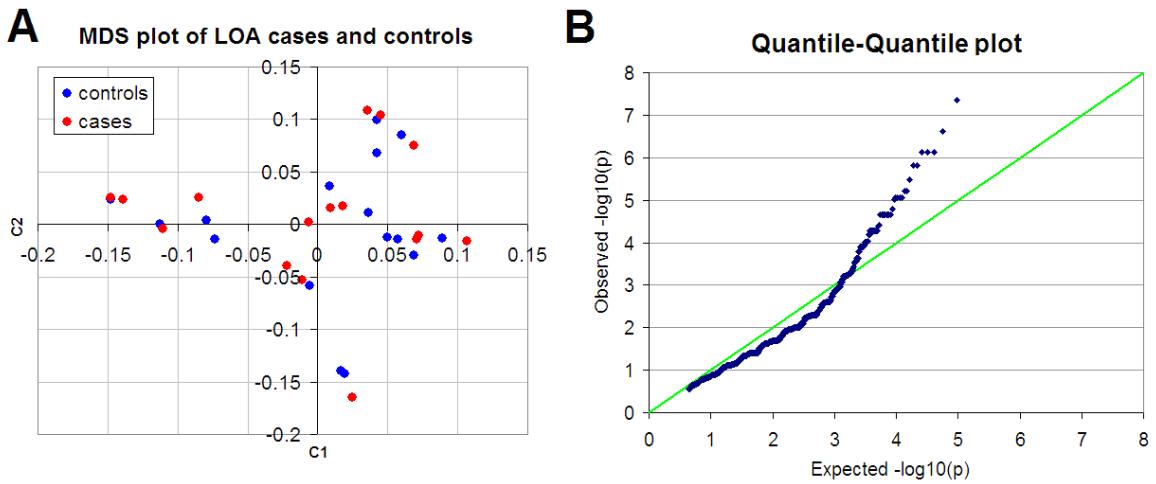


Figure 5.2 MDS and QQ plots for LOA in the PRT

(A) MDS plot showing the relatedness of individuals used for the GWAS based on IBS data calculated from SNP genotypes. Cases and controls are evenly mixed on the plot with no clustering, indicating no population stratification. (B) Quantile-quantile plot of observed versus expected P values. The genomic inflation value for the dataset was 0.81.

Raw genotyping data were visualised across the region of genome-wide significance on chromosome 18. All but one of the 16 cases were homozygous for a shared haplotype, the interval for which was defined by recombination events which had occurred in two individuals resulting in loss of the shared homozygosity (Figure 5.3). The disease-associated interval was a 1.89 Mb region defined as chr18:53,533,360-55,418,743 based on the CanFam2 genome build.

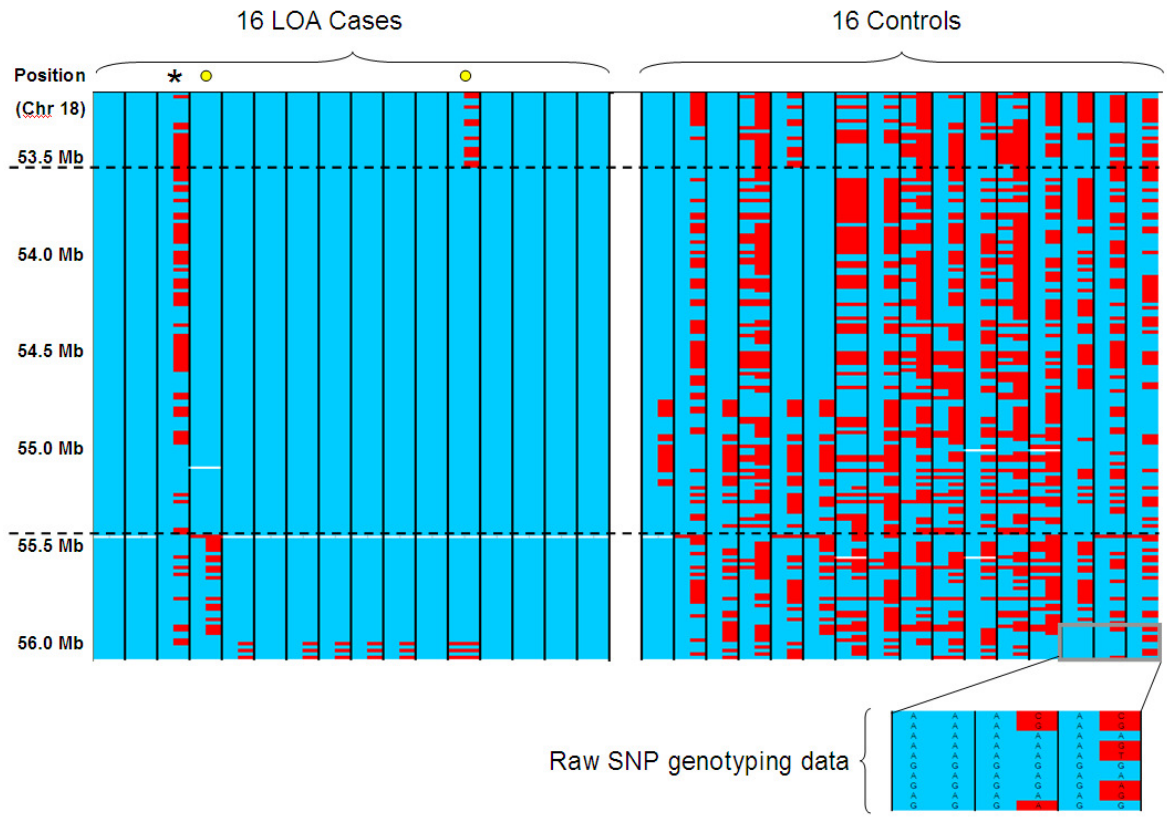


Figure 5.3 Raw genotyping data across the disease-associated region for LOA

Graphical overview of the raw genotyping data across the LOA critical region. Each column represents one individual, and SNP markers are listed in the rows with the approximate positions indicated. Major alleles are highlighted in blue, minor alleles in red and missing genotypes in white. Interval defining cases are marked by a circle at the top of individual columns. The outlying case which is not homozygous for the disease-associated haplotype is marked with an asterisk.

5.2.3. Investigating genes in the disease-associated region

The disease-associated region for LOA contained 91 genes and was syntenic to chromosome 11 of the human genome (see Appendix 9). Genes within the interval were assessed for potential involvement in LOA, with the gene encoding beta-III spectrin (*SPTBN2*) a particularly strong candidate, as mutations in the gene have been shown to cause SCA5 in humans (Ikeda et al., 2006).

5.2.4. *SPTBN2* sequencing

The *SPTBN2* was exon resequenced in an attempt to identify potential causal mutations for LOA. No coding or splice site polymorphisms were identified.

5.2.5. Targeted resequencing of the LOA disease-associated interval

With no coding mutations identified in the most probable candidate gene and with the disease-associated interval containing 91 genes, the most appropriate strategy for investigating the disease-associated interval was to use targeted resequencing of the entire region. To make the investigation more efficient and cost-effective, probes were designed to capture both the disease-associated region for LOA in the PRT and SCA in the IS (Table 4.5).

Two LOA cases and three controls were selected for library preparation (five IS DNA samples were also chosen). Controls were selected based on haplotype analysis across the disease-associated region, and individuals chosen with different haplotype patterns to facilitate polymorphism exclusion. Libraries were made using the SureSelect target enrichment methodology and sequencing performed on the Illumina HiSeq 2000, producing a 19.64 Gb dataset of 51 bp paired-end reads. Further details of the experiment design and enrichment statistics can be found in sections 4.2.8.1 – 4.

5.2.6. Data analysis

Data were analysed in Linux using the NGS pipeline (described in Chapter 3). SNP and indel calls were aligned across all individuals using the SNP Handler Excel macro developed by Dr Mike Boursnell at the AHT. Compiled SNPs were annotated with gene consequence predictions and assigned a segregation score based on how well SNPs segregated according to disease status. In total 7,024 SNPs and 1,507 indels were identified across the LOA disease-associated region, with 541 variants segregating in accordance with disease status. Sequencing data generated from the five Italian Spinoni captured across the same region and ten Siberian Huskies captured across a fully overlapping region (chr18:51,770,000-58,240,000) for an unrelated project (data generated by Louise Pettitt and Sally Debenham) were used as additional controls to help to reduce the number of segregating SNPs to 141. The majority of the SNPs were distributed towards the 3' end of the disease-associated region, although one SNP was located in an isolated intergenic position approximately 650 kb 5' of the next correctly segregating SNP, perhaps indicating a more recent mutation event (Figure 5.4).

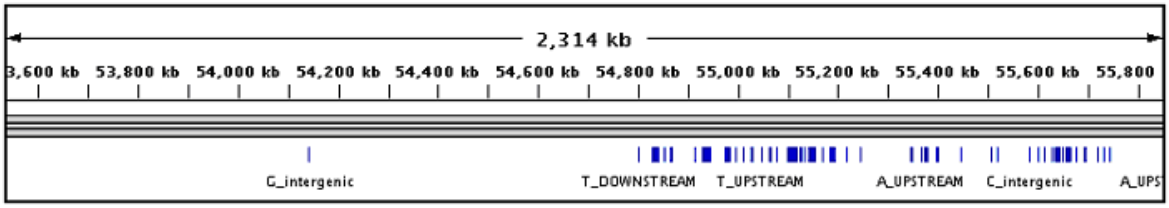


Figure 5.4 Distribution of SNPs across the sequenced region

Segregating SNPs were mostly distributed towards the 3' end of the disease-associated region, apart from a single intergenic SNP located toward the 5' end of the region.

5.2.7. In-house targeted resequencing of additional controls

A second enrichment kit with the same design was used to produce sequencing libraries for ten additional PRT controls in an attempt to narrow down the number of potentially causal variants. In June 2012 an Illumina MiSeq benchtop massively parallel sequencer was acquired by the AHT, presenting an opportunity for the entire workflow of the experiment to be completed in-house. In overview, DNA extracted from buccal swabs was fragmented by in-house sonication yielding fragments in the 100-800 bp range. Indexed libraries were then subsequently prepared using SureSelect target enrichment methodology and quantified for sequencing on the Illumina MiSeq platform. Paired-end sequencing with 150 bp reads was performed yielding a dataset of 2.07 Gb. The individuals used and dataset are summarised in Table 5.1. The data generated were processed through the sequence analysis pipeline and variants aligned across all samples. In total 6,544 SNPs and 1,559 indels were identified across the ten additional controls sequenced. Based on the suggestive haplotypes of the individuals sequenced, 95 SNPs segregated with the predicted disease status. Using the SNP and indel information from the ten additional control samples and the original dataset, 73 SNPs and indels could be excluded from the list of segregating variants. 22 SNPs could not be excluded because of lack of read depth for the ten additional controls. The remaining 70 segregating SNPs and indels were visually checked by browsing the sequence read alignments.

Table 5.1 Summary of target-enriched MiSeq sequencing data of ten additional controls

Libraries from ten additional control individuals were prepared and sequenced completely in-house, to reduce the number of disease segregating variants. Capture success and sequencing results are summarised.

ID	Relation to case	Suggestive haplotype	No. of reads (Million)	Dataset size (Mb)	Target enrichment efficiency (%)	Bases achieving 10x coverage (%)	PCR duplicates (%)
11291	Dam	heterozygous	1.13	170	71.2	58.9	1.9
11294	Half sibling	Wild-type	1.25	188	73.6	59.0	2.6
11491	Half sibling	heterozygous	1.20	180	77.2	59.6	2.4
12125	Half sibling	heterozygous	1.19	179	76.5	60.6	3.0
12899	Half sibling	heterozygous	1.31	197	77.6	60.7	2.8
17477	Unrelated	Wild-type	1.35	203	54.2	50.3	1.6
17529	Unrelated	Wild-type	1.96	294	15.2	31.2	0.4
17553	Unrelated	Wild-type	1.72	258	76.4	63.4	3.5
18032	Unrelated	Wild-type	1.05	158	77.4	59.4	1.9
18501	Dam	heterozygous	1.20	180	62.1	55.2	1.9

5.2.8. Investigation of segregating variants

By visualising sequencing data aligned against gene positions, cross species conservation data, and variant effect prediction data, three provocative candidate SNPs were identified. These were non-synonymous missense mutations in the genes encoding the calcium dependent cysteine protease, calpain 1 (*CAPN1*) and vacuolar protein sorting 51 homolog (*VPS51*) and the isolated intergenic SNP which appeared to be in a semi-conserved DNase1 hypersensitivity region.

The *CAPN1* mutation was predicted to cause cysteine to tyrosine amino acid substitution (C115Y). Multi-species alignment in the UCSC genome browser across 46 vertebrate species shows a very high level of conservation at amino acid 115 of the *CAPN1* protein (Figure 5.5). Of the 38 species with alignable sequence data, 37 had a cysteine residue at position 115. The only species not to have a cysteine residue at position 115 was the alpaca, although a 1 bp insertion at residue 124 puts the transcript out of frame and may suggest that the gene is non functional in this species.

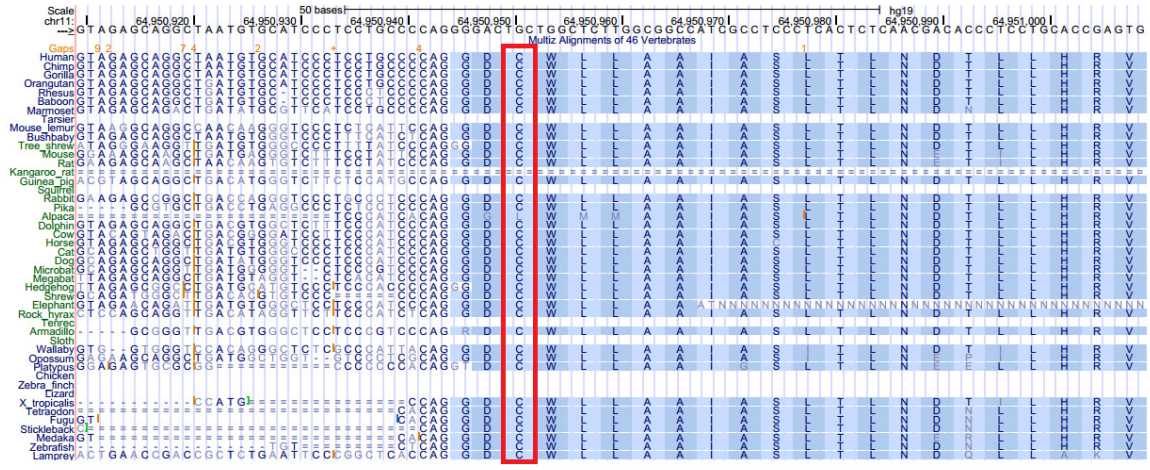


Figure 5.5 Multi-species alignment across residue 115 of the calpain1 peptide

Alignment across 46 vertebrate species was visualised using the UCSC genome browser to assess the level of cross-species conservation for the *CAPN1* gene. The reference genome is human (Hg19).

The mutation in the *VPS51* gene was predicted to result in a glutamic acid to lysine amino acid change (E24K). Multi-species alignment in the UCSC genome browser across 46 vertebrate species shows a high level of conservation at the amino acid level (Figure 5.6). Of the placental mammal group (human to sloth) 29 of the 30 aligned species have a glutamic acid residue (E) at the corresponding locus. The differing species, mouse, has an aspartic acid at amino acid 24 of *VPS51*. Only three of the more distantly related vertebrate species (wallaby to lamprey) have alignment data at the questioned position, however browsing the adjacent amino acids suggests that the protein is not conserved for these species.

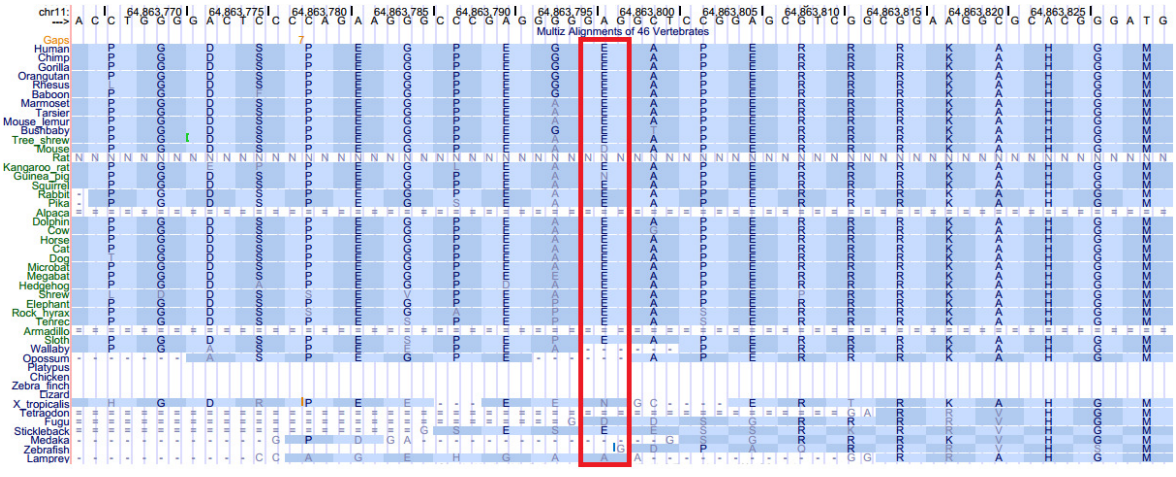


Figure 5.6 Multi-species alignment across residue 24 of the VPS51 peptide

Alignment of the *VPS51* gene across 46 vertebrate species was visualised using the UCSC genome browser to assess the level of cross-species conservation. The reference genome is human (Hg19).

The intergenic SNP identified in a semi-conserved region of the canine genome was defined as chr18:54,147,546A>G based on the CanFam2 genome build. Browsing the corresponding position in the human genome showed that the SNP was positioned in a region of DNase1 hypersensitivity, according to data from the ENCODE (Encyclopedia of DNA elements) project. As regions of DNase1 hypersensitivity are often indicative of regulatory areas of the genome, the SNP identified could potentially be functionally important and therefore worthy of further investigation. Investigation of multi-species alignments at the base pair level showed the A base to be conserved across many placental mammalian species, with 16 out of 20 species possessing an A base at the position (Figure 5.7).

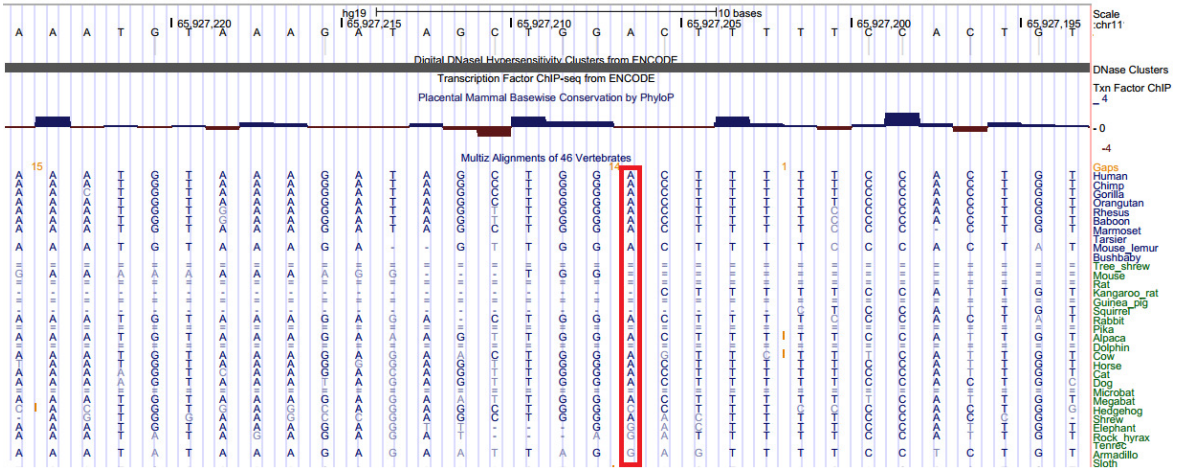


Figure 5.7 Multi-species alignment across the intergenic associated SNP locus.

Base-pair level conservation across 33 placental mammalian species, for an isolated intergenic SNP polymorphism identified in the canine genome (chr18:54,147,546A>G). The reference genome is human (Hg19).

5.2.9. Follow-up investigation of candidate SNPs

All three candidate SNPs were further assessed by genotyping on an additional cohort using allelic discrimination assays (TaqMan). The SNPs were initially genotyped on a multibreed panel of 96 healthy individuals, from 32 different breeds. The non-reference *VPS51* SNP allele was found in four individuals, (one Tibetan Spaniel, one Miniature Poodle and two Dobermann Pinschers), indicating that it was likely to be a common polymorphism. All 96 individuals of the multibreed panel were homozygous for the reference allele for *CAPN1* and intergenic SNPs, which were therefore still potentially causal for LOA. All three SNPs were subsequently genotyped on an expanded cohort of 227 PRTs, including the 36 individuals that were used in the initial GWAS study, ten additional cases and 185 additional controls. Results are summarised in Table 5.2.

Table 5.2 Segregation analysis of the three top LOA associated SNPs

Results of genotyping an extended PRT cohort for the three top LOA associated SNPs

	Intergenic SNP		*CAPN1* SNP		*VPS51* SNP	
Genotype	**Case**	**Control**	**Case**	**Control**	**Case**	**Control**
w/t homozygous	3	131	3	133	3	128
Heterozygous	1	69	1	67	1	68
Mutant homozygous	22	1	22	1	22	5

The intergenic and *CAPN1* SNPs most closely segregated with disease status. Interestingly some of the genotyping results indicated that a recombination event had occurred between the two markers (Table 5.3). The *VPS51* SNP was excluded because five controls were homozygous for the disease-associated allele.

Table 5.3 Evidence of recombination between *CAPN1* and the intergenic SNP

Six individuals gave discordant genotyping results between the *CAPN1* SNP and the intergenic SNP candidate. Results provided evidence of a recombination event as the wild-type *CAPN1* alleles were found in association with non wild-type intergenic SNP alleles and vice-versa.

ID	Intergenic SNP genotype	*CAPN1* genotype
11166	wild-type	Heterozygous
13365	Heterozygous	wild-type
13539	Heterozygous	wild-type
14280	wild-type	Heterozygous
18046	wild-type	Heterozygous
21275	wild-type	Heterozygous

5.2.10. Follow-up of discordant cases and controls

Based on the *CAPN1* variant, the most provocative SNP according to exonic positioning and close segregation with disease status, discordant cases and controls were followed up by communicating with owners and veterinarians. The single individual reported as clinically clear, but homozygous for the disease-associated allele of the *CAPN1* mutation, was still reported to be clinically healthy by the owner. The owner was asked to provide a video of the dog walking, trotting and running, which clearly showed the dog to have an ataxic gait, with a weaving pattern of the hind legs, which is typical for cases of LOA in the PRT. Video footage was provided when the dog was four years of age, perhaps suggesting a slower than normal progression rate for the condition. Three of the four owners of the dogs which were clinically affected, but not homozygous for the *CAPN1* disease-associated allele were also successfully contacted. All three dogs had been

euthanised on welfare grounds due to disease severity. Video footage was not available for the cases, but the clinical signs described were consistent with LOA. The owner of two of the dogs was an experienced breeder and veterinarian who had seen a number of LOA cases. This owner stated that clinical signs were perhaps more severe and the progression rate faster than an average case, but within the expected range. The owner of the fourth dog could not be contacted.

5.2.11. Genotyping of Jack Russell Terriers at the *CAPN1* SNP locus

Five ataxic JRTs were genotyped for the *CAPN1* SNP associated with LOA in the PRT. Of the five dogs three were homozygous wild-type, one was heterozygous and one was homozygous for the disease-associated allele, suggesting that the causal mutation in the LOA locus is not a major cause of ataxia in the JRT.

5.2.12. Predicting functional effects of the *CAPN1* variant

The *CAPN1* SNP was the strongest candidate polymorphism based on conservation data and segregation analysis across an extended cohort of individuals. Three computational tools were used to help assess whether the *CAPN1* amino acid change could have an important functional effect. These were SIFT (Sorting Intolerant from Tolerant), Polyphen (Polymorphism phenotyping) and MutationTaster (Adzhubei et al., 2010, Ng and Henikoff, 2001, Schwarz et al., 2010). The SIFT algorithm is based on comparing protein sequence similarity across species, and assumes that highly conserved amino acids will be highly intolerant to changes in identity. The *CAPN1* C115Y amino acid change was predicted by SIFT to be damaging. The program Polyphen uses both conservation information and data from SWISS-PROT, an annotated database of non-redundant protein sequences. The output from PolyPhen suggested that the C115Y amino acid change was probably damaging with the highest possible level of confidence (Figure 5.8).

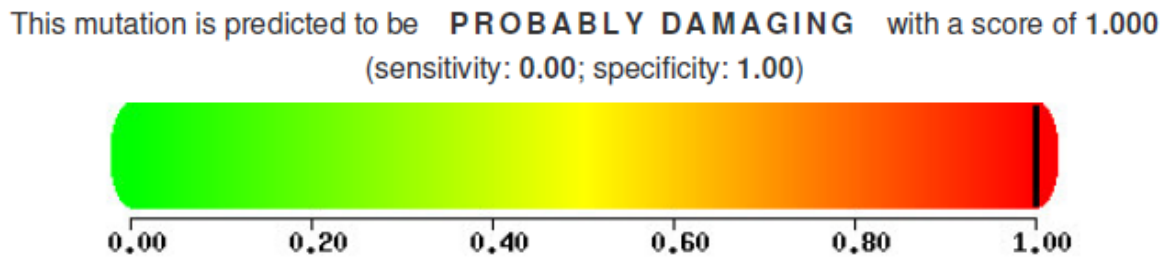


Figure 5.8 Polyphen output for the *CAPN1* C115Y variant

Polyphen consequence prediction for the *CAPN1* non-synonymous SNP. The output suggested the mutation to be probably damaging with the maximum score of 1, where a score of 0 is given for an amino acid substitution that is predicted to be benign with the maximum level of confidence and a score of 1 indicates that an amino acid substitution is likely to be damaging with the maximum level of confidence.

The third prediction tool MutationTaster considers both protein sequence and nucleotide sequence conservation across a number of model organisms, protein features and domains that could potentially cause a reduction in the level of mRNA produced. MutationTaster gave the prediction that the *CAPN1* non-synonymous SNP was disease causing, with a probability score of 1 (maximum score). Residue 115 and the five flanking amino acids either side of residue 115 were identical on alignment of human, chimp, cat, mouse, *Xenopus*, zebrafish, *Fugu*, *C.elegans* and *Drosophila* protein sequences (ie conserved even in two invertebrate species). On nucleotide sequence alignment base identity of the altered base was identical for human, chimp, rhesus, cat, mouse, Xenopus, *Fugu*, although no sequence alignments could be retrieved for zebrafish, *C.elegans* and *Drosophila*. Protein domain analysis by the MutationTaster package predicted loss of the calpain catalytic domain (aa55-354) and loss of a helix domain (aa115-124). A score for chemical dissimilarity of 5.29 out of 6 was calculated by MutationTaster, based on the Grantham matrix, for cysteine to tyrosine residue change. Using the original Grantham matrix of amino acid dissimilarity, cysteine versus tyrosine comparison gave a score of 194 (range 5-215, mean 100) (Grantham, 1974).

A BLOSUM (BLOcks SUbstitution Matrix) table was used to investigate the evolutionary likelihood of a cysteine to tyrosine substitution. Using the BLOSUM64 matrix a C>Y substitution has a logarithm of odds value of -3, and indicated that a tyrosine is unlikely to be substituted for a cysteine (Henikoff and Henikoff, 1992).

5.2.13. mRNA-seq data analysis

A genome-wide mRNA-seq dataset was generated using cerebellum tissue from a LOA case and cerebellum tissue from a clinically normal control (See section 6.2.5.). Data confirmed expression of *CAPN1* in cerebellum tissue. By counting aligned reads across genes in the LOA disease-associated region potential differences in the expression between the case and the control could be assessed. Significant changes at the 5% level before adjustments for sample size and read count are displayed in Table 5.4 (Data analysis performed at the Wellcome Trust Centre for Human Genetics, Oxford).

Table 5.4 Gene expression changes in a LOA case across the disease-associated region

List of significant changes in gene expression before P value adjustment for cohort size and read count. Fold change values above 1 indicate a reduction in expression in the LOA case cerebellum.

Gene	Position	Read count case	Read count control	Fold Change	P value	P adjusted
SPTBN2	chr18:53664195-53696100	14322	9912	0.69	0.01	0.37
RBM14	chr18:53746046-53754057	2219	1565	0.71	0.02	0.49
BBS1	chr18:53843758-53860191	232	105	0.45	0.02	0.41
NPAS4	chr18:53933238-53936598	67	17	0.25	0.05	0.68
FAM89B	chr18:54658069-54659575	1198	1678	1.40	0.04	0.62
SSSCA1	chr18:54660687-54661481	869	1247	1.44	0.04	0.63
SCYL1	chr18:54693198-54705475	949	1430	1.51	0.02	0.42
POLA2	chr18:54901204-54917024	605	964	1.59	0.01	0.38
MRPL49	chr18:55093567-55097563	759	1092	1.44	0.04	0.66
FAU	chr18:55098084-55099271	4892	6692	1.37	0.03	0.53
VPS51	chr18:55108993-55130845	1846	2535	1.37	0.04	0.60
PYGM	chr18:55423235-55435168	6560	8921	1.36	0.03	0.53
NRXN2	chr18:55464994-55562401	14452	11026	0.76	0.04	0.65

Although some small changes (less than 2 fold increases/decreases in gene expression) were seen across the disease-associated region, none of the changes reached significance after adjusting for the small sample size.

Read alignments for the cases and control mRNA-seq datasets were visually inspected as a way of checking for mutations in regions not captured by target enrichment. No additional polymorphisms were identified in the mRNA-seq datasets. Read depths across genes were visually compared for the case and control datasets, confirming no large changes in expression across the disease-associated region.

5.2.14. Copy number investigation

The sequence dataset was interrogated for copy number variation by assessing the ratio of case to control aligned read count across overlapping 5 kb windows. Windows containing less than 100 reads for both the case and control dataset were excluded. Across the disease-associated region the ratio remained close to 1, excluding copy number variants of 5 kb or greater (Figure 5.9).

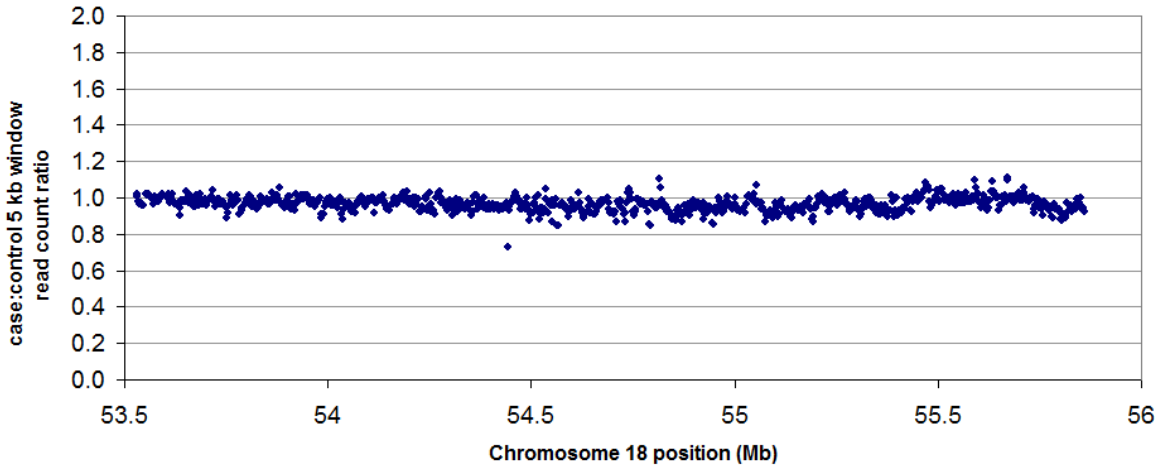


Figure 5.9 Copy number variation investigation

Potential copy number variants were investigated by comparing the count of read alignments for one case versus one control across a sliding 5 kb window. Ratios of greater than one are suggestive of an increase in copy number, and ratios of 0.5 and 0 suggestive of heterozygous and homozygous deletions respectively.

5.2.15. DNA test launch

A diagnostic test was launched in November 2012 based on the *CAPN1* SNP for the PRT but not the JRT, using the allelic discrimination genotyping assay. Dogs with two copies of the disease-associated allele were defined as affected, with one copy of the disease-associated allele as carriers, and with no copies of the disease-associated allele as normal.

5.3. Comments and conclusions

5.3.1. Genome-wide association study

A GWAS approach using a high-density SNP array to genotype 16 cases and 16 controls was successfully used to identify a single locus associated with LOA in the PRT. Reports in the scientific literature suggested that a complex pattern of inheritance was most likely for LOA in the PRT, so results indicating a single associated locus largely contradicted this prediction. Of the 16 cases, 15 were homozygous for a shared haplotype, leading to a high level of statistical significance and suggesting that the locus identified was the major cause of LOA in the breed.

Because of the high levels of genomic inflation seen in previous association studies, particular care was taken when selecting cases and controls for use in the LOA study. Where possible first degree relatives were used (ie parents or siblings), and failing that half-siblings or unrelated individuals were used as controls. This led to a genomic inflation value of 0.81 based on the median chi squared value. This suggested slight genomic deflation, meaning that the range of chi squared values were overall lower than expected for a randomly selected cohort of cases and controls with no population stratification. The low genomic inflation value is likely to be due to first degree relatives being more similar genetically in the purebred dog than in human studies, leading to lower than expected probability values. Nevertheless a single strong statistical signal remained, and the study appeared to be unaffected. Genomic deflation in studies of complex disease could however be potentially problematic as significant signals could be masked due to high levels of genetic similarity between cases and controls, meaning a high proportion of risk factors for disease would be shared in both sets.

5.3.2. Exclusion of the *SPTBN2* gene as the cause of LOA in the PRT

The gene encoding beta-III spectrin (*SPTBN2*) provided a strong candidate gene within the disease-associated region. Mutations in *SPTBN2* have been shown to cause SCA5 in humans (Ikeda et al., 2006). Spectrins are important structural components of the plasma membrane and play a significant role in restricting and stabilising membrane spanning proteins within specific subdomains of the plasma membrane. Beta-III spectrin is primarily expressed in the nervous system and the highest levels of expression are found in Purkinje cell soma and dendrites (Sakaguchi et al., 1998). Beta-III spectrin has been shown to stabilise the glutamate transporter *EAAT4* at the plasma membrane of the Purkinje cells (Jackson et al., 2001), facilitate protein trafficking by linking the microtubule motor to vesicle-bound cargo (Holleran et al., 2001) and maintain a high density of sodium

channels within the soma and dendrites of Purkinje cells (Perkins et al., 2010). Beta-III spectrin is critical for development of Purkinje cells (Gao et al., 2011).

On exon resequencing of *SPTBN2* no non-synonymous, frameshift or splice site variants were identified excluding coding changes as a potential cause of LOA in the PRT. Targeted resequencing of the LOA disease-associated region enabled coding and non-coding regions of the *SPTBN2* to be investigated, although no potentially causal variants were identified. Data from mRNA-seq experiments provided a further check on the *SPTBN2* transcript helping to exclude mis-splicing and changes in expression levels as potential causes of LOA. The *SPTBN2* gene was therefore excluded.

Although similar clinical signs of ataxia are shared in both LOA and SCA5, other clinical features suggest potentially different mechanisms. Clinical features of SCA5 suggest a predominately cerebellar disease (Bauer et al., 2006), whereas histopathological examination of LOA cases suggest limited cerebellar pathology, with degeneration of the brain stem and spinocerebellar tract involved in disease progression. This evidence supports a different genetic cause for LOA and SCA5, although it is possible for different mutations in the same gene to result in variable phenotypes.

5.3.3. Targeted resequencing of the LOA disease-associated region

Targeted resequencing was successfully used to identify potential causal variants across the LOA disease-associated region. In the first round of sequencing on average 79% of bases in the target region achieved 10x coverage. This was a high level of coverage for a target enrichment experiment, but this still left 21% of the region either unsequenced or sequenced at a low level of coverage. Although the non-sequenced regions are largely repetitive in nature, which would have been masked during bait design, the missing data highlights how mutations in repetitive regions of the genome could be missed.

The second target enriched massively parallel sequencing experiment to sequence ten additional controls was completed entirely in-house, satisfying one of the targets set out at the start of this thesis. Ten fold less data were produced from one run of the MiSeq platform, resulting in 56% coverage of the target region, 23% lower than for the first HiSeq run. The data however still served the required function of reducing the number of disease segregating variants.

One interesting difference between the MiSeq and HiSeq runs was the average level of PCR product duplicates seen, which was 37% for the HiSeq run and 2% for the MiSeq

run. The high level of PCR product duplicates for the HiSeq run suggests that the level of data produced is approaching the useful limit, as any further increase in data output would likely result in further duplicates being produced which are removed by the sequencing analysis pipeline. This highlights that although competition in massively parallel sequencing technology development has led to an exponential increase in sequence data output, usefulness of the data may be limited by the method used to create sequencing libraries.

5.3.4. *CAPN1* as a candidate for LOA in the PRT

5.3.4.1. The calpain family

The calpain 1 gene (*CAPN1*) encodes an intracellular calcium dependent cysteine protease. Calpains are found throughout the plant, animal and fungal kingdoms, and belong to the papain superfamily of cysteine proteases with a catalytic site related to caspase and cathepsin family members (Barrett and Rawlings, 2001, Berti and Storer, 1995). Work to characterise calpain structurally was largely carried out with two conventional forms known as μ-calpain (*CAPN1*) and m-calpain (*CAPN2*), which subsequently was used to define the calpain family in classical and non-classical forms. Classical calpains consist of an N-terminal anchor helix region, two protease core domains (PC1 and PC2), a C2 like domain (C2L) and a penta EF-hand calcium binding domain (PEF) (Figure 5.10).

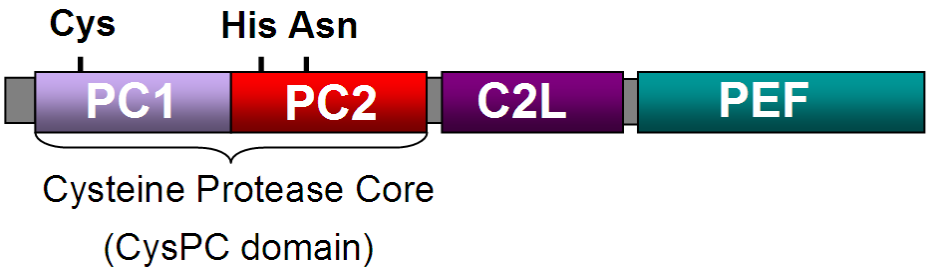


Figure 5.10 The structure of classical calpain

Classical calpain consists of two protease core domains (PC1 and PC2), a C2 like domain (C2L) and a penta EF-hand domain (PEF). The PC1 and PC2 make up the cysteine protease core. The cysteine, histidine and asparagine residues make up the catalytic site of the enzyme.

Both *CAPN1* and *CAPN2* proteins form heterodimeric structures with the small regulatory subunit CAPNS1, interacting with the fifth EF-hand motif. On binding of calcium, conformational changes result in formation of a catalytic triad of cysteine, histidine and asparagine residues and activation of the enzyme (Hosfield et al., 1999).

Calpains have limited protease activity and are therefore classified as modulatory enzymes that precisely and irreversibly cleave specific protein substrates (Sorimachi et al., 1997), although the exact physiological functions of calpains remain unclear.

5.3.4.2. Calpain gene knockouts

Advances in the understanding of calpains have been made by analysing targeted gene knock-outs. The importance of the ubiquitously expressed μ/m calpains (*CAPN1* and *CAPN2*) was shown by targeted knockout of the gene encoding the small regulatory subunit CAPNS1 (*CAPN4*), which resulted in embryonic lethality (Tan et al., 2006). Targeted deletion of mouse *CAPN2* also resulted in embryonic lethality, although *CAPN1* null mice appear phenotypically normal, apart from an observed reduction in platelet aggregation (Azam et al., 2001, Dutt et al., 2006).

5.3.4.3. Calpain associated disease in human patients

Mutations in *CAPN3* have been associated with limb girdle muscular dystrophy type 2A, a progressive atrophy and weakness of the shoulder and pelvic girdle muscles in humans (Richard et al., 1995). Calpains have been associated with neuronal necrosis, with proteolytic activity increasing as cellular calcium levels rise due to loss of homeostasis after trauma (Kampfl et al., 1996). Experimentally induced brain trauma in mice by ischaemia (reduction of blood supply to a region of tissue) was shown to result in increased proteolysis of frodin (alpha spectrin), a major cytoskeletal protein, an event hypothesised to be part of a cascade leading to neuronal cell death (Saido et al., 1993). Calpain inhibitors have therefore been suggested as potential therapeutic agents for traumatic brain injury (Bralic et al., 2012, Pignol et al., 2006). Calpains have also been linked to roles in long term potentiation and Alzheimer's disease. Despite links to neuronal death and disease, there is evidence that calpains may actually contribute to dendrite remodelling after neural injury, suggesting a maintenance role in the nervous system (Faddis et al., 1997).

5.3.4.4. The disease-associated *CAPN1* mutation

The disease-associated *CAPN1* mutation is a non-synonymous G to A base substitution at position 344 of the *CAPN1* transcript, resulting in substitution of a cysteine residue for a tyrosine (C115Y). This 155 cysteine residue is analogous to the catalytic cysteine residue that forms part of a catalytic triad with histidine and asparagine, and is therefore critical to the enzymatic properties of the protein (http://www.uniprot.org/uniprot/P07384). This suggests that substitution of this residue is likely to have a detrimental effect on enzyme activity, and a potential loss of function. Of the three candidate SNPs from resequencing of the disease-associated region, the *CAPN1* variant showed the strongest pattern of segregation on genotyping of an extended cohort of PRT individuals. An extremely high

level of conservation was seen across species for *CAPN1* orthologues at the amino acid level for the 115 cysteine residue. High levels of conservation within species for calpain family members (paralogues of *CAPN1*) further suggest the critical importance of the residue. Predictive tools suggest the mutation to be potentially pathogenic and use of Grantham and BLOSUM tables suggest tyrosine to have different chemical characteristics and cysteine to tyrosine base substitutions to be evolutionarily uncommon. Collectively the evidence presents a strong case for the *CAPN1* SNP to be deleterious.

5.3.4.5. The *CAPN1* mutation and LOA phenotype

Although genomic comparisons and predictive tools strongly suggest that the *CAPN1* mutation is likely to be pathogenic, this contradicts the clinically normal phenotype seen in the *CAPN1* null mouse. Conversely, there have been no suggestions of platelet disorders in dogs homozygous for the *CAPN1* variant, although this potential defect has not been investigated in the dog and does not appear to significantly affect bleeding times in the mouse. One possibility is that the null mouse does not have significant longevity for clinical signs to be noticed or mice are euthanised before manifestation of clinical signs. Another possibility is that the *CAPN1* gene has a slightly different role in the mouse with the calpain family having a level of redundancy, allowing other family members to compensate for loss of calpain 1 activity. A third possibility is that the *CAPN1* mutation is not the cause of LOA but in fact a marker in linkage disequilibrium with the disorder, although the exact residue changed by the *CAPN1* variant is highly suggestive.

A possible role for calpain 1 in neuronal maintenance and remodelling best fits the potential for a defective calpain 1 protein to be the cause of LOA in the PRT. LOA is largely a disease of motor neurone degeneration in the spinocerebellar tract, with Wallerian-like degeneration observed on histopathological examination. Defective maintenance mechanisms due to lack of calpain 1, leading to neurite degeneration and necrosis, would explain these observations. However, processing of substrates by calpain 1 have been implicated in neurite degeneration, confusing the roles of calpains and potentially suggesting involvement in multiple molecular processes and pathways (Demarchi and Schneider, 2007, Touma et al., 2007).

5.3.5. Exclusion of the intergenic SNP

Although the *CAPN1* SNP was considered to be the strongest candidate for the cause of LOA based on the exonic location and the predicted effect on protein function, there was insufficient evidence to formally exclude the intergenic SNP as causal. In order to exclude one of the SNPs an individual that was homozygous for the disease-associated allele, but clinically normal would need to be identified. Conversely functional work would be needed,

such as the use of a knock-out mouse model, to help confirm a SNP as the causal mutation.

5.3.6. Discordant cases

Four PRTs and four JRTs which displayed clinical signs consistent with LOA, but were not homozygous for the *CAPN1* disease-associated allele were identified in the study. Diagnosis was not changed on follow-up investigation, although full neurological examinations had not been performed to rule out other causes of disease such as inflammatory or infectious brain diseases. Six of the discordant cases were homozygous for the wild-type *CAPN1* allele and two were heterozygous. Genotyping of SNPs across the disease-associated region would be required to determine whether recombination events had occurred to potentially rule out the *CAPN1* SNP. However it is impossible to rule out the existence of a genetically distinct, but clinically very similar form of ataxia to LOA in the breed. Along with an early onset form that has been reported (personal communication, Gary Johnson, University of Missouri), this may indicate that there are potentially three genetically distinct forms of ataxia that segregate within the PRT and JRT breeds.

5.3.7. DNA testing

Given that the *CAPN1* variant is at least highly associated with LOA in the PRT, a DNA test based on *CAPN1* was launched in November 2012. The test would allow breeders to identify asymptomatic carriers and by selective breeding reduce the frequency of the disease-associated allele in the breed population. The DNA test would also potentially benefit the research programme by identifying further clinically affected individuals that are not homozygous for the disease-associated mutation. These cases would allow the effectiveness of the DNA test to be monitored and if enough cases were collected, allow an additional GWAS to be undertaken for a third potential cause of ataxia in the breed. Additionally, identification of clinically normal individuals, which are homozygous for the disease-associated allele may indicate that the *CAPN1* mutation is not causal or could imply involvement of a modifier gene affecting age of onset, as demonstrated for progressive retinal atrophy in the Miniature Long Haired Dachshund (Miyadera et al., 2012). Both scenarios would require SNP array genotyping of additional individuals to either refine the disease-associated region or to perform an association study to identify a second modifying locus.

5.3.8. Summary

Using a GWAS approach and target enriched massively parallel sequencing a disease-associated SNP in *CAPN1* has been identified. The SNP is a missense mutation causing

a cysteine to tyrosine substitution at residue 115 of the calpain 1 protein. Cysteine 115 is a highly conserved residue and forms a key part of a catalytic triad of amino acids that are crucial to the enzymatic activity of cysteine proteases. Given the function and high level of conservation, substitution of the cysteine residue is highly likely to have a negative effect on the activity of the enzyme, although functional studies would be required to assess the extent of activity loss. Loss of *CAPN1* activity as a cause of LOA is difficult to prove, although a suggested role for *CAPN1* in neuronal maintenance fits with the pathogenesis of the disease based on histopathological evidence. A DNA test has been launched as a breeding tool to help reduce the prevalence of LOA within the breed population.

Chapter

6. Neonatal cerebellar cortical degeneration in the Beagle

6.1. Background

At the Centre for Small Animal Studies at the AHT isolated and clinically distinct cases are often seen, which represent either rare conditions or previously unseen novel disorders. When a four week old Beagle presented with clinical signs of cerebellar ataxia with an inability to ambulate, a full clinical investigation was carried out to categorise the condition accurately and to rule out other underlying causes, although an inherited basis was deemed likely, based on family history. Because of the poor prognosis for the puppy the owner elected euthanasia. Prompt post-mortem investigation allowed fresh tissue samples to be obtained for histopathological and gene expression studies. No clinically similar Beagle cases had been seen previously at the AHT.

6.1.1. Clinical investigation

The clinical investigation was carried out at the AHT by Elsa Beltran and Dr Luisa De Risio (veterinary neurologists). The four week old puppy had a ten day history of severe cerebellar ataxia, and was unable to ambulate from the normal onset age of walking. The clinical signs had remained stable since then. The puppy was eating and drinking well, with no other gross abnormalities on physical examination.

Neurological examination revealed severe cerebellar ataxia, with a tendency to lean and fall towards both sides, resulting in an inability to walk without assistance. Proprioceptive positioning was normal while hopping reactions were abnormal with delayed onset of protraction and exaggerated response, once initiated. Spinal reflexes were normal in all four limbs. Cranial nerve examination revealed an absent menace response bilaterally with normal vision. Occasionally when the head was positioned in extension, spontaneous rotatory nystagmus was observed. A lesion involving mainly the cerebellum and spinocerebellar tracts was suspected. The main differential diagnoses included degenerative central nervous system disease, such as neonatal cerebellar cortical degeneration (NCCD) and less likely inflammatory/infectious central nervous system disease, metabolic disease and neoplasia. Haematology and comprehensive biochemistry investigations did not reveal any significant abnormalities. Brainstem auditory evoked responses (BAER hearing test) identified clear waves I to V (normal). Based on the severity of the clinical signs, normal haematology and biochemistry results, a degenerative condition was considered the most likely underlying cause and the breeder elected euthanasia. Post-mortem examination was performed an hour after euthanasia, and failed to reveal gross pathology. The brain *in toto* weighed 42 g, whilst the cerebellum weighed 5 g (12%, normal 10-12%). Narrowing of folia was not noted.

There were no signs of systemic illness in the sire, dam or littermates. Physical and neurological examination of the dam, sire and six clinically unaffected littermates, was carried out by Elsa Beltran at the AHT, and did not reveal any abnormalities.

The only previous litter from the same sire and dam mating consisted of seven puppies, of which two (one female and one male) had clinical signs consistent with NCCD and the remainder were clinically normal based on clinical history and video footage provided by the breeder when the puppies were eight weeks old. Both clinically affected puppies were euthanised at eight weeks of age and paraffin-embedded sections of cerebellum from the female were available for histopathological examination.

6.1.2. Histopathological investigation

Histopathologial investigations were carried out by Jennifer Stewart at the AHT. Histopathologically, the lesions were confined to the cerebellum. Examination of serial cerebellar sections of the four week old puppy identified mild loss of Purkinje cells, with increased numbers of astrocytes. Moderate numbers of Purkinje cells were shrunken with angular cell margins, hypereosinophilic cytoplasm, and condensed nuclei (Figure 6.1). Occasional associated swollen dendritic processes were identified. Spheroids were rarely seen. Mild spongiosis was present in the granular cell layer.

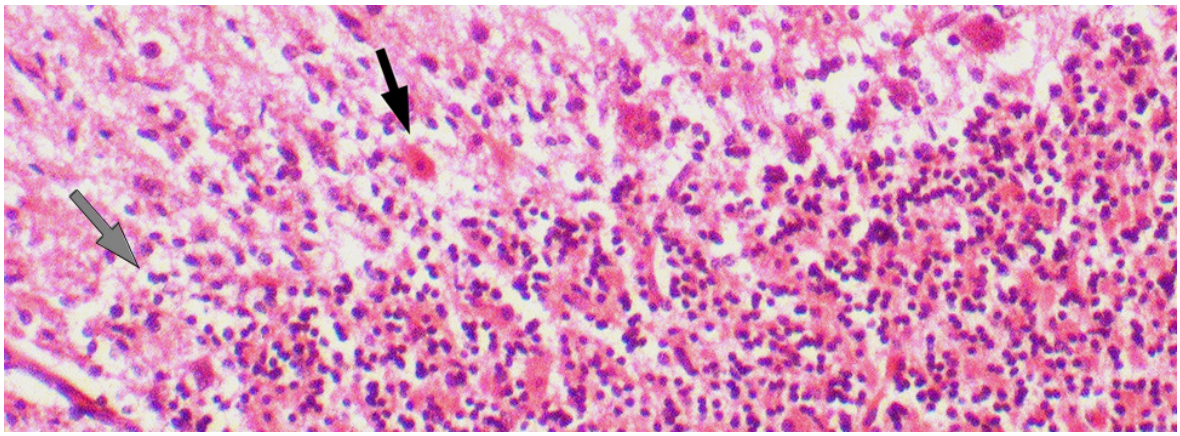


Figure 6.1 Cerebellar folia of a four week old Beagle with NCCD

Loss of occasional Purkinje cells is indicated by the grey arrow. The black arrow points to a degenerating Purkinje cell with hypereosinophilic cytoplasm and a condensed nucleus. 100x magnification. Image provided by Jennifer Stewart, Animal Health Trust.

Purkinje cell interface Bielschowsky fiber stain was performed by Dr. Kaspar Matiasek at the Institute of Veterinary Pathology, Ludwig-Maximilians University, Munich and demonstrated the subacute loss of Purkinje cells, also called “empty baskets” (Figure 6.2).

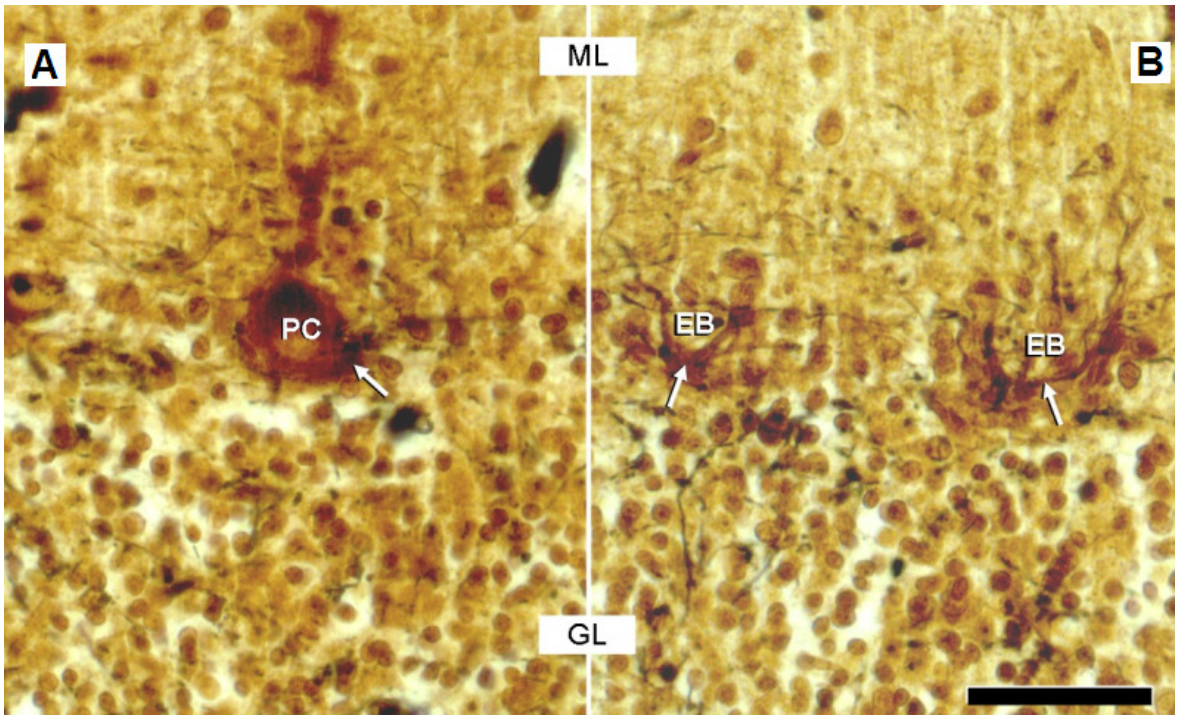


Figure 6.2 Bielschowsky staining of cerebellum tissue from the Beagle NCCD case

(A) A normal Purkinje cell. (B) Subacute loss of Purkinje cells (PC) is indicated by the presence of "empty baskets" (EB). Scale bar: 40 µm. ML: molecular layer; GL: granule cell layer. Image provided by Dr. Kaspar Matiasek, Ludwig-Maximilians University.

Examination of slides (Jennifer Stewart, AHT) made from the paraffin embedded tissues of the eight week old female from the previous litter identified marked Purkinje cell loss with rare remaining, often abnormal, Purkinje cells. Variably swollen or shrunken, hypereosinophilic Purkinje cells were identified within the remaining population. Correlating astrocytosis replaced previously lost Purkinje cells. Swollen dendritic processes and small numbers of spheroids were present. There was moderate to marked thinning of the subjacent granular layer. Cerebellar nuclei in both puppies were normal. Neuronal storage products were not identified in either puppy. The clinical and histopathological investigations confirmed the diagnosis of NCCD in both puppies.

6.1.3. Previous reports of NCCD in the veterinary literature

There have been two previous reports of conditions similar to NCCD in Beagle in the veterinary literature. Cerebellar cortical degeneration in Beagle dogs was reported by Yasuba and colleagues in 1988 (Yasuba et al., 1988). The cases in the study were born as a part of a breeding colony so the condition of the puppies could be assessed from birth. Three out of a litter of eight puppies were affected and clinical signs were first seen from three weeks of age, consistent with the case seen at the AHT. The clinical signs showed a slow progression until the affected dogs were euthanised at 14 weeks of age.

Histopathological examination revealed thinning of the folia and widened sulci. Degeneration to loss of Purkinje cells was observed, and the molecular and granular layers were reduced in thickness.

A report of cerebellar cortical abiotrophy by Kent and colleagues was published in 2000 (Kent et al., 2000). The case was a four and a half month old female, which was one of two affected individuals from a litter of five. The owners reported that the affected puppies were unable to ambulate normally from the onset of walking, but no progression of clinical signs was noted by the breeder. At four and a half months of age the dog was ambulatory but had clinical signs of bilateral cerebellar ataxia, with a dysmetric gait and an inability to regulate the rate and range of movement, with both hypermetric and hypometric striding. Balance was affected, and the dog would fall in all directions (ie sideways, forwards and backwards). Purkinje cell loss was observed histopathologically, which was consistent with the previously reported case and the AHT case. All areas of the cerebellum were more uniformly affected than reported by Yasuba and colleagues, although this may be related to age at examination.

6.1.4. Study approach

The introduction of massively parallel sequencing techniques has facilitated the use of many new DNA sequencing applications. Genome-wide sequencing such as whole genome resequencing, whole exome sequencing and genome-wide mRNA sequencing (mRNA-seq) can now be achieved in a single, cost effective experiment. Next generation sequencing technologies coupled with target enrichment techniques (Asan et al., 2011) allow for the simultaneous sequencing of several exomes that can be used to scan coding regions of the genome for disease-associated mutations in a case-control approach. An example of a successful use of this method is the identification of *AFG3L2* mutations in spastic ataxia-neuropathy syndrome (Pierson et al., 2011). In the spastic ataxia-neuropathy syndrome study, DNA from just the father, mother and two affected siblings of a consanguineous family was subjected to whole exome sequencing. Although an average of 8,585 missense and 87 nonsense changes were identified per individual, use of an autosomal recessive model and exclusion of SNPs catalogued in the dbSNP database, left just two candidate variants, only one of which was plausibly causal, illustrating the potential use of genome-wide sequencing approaches using a minimal set of individuals to identify causal variants.

Genome-wide mRNA sequencing is widely used in quantitative gene expression studies and can be used to improve genome annotation. This approach has recently been

suggested for the canine genome which relies heavily on predictive methods for genome annotation (Derrien et al., 2012). Genome-wide mRNA sequencing has the potential to confirm exon boundaries, classify previously undiscovered genes, and identify gene isoforms resulting from alternative splicing, but its use as a method of identifying coding changes associated with inherited disease is not widely reported.

Following identification of a case of NCCD, an mRNA-seq experiment was planned as a method of candidate gene sequencing. Ataxia in humans is well characterised and a large number of genes and loci have been associated with both autosomal dominant (SCA1-36) and autosomal recessive (SCAR1-12) spinocerebellar ataxias. The mRNA sequencing data were generated to interrogate canine orthologues of spinocerebellar ataxia associated genes identified from human studies with the aim of identifying the causal genetic variant responsible for NCCD in the Beagle.

6.1.5. Summary

A four week old Beagle with clinical signs of cerebellar ataxia was seen at the AHT. The prognosis for the puppy was poor and the owner elected euthanasia. Neurological and histopathologal examinations led to a diagnosis of NCCD. The post-mortem examination was performed within an hour of euthanasia, allowing fresh cerebellum tissue to be stored in RNAlater. The standard procedure for investigating genetic diseases is to collect a large case-control cohort, before undertaking a GWAS approach to identify disease-associated loci, which can then be screened for causal mutations. However, the tissue sample obtained allowed the trialling of mRNA-seq as a method of candidate gene sequencing in an attempt to identify the causal mutation for this disorder and develop a diagnostic DNA test.

6.2. Results

6.2.1. Pedigree analysis

Of a total of 14 dogs from the same sire and dam mating, 3 dogs (2 male and 1 female) were affected by NCCD, which was consistent with an autosomal recessive mode of inheritance (Figure 6.3).

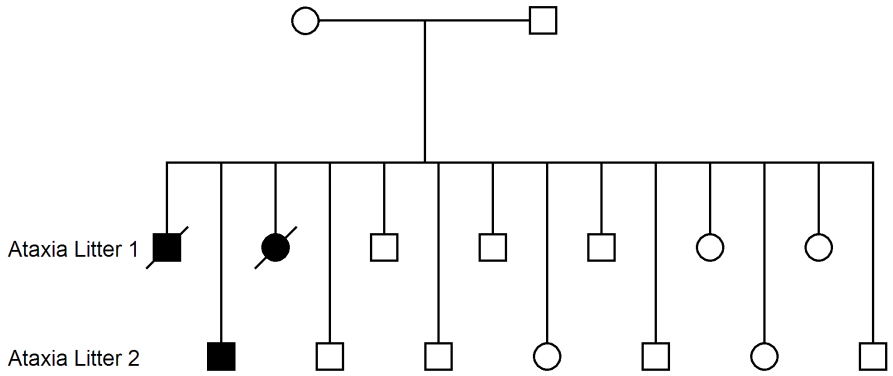


Figure 6.3 Pedigree of the three Beagle NCCD puppies

All three affected puppies were from the same sire and dam mating, and included both male and female puppies, with ratios of affected to unaffected littermates suggestive of an autosomal recessive mode of inheritance.

6.2.2. RNA integrity

A crucial factor for successful mRNA-seq experiments is integrity of RNA. If the RNA is degraded, sequencing libraries are biased towards the 3' end when using a polyA capture method of purifying the mRNA from total RNA. RNA for use in the study was DNase treated to ensure no sequenced library fragments originated from genomic DNA. Five total RNA samples were extracted from cerebellum tissue of an ataxic Parson Russell Terrier, two ataxic Norwegian Buhunds, the NCCD case and from a neurologically normal dog of unknown breed. RNA integrity was assessed using the Agilent Bioanalyser RNA assay at Cambridge Genomic Services, Cambridge (Figure 6.4).

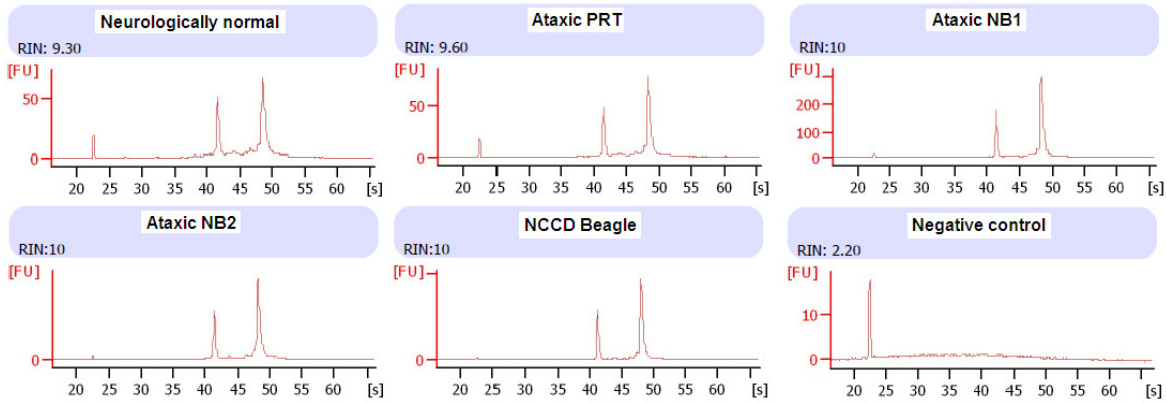


Figure 6.4 RNA integrity numbers (RINs)

Cerebellum RNA samples were assayed on the Agilent Bioanalyser to assess concentration and RNA integrity. The peaks on the electropherogram after 42 and 48 seconds represent the 18s and 28s rRNA respectively. An RNA integrity number (RIN) between 0 and 10 is given with 0 indicating completely degraded RNA or absence of RNA and 10 indicating fully intact RNA.

Illumina recommend use of RNA with RNA integrity number (RIN) values of greater than 7.00 for use in library preparation. All total RNAs had a RIN of 9.30 or greater and were suitable for library preparation.

6.2.3. Libraries

Six libraries were constructed in total. Initially a trial library was made using the NCCD Beagle RNA which would be multiplexed with SureSelect libraries for sequencing to gauge the effectiveness of library preparation and to calculate the amount of data required to successfully use mRNA-seq as a method of sequencing the selected candidate genes. After successful sequencing of the trial library, a second batch of libraries was constructed, one for each of the samples described in section 6.2.2., with a ~150 bp insert size.

6.2.4. Assessing library content by cloning

Libraries were assessed for mRNA enrichment success by molecular cloning followed by Sanger sequencing. A summary of cloning results is shown in Appendix 13. No cloned fragments were ribosomal RNA in origin suggesting successful polyA enrichment. A high percentage of reads mapped to 3' UTR regions of genes (59%), which was suggestive of a possible 3' bias in the libraries.

6.2.5. Illumina sequencing of mRNA libraries

Sequencing of mRNA-seq libraries was performed in three stages. Initially the trial Beagle library was multiplexed with SureSelect libraries constructed for unrelated projects and sequenced on the Illumina HiSeq 2000 platform. The aim was for 5% of the reads generated to be attributed to mRNA-seq and the remaining reads for unrelated projects.

The sequencing lane produced a 19.68 Gb dataset with 1.39 Gb (7.1%) belonging to the mRNA-seq project.

In the second stage the PRT mRNA-seq library (~150 bp inserts) was multiplexed with a second batch of SureSelect libraries, aiming for ~10% of the data to be attributed to mRNA-seq. The sequencing lane produced a 20.85 Gb dataset with 2.15 Gb (10.33%) belonging to mRNA-seq.

To produce a larger dataset for all five libraries and to resolve a demultiplexing issue which arose when multiplexing the mRNA-seq libraries with the SureSelect libraries, all five libraries were sequenced on a dedicated lane of a HiSeq 2000. A 17.38 Gb dataset was produced, which was equally shared between all five libraries. Datasets are summarised in Table 6.1.

Table 6.1 Summary of the mRNA-seq datasets

mRNA-seq datasets were produced through three runs of single lane sequencing on the Illumina HiSeq 2000. For runs 1 and 2, partial lanes were allocated to mRNA with the remainder of the lane allocated to sequencing of SureSelect libraries. For run 3 the entire lane was dedicated to mRNA-seq. PRT – PRT LOA case; NB – Norwegian Buhund ataxia case; U – neurologically normal individual of unknown breed.

mRNA-seq dataset	Sequencing lane content	Libraries sequenced	mRNA-seq dataset size (Gb)
1	Multiplexed SureSelect & mRNA-seq	Beagle cerebellum (Trial library)	1.39
2	Multiplexed SureSelect & mRNA-seq	PRT cerebellum	2.15
3	mRNA-seq only	Beagle, PRT, NB1, NB2, and U	17.38

6.2.6. Assessing the quality of mRNA-seq data

The mRNA-seq data were available from the Wellcome Trust Centre of Human Genetics, Oxford as pre-aligned bam files. The quality of the aligned data were assessed by viewing in IGV. No intergenic reads were detected suggesting RNA extraction and DNase treatment had successfully removed genomic DNA from the library preparation. Data were assessed for 3' read bias by visualising long transcripts. Figure 6.5 shows the 5', central and 3' regions of the dystrophin gene (*DMD*). The *DMD* gene has a long 14 kb transcript spanning a genomic region of 2.2 Mb. The gene is primarily expressed in skeletal muscle, but some expression is seen in brain tissue. Although a four fold drop in read depth was seen from the 3' to the 5' end of the gene, coverage was achieved for all exons,

suggesting exonic coverage would be achieved for other genes of a similar length or shorter, and expressed at a comparable level.

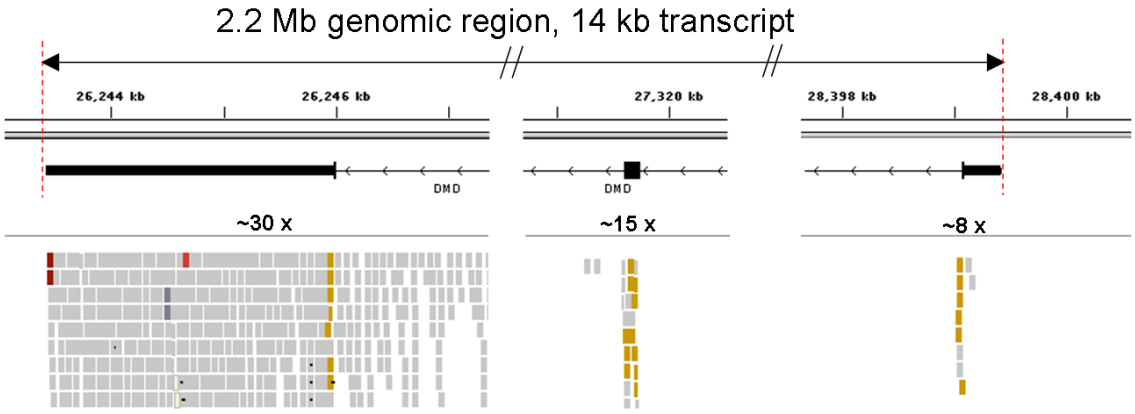


Figure 6.5 Assessment of 3' bias by visualising a long transcript in IGV

Long transcripts such as the dystrophin (*DMD*) transcript were used to assess the level of 3' bias. Over the longest transcripts there appeared to be 3-4 fold decrease in coverage from the most 3' to 5' regions.

As several regions of the genome were assessed visually in IGV, it was noticed that there appeared to be reads mapping to the exact position of SureSelect baits in the mRNA-seq dataset (Figure 6.6). Suggested causes of the problem included contamination of the pre-amplified mRNA-seq library with a SureSelect library, faulty allocation of reads during demultiplexing or cross-contamination of indexing primers.

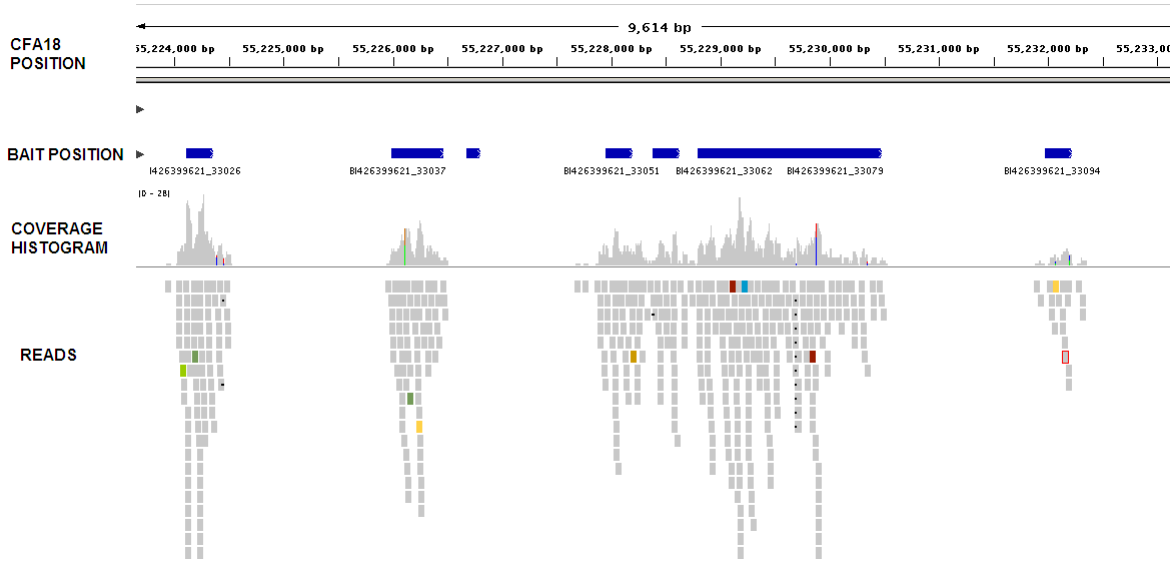


Figure 6.6 Presence of SureSelect reads in mRNA-seq data

It was noted that in some intergenic regions it appeared that data in the position of SureSelect baits was present in mRNA-seq datasets. In exonic regions this issue presented a problem in distinguishing mRNA-seq reads from SureSelect reads, especially when expression levels for a particular gene were low.

In the first mRNA-seq experiment the trial Beagle cerebellum library was multiplexed with SureSelect libraries for the PRT and IS disease-associated regions. Reads in intergenic regions were seen at the location of SureSelect baits. In the second mRNA-seq experiment the PRT cerebellum library (part of the batch of five mRNA-seq libraries) was multiplexed with SureSelect libraries for an unrelated project. Reads in intergenic regions were also seen at the location of SureSelect baits, but contamination of the pre-amplified mRNA-seq library with material from the SureSelect libraries could be ruled out as the mRNA-seq libraries were made prior to synthesis of the SureSelect libraries. The problem was resolved by sequencing the mRNA-seq libraries independently on a single lane of an Illumina HiSeq 2000. This also presented the opportunity to generate mRNA-seq data for a control cerebellum sample from a clinically normal dog and mRNA-seq data for two libraries attributed to an unrelated project (Norwegian Buhund libraries).

6.2.7. Candidate gene selection

A review of the scientific literature and the Online Mendelian Inheritance in Man (OMIM) database indicated 41 human ataxia loci had been identified for which 28 causal genes had been characterised. Twenty seven of the genes causing human ataxia had orthologous canine genes, and these genes were considered as candidates. Candidate genes could only be considered if they were expressed sufficiently to give complete exonic coverage from the mRNA-seq dataset.

6.2.8. Candidate gene analysis

The dataset used in the initial analysis was created from the Beagle trial library and consisted of 13.64 millions reads, with 97.1% of the reads mapping to the dog genome. The dataset was sufficient for complete exonic coverage of 24 of the 27 candidate genes. No polymorphisms were identified in 11 of the genes. Three genes contained polymorphisms in non-coding regions only. Heterozygous SNPs were identified in four genes, excluding association with NCCD. Two genes contained synonymous SNPs. Non-synonymous changes were identified in *ITPR1* (chr20:15,780,361G>C;p.E2491Q), *BEAN1* (brain expressed, associated with NEDD4, 1) (chr5:85,782,181C>T;p.R247Q), and *ADCK3* (aarF domain containing kinase 3) (chr7:41,059,467A>G;p.S328P). For *ITPR1* and *ADCK3* the non-reference residue is highly conserved amongst vertebrate species and therefore could be ruled out as causal. Alignment across vertebrate species at the site of the *BEAN1* polymorphism indicated that both glutamic acid and glutamine residues occur naturally, enabling that variant to also be excluded. An 8 bp deletion was detected in exon 29 of *SPTBN2* (chr18:53,691,704_53,691,711del) and is shown in Figure 6.7. The deletion was confirmed by Sanger sequencing. The frameshift is predicted to result in a run of 27 aberrant amino acids, followed by premature termination with a 410

amino acid truncation (p.G1952insRDRGQGRPLLLMHRHGAGAACQEPLCS*). A summary of candidate genes and variants is shown in Appendix 14.

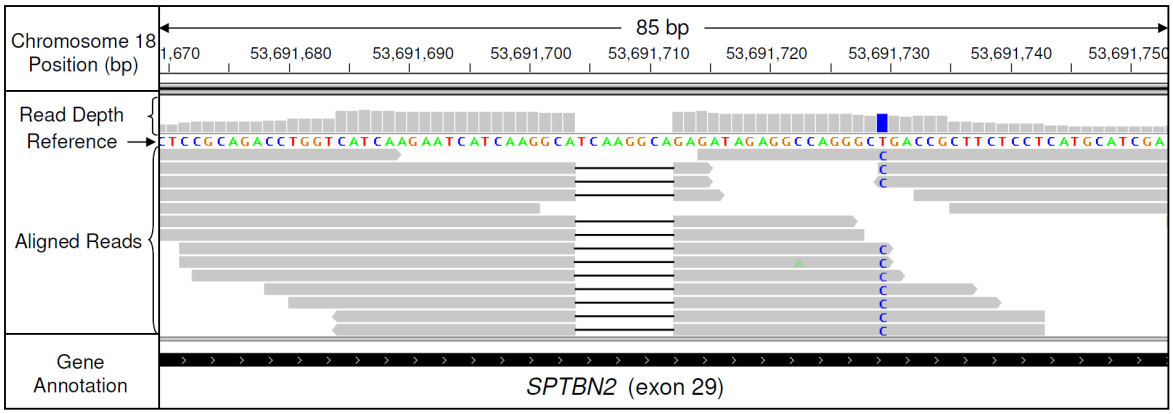


Figure 6.7 Identification of an 8 bp deletion in *SPTBN2*

Display in IGV of the 8 bp deletion in exon 29 of *SPTBN2.*

6.2.9. Genotyping

Genotyping experiments were performed to establish whether the 8 bp *SPTBN2* deletion could be potentially causal. Example results are shown in Figure 6.8. The sire and dam of the affected dogs were both heterozygous for the 8 bp deletion, and out of the ten clinically unaffected siblings tested, seven were heterozygous and three were homozygous for the wild-type allele. DNA extracted from FFPE tissue of a previous NCCD case from the same sire and dam mating was homozygous for the deletion. Seven other clinically unaffected half-siblings, with the same sire as the affected dogs, were either heterozygous or wild-type homozygous. An additional 145 Beagles, which were collected for an unrelated project and clinically normal with respect to NCCD, were also genotyped. Eight dogs were heterozygous for the deletion, and the remaining 137 dogs were homozygous wild-type, in full concordance with the mutation being causal. In addition 513 dogs from 37 other breeds were also genotyped; all were homozygous for the wild-type allele.

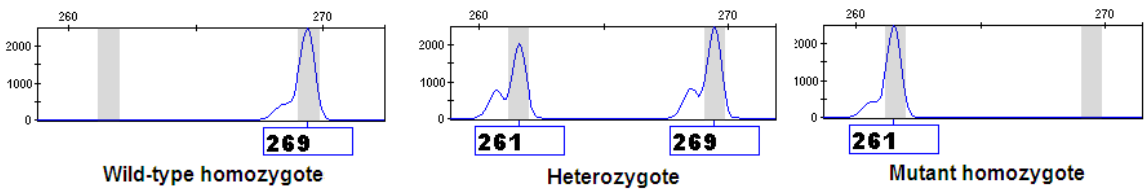


Figure 6.8 NCCD diagnostic test genotype display

The 8 bp deletion created a PCR product fragment length polymorphism, which individuals could be genotyped for as a method of diagnostic testing.

6.2.10. qPCR assessment of *SPTBN2* levels

Limited qPCR experiments to assess the expression levels of *SPTBN2* in affected and normal cerebellum tissues suggested a 63 fold reduction in *SPTBN2* transcript levels in the affected Beagle (Table 6.2).

Table 6.2 Relative expression analysis data

Expression levels of *SPTBN2* were measured relative to the reference genes beta-actin (*ACTB)* and tata-box binding protein (*TBP*) using qPCR. Fold change was calculated based on changes in threshold cycle (Ct) measurements within (ΔCt) and between (ΔΔCt) the case and control.

Sample Name	Assay Name	Ct Mean	Std Dev.	ΔCt	ΔΔCt	Fold Δ	Mean Fold Δ
Control	*ACTB*	21.54	0.01	2.14			
Control	*TBP*	24.34	0.01	-0.65			
Control	*SPTBN2*	23.69	0.07				
Beagle NCCD	*ACTB*	17.54	0.03	8.44	6.30	79 x	
Beagle NCCD	*TBP*	21.12	0.05	4.86	5.52	46 x	63 x
Beagle NCCD	*SPTBN2*	25.98	0.05				

6.2.11. Western blot analysis of *SPTBN2* protein

Using western blot analysis with primary antibodies targeting the N-terminal region of beta-III spectrin, no full length or truncated beta-III spectrin could be detected in cerebellum tissue of the NCCD case, suggesting expression of the protein may have been abolished (Figure 6.9).

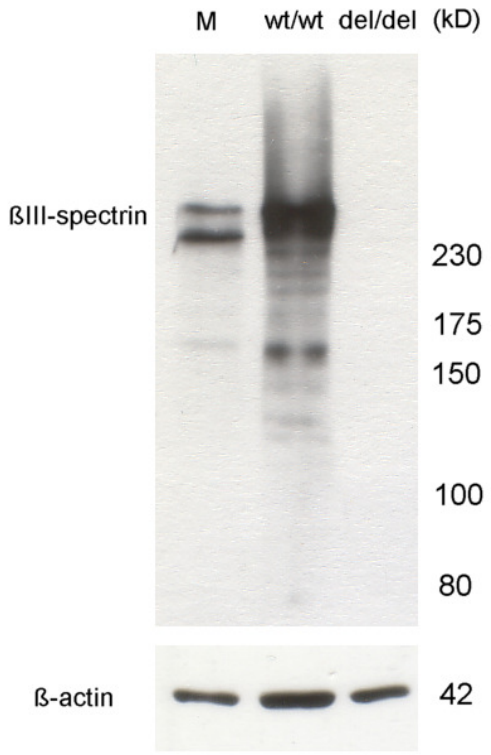


Figure 6.9 Western blot analysis of beta-III spectrin

Western blot analysis of wild-type individual (wt/wt) and affected Beagle (del/del) cerebellum tissue homogenates revealed no detectable beta-III spectrin in the affected Beagle cerebellum. Mouse cerebellum (M) was used as a control and beta-actin as a loading control.

6.2.12. Genome-wide comparison of expression levels

The availability of a control cerebellum sample allowed for a genome-wide case-control comparison of gene expression levels. Based on the comparison between the NCCD Beagle cerebellum and the control cerebellum a 10.8x reduction in *SPTBN2* expression was indicated. This expression change was ranked the 88th largest reduction in expression, and 10th most significant change based on P-values genome-wide (Table 6.3). Data analysis was performed at the Wellcome Trust Centre for Human Genetics, University of Oxford.

Table 6.3 Top 20 most significant changes in gene expression

20 most significant differences in gene expression between the control cerebellum and the Beagle NCCD cerebellum. Fold change is shown as fold decrease in expression. There are 30,194 annotated genes for the CanFam2 genome build.

Rank	Gene ID	chr	position	baseMean	baseMeanA	baseMeanB	foldChange	P value	P Adjusted
1	*PLK5*	20	60,478,581	4160	5	8315	1663.67	9.06E-86	1.76E-81
2	*TTR*	7	60,920,284	5368	63	10672	169.48	8.58E-78	8.34E-74
3	*DSP*	35	10,486,761	1301	48	2555	53.25	3.06E-67	1.98E-63
4	*KIAA1199*	3	59,591,949	1353	2598	107	0.04	3.45E-54	1.68E-50
5	*THY1*	5	17,512,680	6863	1250	12476	9.98	2.16E-52	7.41E-49
6	*USH1G*	9	8,528,280	989	90	1887	20.98	2.29E-52	7.41E-49
7	*SCN1B*	1	120,412,273	13827	2802	24852	8.87	1.20E-50	3.33E-47
8	*FAT4*	19	18,332,210	1590	3067	114	0.04	7.46E-50	1.81E-46
9	*FBLN2*	20	6,920,237	2937	500	5375	10.75	1.14E-48	2.45E-45
10	*NAV1*	7	4,136,024	1573	2977	169	0.06	1.81E-48	3.53E-45
11	*SPTBN2*	18	53,664,219	5414	917	9912	10.81	5.89E-47	1.04E-43
12	*SEMA6A*	11	9,123,714	2141	3961	321	0.08	5.16E-45	8.37E-42
13	*NYNRIN*	8	7,374,533	678	1274	82	0.06	6.43E-43	9.62E-40
14	*Novel*	Un	47,788,156	33270	8388	58152	6.93	1.71E-42	2.38E-39
15	*IFI44*	6	71,424,246	1145	2207	83	0.04	1.94E-42	2.51E-39
16	*Novel*	31	11,457,222	1863	3409	317	0.09	4.96E-41	6.03E-38
17	*C3orf18*	20	41,801,054	1243	185	2301	12.44	5.34E-41	6.11E-38
18	*AOAH*	14	50,856,684	602	31	1172	37.83	8.26E-41	8.93E-38
19	*COL6A1*	31	41,618,379	1082	2040	124	0.06	2.50E-40	2.55E-37
20	*OAS1*	26	13,402,378	1197	2271	122	0.05	2.87E-40	2.79E-37

Independent analyses of genome-wide gene expression were carried out for the PRT transcriptome against the control transcriptome and the NB transcriptomes against the control transcriptome. Fold change values obtained were compared to the top 20 genes showing the largest fold changes for the Beagle transcriptome (Table 6.4). Although there are some Beagle specific expression changes, many of the same genes show large changes in expression for the PRT and NB, suggesting that the changes may due to the small sample size, rather than being of true significance.

Table 6.4 Fold change comparisons between PRT, NB and BE libraries

Comparison of expression fold changes associated with the PRT, NB and BE transcriptomes against the control transcriptome. Fold change is shown as fold decrease in expression.

Rank	Gene ID	Chr	Position	PRT vs control fold change	NB vs control fold change	BE vs control fold change
1	PLK5	20	60,478,581	1.65	2.65	1663.67
2	TTR	7	60,920,284	51.35	34.45	169.48
3	DSP	35	10,486,761	1.57	22.44	53.25
4	KIAA1199	3	59,591,949	2.39	0.09	0.04
5	THY1	5	17,512,680	-	-	20.98
6	USH1G	9	8,528,280	1.39	2.35	9.98
7	SCN1B	1	120,412,273	1.82	2.27	8.87
8	FAT4	19	18,332,210	-	0.12	0.04
9	FBLN2	20	6,920,237	2.14	1.64	10.75
10	NAV1	7	4,136,024	0.53	0.17	0.06
11	SPTBN2	18	53,664,219	0.69	0.50	10.81
12	SEMA6A	11	9,123,714	0.45	0.25	0.08
13	NYNRIN	8	7,374,533	0.05	0.04	0.06
14	Novel	Un	47,788,156	4.16	5.05	6.93
15	IFI44	6	71,424,246	-	-	0.04
16	Novel	31	11,457,222	0.29	0.23	0.09
17	C3orf18	20	41,801,054	1.47	2.26	12.44
18	AOAH	14	50,856,684	1.59	7.86	37.83
19	COL6A1	31	41,618,379	0.57	0.16	0.06
20	OAS1	26	13,402,378	0.46	0.39	0.05

6.3. Comments and conclusions

6.3.1. Study approach

Genome-wide mRNA sequencing of a single Beagle NCCD case was used as a method of candidate gene sequencing. A number of factors made the choice of genome-wide mRNA sequencing an appropriate study approach. DNA was available from only two cases, therefore to perform a GWAS additional cases would have been needed, which would have delayed the outcome of the study. Very few cases of NCCD have been reported in the UK, so even collecting a modest cohort of 12 cases and 12 controls would have taken a considerable length of time, whereas an approach using a single case allowed the project to commence immediately. The phenotype for NCCD was both severe and early onset. With a severe disease it is reasonable to suspect a causal mutation with a severe consequence, such as one that causes a significant change in gene expression or that causes a deleterious effect on the function of a protein. The number of affected individuals across the two litters was highly suggestive of a simple autosomal recessive mode of inheritance. This meant that a single homozygous gene mutation was probably responsible for the condition, and because of LD, genes which contained heterozygous polymorphisms could be ruled out immediately. A fresh tissue resource was available for the case. Post-mortem examination was carried out immediately after euthanasia, and cerebellum tissue stored immediately in RNAlater solution. This enabled highly intact RNA to be extracted with a RIN value of 10. And finally a large number of genes had previously been associated with ataxia in human studies, providing a number of candidate genes to be investigated.

6.3.2. Advantages of mRNA-seq

As well as being an appropriate approach for studying NCCD in the Beagle, mRNA-seq has additional advantages for use in mutation identification. As previously stated, the sample collection stage required when performing a GWAS can be avoided when using the mRNA-seq approach, which shortens study time-frames especially when cases are rare. The approach has advantages over exome enrichment and sequencing methodology, in that the method can be performed in all species with reference genome sequence builds, without the need for a proprietary kit. The mRNA-seq approach is also not dependent on reference genome annotation, which may be inaccurate or incomplete in some species. In addition a much smaller dataset is required for an mRNA-seq approach in comparison to using whole genome sequencing. mRNA-seq also gives some information about gene expression levels, although this can be a disadvantage if disease causing genes show limited expression or are drastically reduced in diseased tissues.

Because only an mRNA-enrichment stage is required followed by a standard library preparation the approach is highly cost-effective. Only a small amount of data is required (for example, data generated from a partial lane of an Illumina HiSeq 2000 is adequate) and therefore small scale mRNA-seq approaches can “piggy-back” by multiplexing with other projects requiring larger amounts of data, such as SureSelect target enrichment sequencing experiments.

6.3.3. The *SPTBN2* gene

Using mRNA-seq an 8 bp deletion in the *SPTBN2* gene encoding beta-III spectrin was identified, that segregated consistently with NCCD in the Beagle. Spectrins are a family of cytoskeletal proteins, with tetrameric structures comprising two alpha and two beta subunits, with diversity and specialisation of function. Spectrins are important structural components of the plasma membrane and play a significant role in restricting and stabilising membrane spanning proteins within specific subdomains of the plasma membrane. The spectrin cytoskeleton was first discovered in erythrocytes and has since been identified in a variety of cells (Bennett and Baines, 2001). Beta-III spectrin is primarily expressed in the nervous system and the highest levels of expression are found in Purkinje cell soma and dendrites (Sakaguchi et al., 1998). Beta-III spectrin has been shown to stabilise the glutamate transporter *EAAT4* at the plasma membrane of the Purkinje cells (Jackson et al., 2001), facilitate protein trafficking by linking the microtubule motor to vesicle-bound cargo (Holleran et al., 2001) and maintain a high density of sodium channels within the soma and dendrites of Purkinje cells (Perkins et al., 2010). Beta-III spectrin is critical for development of Purkinje cells (Gao et al., 2011). The identification of an 8 bp deletion in *SPTBN2*, a gene associated with spinocerebellar type 5 (SCA5) in humans (Ikeda et al., 2006), that fully segregates with the disease provides a strong candidate variant for NCCD in the Beagle.

6.3.4. Human mutations in *SPTBN2*

In humans, three mutations in *SPTBN2* have been shown to cause autosomal dominant spinocerebellar ataxia type 5 (SCA5) (Ikeda et al., 2006). The causal mutations identified include two in-frame deletions of 39 and 15 bp which alter the structure of the 3rd of 17 spectrin repeats, and a single base pair substitution causing an amino acid change (L253P) in a highly conserved region of the calponin homology domain. The consequence of the two in-frame deletions in beta-III spectrin is predicted to be disruption of the highly ordered triple alpha helical structure of the spectrin repeat, causing conformational changes in the tetrameric alpha-beta spectrin complex (Ikeda et al., 2006). Studies suggest the resulting mutant protein may affect the localisation of EAAT4 and GluRδ2, one possible outcome of which is glutamate signalling abnormalities and Purkinje cell

death (Ikeda et al., 2006). The L253P missense mutation has been shown to result in loss of interaction with the Arp1 subunit of the dynactin-dynein complex, affecting the role of beta-III spectrin in vesicle trafficking, preventing transport of both beta-III spectrin and EAAT4 to the cell membrane from the Golgi apparatus in Purkinje cells causing cell dysfunction and death (Clarkson et al., 2010).

More recently a homozygous nonsense mutation in *SPTBN2* has been associated with developmental ataxia and cognitive impairment in humans (Lise et al., 2012). Three individuals from a consanguineous family were observed to have the disorder. Onset of walking was typically delayed to seven years of age and an obvious dysmetric gait was observed, consistent with the affected Beagle puppy's inability to ambulate. All affected individuals had intelligence quotient (IQ) scores falling within the learning disabled range. Genome sequencing was performed to rule out any other potential genetic causes of the cognitive impairment, and use of a mouse model confirmed a role for *SPTBN2* in both cognitive development and function. No obvious cognitive defects were observed for the affected Beagle puppy, although cognitive function was not fully investigated.

6.3.5. Beta-III spectrin knock-out mice

Experimentally induced beta-III spectrin deficiency in mice from two independent studies resulted in phenotypes that resemble NCCD in Beagle dogs (Perkins et al., 2010, Stankewich et al., 2010). One beta-III spectrin deficient strain was produced by targeting replacement of exon 3 to 6 of *SPTBN2* with the neomycin-resistance gene, resulting in a frameshift and a premature stop codon in exon 7. As a result no full length beta-III spectrin is produced in beta-III spectrin deficient mice, although a low level of near full length protein is produced due to novel exon 1 (rather than exon 2) to exon 7 splicing (Perkins et al., 2010). Homozygous beta-III spectrin deficient mice develop characteristics of progressive cerebellar ataxia from a few weeks of age with cerebellar atrophy and Purkinje cell loss. In the parallel study the beta-III spectrin deficient mouse strain is the result of beta-geo insertion between exons 25 and 26 resulting in premature termination in spectrin repeat 14, which is closer to the position of the Beagle mutation, although results in the loss of the ankyrin binding domain (Stankewich et al., 2010). The beta-III spectrin deficient mice from this study display a mild non-progressive ataxia by 6 months and a myoclonic seizure disorder by one year (Stankewich et al., 2010). It is apparent that onset of ataxia is later for the beta-III spectrin deficient mouse in comparison to Beagle NCCD cases, with mice not showing significant signs of ataxia until six months of age (past sexual maturity). This is more comparable to the human disease, though the differences in the modes of inheritance suggest different mutational effects. Deficient mice also show

only a mild ataxia and remain ambulatory, while the dogs described are more severely affected both in terms of degree of ataxia and Purkinje cell loss. In the study by Stankewich and colleagues, no Purkinje cell loss was documented in mice by 18 months of age, only atrophy of the dendritic arbor. Disparity in phenotype between species may suggest differences in cerebellar development, function, and potentially the involvement of beta-III spectrin. Further understanding of canine cerebellar function would be required to shed light on the described differences and common principles.

It has been shown that heterozygous mice, generated by exon 2-6 replacement, do not display any characteristics of cerebellar ataxia (Clarkson et al., 2010), in common with heterozygous dogs in the Beagle population, suggesting that SCA5 in heterozygous humans is caused by dominant negative effects of mutant beta-III spectrin, rather than haploinsufficiency. Histopathological examination in heterozygous mice revealed normal size and morphology of the cerebellum and immunostaining studies showed no changes in Purkinje cell morphology. These histopathological findings cannot be correlated with heterozygous Beagle dogs as none underwent post-mortem examination and all of them are currently alive and clinically unaffected. Interestingly, slight motor impairments were reported for heterozygous mice generated by beta-geo insertion between exons 25 and 26, perhaps indicating that the truncated protein is having a slight dominant negative effect, and illustrates how disease progression is dependent on the positioning of *SPTBN2* mutations.

6.3.6. Deletion mechanism

The 8 bp deletion in the dog is located at a tandem repeat sequence, suggesting homologous recombination as the deletion mechanism. The position of a SNP (c.5580T>C) 18 bp downstream of the deleted sequence removes a possible termination site for the mutant protein and extends the sequence of potential aberrant amino acids from 6 to 27.

6.3.7. Expression analysis

6.3.7.1. Quantitative PCR approach

Expression analysis was limited, due to the availability of only one case and one control. Results are suggestive of a 68x reduction in the relative expression levels of *SPTBN2* in the NCCD case cerebellum, which may be due to nonsense mediated decay. Even though *SPTBN2* expression is greatly reduced in NCCD affected cerebellum tissue, sufficient read depth from the mRNA-seq experiment was still achieved, because of the high levels of *SPTBN2* expression normally seen in the cerebellum.

6.3.7.2. Genome-wide expression analysis

Because of the availability of cerebellum mRNA-seq data from a clinically unaffected individual from the third mRNA-seq dataset, genome-wide expression analysis was carried out. Insufficient cases and controls were available to perform a complete investigation, but results did show that even on a genome-wide scale with a one case and one control approach the change in expression level of *SPTBN2* was highly significant and one of the greatest fold changes genome-wide. Further investigation of highly significant fold changes by comparing other case-control analyses to the Beagle case-control analysis showed that many of the significant changes were shared. Although these changes could account for compensatory gene regulation changes in ataxic brains, it is also very possible that these changes were specific to the control. This is a consequence of some genes that can show large fluctuations in gene expression and further controls would be needed to mitigate against this problem. It is important to reiterate that the experimental design was far from perfect. The control dog in the investigation was of unknown breed. Differences in cerebellum gene expression across breeds has not been investigated, but it is feasible that dogs that are bred for increased balance and poise, such as herding dogs, may have changes in cerebellum gene expression as a result of this selection process. Results of the qPCR expression analysis suggested a higher fold change in *SPTBN2* gene expression than the mRNA-seq data. qPCR is often considered the gold standard when investigating gene expression. It is possible that there was a slight increase in 3' bias for the control library, this could result in fewer reads across the entire transcript and would result in less of a difference between the NCCD Beagle and the control read count across the gene.

6.3.8. Protein analysis

Further to a reduction in mRNA levels, no full length or truncated beta-III spectrin was detectable in NCCD affected cerebellum tissue by western blot analysis. This may indicate that the 8 bp deletion results in a full knock-out of *SPTBN2*. A full gene knock-out eliminates the possibility of a dominant negative effect that could be caused by a truncated form of the beta-III spectrin protein, which is consistent with the observation that heterozygous dogs do not show clinical signs.

6.3.9. NCCD in other breeds

Although NCCD is likely to be heterogeneous in different canine breeds, screening for the *SPTBN2* deletion in non-Beagle cases has not been investigated to confirm this. It is possible that the mutation could exist at very low frequencies in other breed populations,

especially those closely related to the Beagle, but extensive screening of large numbers of individuals would be required to fully investigate this possibility.

6.3.10. Summary

Genome-wide mRNA-seq was used to successfully identify the causal mutation for NCCD in the Beagle. This novel study approach was applicable due to the availability of a fresh tissue resource and many suitable candidate genes, and allowed genetic investigations to commence in a time and cost effective manner without the need to collect a large case-control sample cohort.

Chapter

7. General discussion

7.1. Overview

This study aimed to use the dog as a model to demonstrate the advances in genetic mapping and DNA sequencing techniques. This thesis has demonstrated how the available techniques can be adapted to suit the particular disease under investigation, and how choice of technique can be dependent on sample availability and resources. In this section the choice of techniques and study limitations is discussed. Potential for further work is also discussed, highlighting how advances in technology may influence strategic choices in the future.

7.2. Genetic mapping strategies

Two genome-wide mapping strategies were used during the course of the investigations. These were homozygosity mapping using microsatellite markers and genome-wide association analysis using a high-density SNP array. Both techniques were successfully used to map disease loci to distinct regions of the canine genome. The homozygosity mapping approach demonstrated how a very small sample cohort of six cases and six controls can be sufficient to successfully map an autosomal recessive condition in the dog. As the technique uses PCR followed by fragment length analysis using capillary electrophoresis, both the experimental work and data analysis could be carried out in-house. The high levels of LD in the dog were essential for the homozygosity mapping study to succeed. Only 300 genome-wide microsatellites were used for the analysis, meaning that on average markers were spaced 8 Mb apart. Fortunately one of the markers analysed was located within the interval of shared homozygosity, allowing the pattern of linkage disequilibrium between cases and controls to be observed.

Genome-wide association study approaches using a high-density SNP array were used to investigate three of the conditions presented in this thesis. A huge advantage of using whole genome SNP genotyping arrays is the level of throughput that can be achieved. The Illumina CanineHD SNP assays 173,000 genome-wide SNPs equating to approximately one every 14 kb. Although it is impossible to avoid outsourcing of the genotyping work when in-house facilities are not available, results can be produced for small cohort studies within a few weeks of sample submission to a service provider. Linux based data analysis is rapid, enabling basic allelic association analysis results to be obtained within a few hours. The density of the CanineHD SNP array means that the fine mapping required after microsatellite-based approaches can be omitted completely. Raw genotyping data can be visualised to clearly identify the position of recombination events marking the disease-associated interval and also gives an indication of potential outliers.

Although use of high-density SNP arrays holds many advantages for carrying out disease mapping studies, there are some potential obstacles to consider. Because of the high number of markers tested it is more likely that false positive associations could be detected by chance. This can be relatively easily corrected for by performing permutations analysis or carrying out Bonferroni correction on allelic association analysis results. A greater problem is the risk of producing false positive results due to high levels of genomic inflation, which essentially means that ranked observed probability values are greater than expected values. Genomic inflation is caused by population stratification, where in the most extreme situation, cases and controls cluster in separate groups of closely related individuals. Population stratification is particularly problematic when performing disease mapping studies in the purebred dog. Because breed groups are genetically isolated, with only a small proportion of individuals within a breed used as breeding stock, it is not unusual for the majority of cases to originate from particular breeding lines. Case DNA samples received for studies often come from the pet dog population however, so obtaining healthy related individuals as controls is not always possible, especially as individual details have to be kept confidential. Although a high level of genomic inflation was seen when parallel mapping two disorders in the Cavalier King Charles Spaniel, the results did not appear to be greatly affected. This was mainly due to the strength of statistical signal obtained when mapping these autosomal recessive disorders and also because of the ability to apply analysis tools to correct for the effects of stratification in datasets. The opposite problem was seen when mapping the locus for late onset ataxia in the Parson Russell Terrier, where genomic deflation was seen. Again the overall conclusions were not adversely affected because of the strength of the signal. Population stratification is a much more significant problem to consider when mapping complex inherited diseases. High levels of genomic inflation could result in a large number of false positive results being produced. Although corrective tools are available they are modelled for use in human association analysis studies where the high levels of genomic inflation seen in dog studies are not evident. Very careful selection of cases and controls is arguably the best method of controlling the problem, such as selecting first degree relatives as controls. However for complex disorders many asymptomatic relatives may share risk factors for disease, and therefore use of first degree relatives may reduce the strength of true positive signals. Although extremely rapid, the use of whole genome SNP arrays is an expensive approach usually costing around £120 per sample. In addition a large lump sum is usually required to fund a project as some service providers only offer to process a minimum of 48 samples in one batch. There are some situations where the use of microsatellite markers may be more appropriate. Genotyping using individual microsatellite markers is inexpensive and data can be analysed relatively rapidly.

Although individuals SNP markers can be genotyped rapidly using an allelic discrimination approach (TaqMan), probes are relatively expensive. Therefore microsatellite markers are more appropriate for use in a candidate gene study approach when only a few individuals need to be genotyped for a small number of genetic markers. Microsatellite markers also tend to be more informative, due to a greater number of potential alleles.

7.3. Candidate gene study approaches

In this thesis two very different candidate gene studies are described, one of which successfully led to the identification of a disease-associated mutation and the other which was unsuccessful. The unsuccessful attempt used microsatellite markers to assess candidate genes for linkage to EF in the CKCS. Although there were many strong candidate genes and well defined cases were available, none were found to be linked to EF. Subsequently a novel disease-associated locus was identified using a GWAS approach and a mutation identified in a gene not previously associated with a muscle hypertonicity condition. This unsuccessful attempt highlights the risks of using a candidate gene approach in situations where the causal gene has not previously been associated with a similar condition in other species, or is not a functional candidate. Candidate gene studies are still however a cost-effective method of excluding many strong candidates, before more expensive GWAS approaches are undertaken, especially for conditions with strong/early onset phenotypes.

In the second candidate gene study a novel approach of genome-wide mRNA-seq was used. Although the approach is unconventional, it is a highly cost effective method of enrichment. Unlike for exome enrichment of genomic DNA, a proprietary kit is not required as mRNA can be isolated using a simple method of polyA enrichment. For partially degraded total RNA samples, ribosomal RNA removal kits are also available for enriching mRNA levels. Unlike exome sequencing using probe-based genome enrichment which relies heavily on the extent and accuracy of genome annotation, with mRNA all coding regions expressed in a particular tissue are enriched including all exons and splicing isoforms. There is also selective enrichment towards genes that are more highly expressed, so only genes which are important in the function of the tissue under investigation are analysed. Using mRNA-seq allowed the candidate gene study to commence with just a single case. This is especially significant for rare or emerging disorders where the collection of additional cases may be time consuming or even impossible. If the study had not led to the direct identification of the causal mutation the dataset may still have become useful at a later date if a disease-associated locus had been identified using a GWAS approach. The sequencing data across the associated

interval could then have been analysed for disease-associated mutations without the need for an additional target enrichment sequencing approach. Alternatively all the other genes sequenced by the mRNA-seq approach could have been analysed for non-synonymous mutations, rather than just the initial candidates, and any provocative mutations identified investigated in further cases and controls.

7.4. Use of massively parallel sequencing techniques

The use of massively parallel sequencing techniques forms a large part of the research work contributing to this thesis and clearly demonstrates the development of next generation sequencing technology over the last five years. Sequencing projects completed as part of this thesis are summarised in Table 7.1.

Table 7.1 Summary of massively parallel sequencing projects

The massively parallel sequencing projects completed as a part of this PhD thesis. 1 x represents single-end sequencing and 2 x represents paired-end sequencing.

Project	Date	Platform	Read length (bp)	Dataset size (Gb)
ITPR1 sequencing	Mar 2008	Illumina GAII	1 x 33	0.05
CKCS projects	Sep 2010	Illumina GAIIx	2 x 51	3.47
LOA/SCA projects	Sep 2011	Illumina HiSeq	2 x 51	19.64
mRNA-seq	Mar 2012	Illumina HiSeq	2 x 51	17.38
LOA additional controls	Jun 2012	Illumina MiSeq	2 x 150	2.07

7.4.1. Target enriched massively parallel sequencing

When *ITPR1* was sequenced, in March 2008, massively parallel sequencing technology was still in its infancy. Reads were short and although paired-end sequencing had been developed, it was not routinely available through the limited number of service providers. As a new and unfamiliar user of the technology the PCR product template was sent to the service provider (Fasteris) for library preparation and sequencing. Analysis of the data was performed in-house with support from Fasteris, although the amount of freely available software was limited at that time. The software package, Maq, was chosen based on recommendation, which could be implemented relatively easily through an “easyrun” Perl script. SNP calling was accurate, but calling of indels was not reliable and contained many false positives. Although a data visualisation tool called Maqview was available (Figure 7.1), the text like display made visualisation of sequence variants difficult. The viewer could therefore only be used as a means of variant verification after automated SNP and indel calls had been made.

Figure 7.1 The Maqview sequence alignment visualisation tool

The Maqview sequence alignment viewer: a Linux based visualisation tool with a text based display output.

Limitations in massively parallel sequencing and data analysis were the reason why the GAA repeat expansion in the *ITPR1* gene was not identified in the first sequencing experiment. The consensus sequence output from Maq was identical to the reference sequence over the expanded region. Although a drop in read depth was observed across the expanded region, some of the short reads had been aligned into the repeated region and because the dataset consisted of single-end reads, no singleton reads were present which would have warranted further investigations.

In-house library preparation needed to be fully investigated for the second attempt at massively parallel sequencing for the CKCS projects. The SureSelect solution based target enrichment system had been launched by Agilent technologies, which enabled in-house capture in the standard laboratory for the first time. In the early edition kit, however, no reagents or guidelines were provided for library preparation pre-capture. Illumina kits were extremely expensive at the time and the required primer and adapter sequences had to be purchased separately making library preparation potentially costly. It was therefore decided to use the NEBnext library preparation kit with library amplification and adapter oligos being produced by an oligo manufacturer, enabling the project to go ahead without the need to raise additional funds. A number of fragmentation methods were investigated, and an enzymatic method was ultimately chosen because, unlike nebulisation (the method initially recommended by Illumina), samples could be processed in parallel. Sonication was also considered but excluded as an option because of concerns it may cause too many damaging nicks to the DNA backbone making library preparation inefficient. As recommended by Agilent at the time an agarose base size selection stage was also included.

Developments to the Agilent kit resulted in changes to the methodology for the second target enriched library preparation experiment. In the new 'XT' version of the kit all reagents for library preparation were included, meaning that no supplementary kits or custom made adapters were required. The newly developed Covaris DNA shearing apparatus was also recommended for DNA fragmentation omitting the need for a size selection stage. This was a convenient development because the only available DNA samples from the PRTs selected for sequencing were from buccal swab samples. Although DNA from buccal swabs is suitable for most molecular applications, it is highly fragmented and contains fragment sizes evenly distributed from 20 kb+ to 100 bp. Enzymatic methods would have been unsuitable as too much material would have been lost during fragmentation, and libraries would require additional rounds of PCR amplification leading to more PCR duplicates in the final sequencing dataset. High molecular weight DNA extracted from whole blood was however available from the ISs selected for target enrichment. Enzymatic fragmentation was therefore attempted for that investigation and a fragment size range was achieved that was highly comparable to Covaris shearing. Buccal swab DNA was therefore outsourced for Covaris shearing. Very similar size ranges could be achieved with both methods demonstrating how methodology can be selected to best suit the available DNA resources, with the same end result.

Unlike for previous target enriched sequencing attempts which required some outsourcing for library preparation and massively parallel sequencing, the workflow for the third target enriched massively parallel sequencing experiment, which was part of the LOA in the PRT project, was completed entirely in-house. Although the DNA to be used in the experiment originated from buccal swabs, outsourcing of fragmentation was avoided by reinvestigating the use of sonication. It was found that libraries could be successfully created using sonicated DNA and although the fragment size range was slightly greater than that achieved by Covaris shearing, the size range was suitable for target enrichment. Sequencing was achieved using a newly acquired Illumina MiSeq.

7.4.2. Development of a sequence analysis pipeline

The amount of data produced in the later massively parallel sequencing projects made the development of a custom sequencing analysis pipeline essential, not only for the projects described in this thesis but also for future projects to be undertaken at the AHT. Although Maq, the program initially used in data analysis, was still functional, many other alignment, sorting and handling programs had been developed which were faster and more accurate. Also as targets increased from kilobase through to megabase regions there was a need for a greater level of automation for analysing data and annotating variant calls. Most

freely available software tools for massively parallel sequence data analysis are Linux based and require implementation at the command prompt by entering the program address and listing a number of parameters. Furthermore, the programs tend to work in a modular fashion making implementation of each program by entering commands laborious and time consuming as the run time for each module is not easy to predict. Therefore to make analysis of the sequencing data easier to implement and more accessible to other users a sequence data analysis pipeline was developed. The pipeline was written as a Perl script which could be implemented with a simple one word alias at the command prompt (NGS ENTER), meaning that no expert knowledge of Linux was required to use it. Users were then prompted to enter the parameters for analysis such as file names and analysis details. The pipeline was flexible, allowing users to choose which modules were included in the analysis, from a simple alignment to full variant analysis and annotation. Error checking was incorporated into the Perl script, so if an error message was reported by the program the script would terminate and notify the user by email. Users were also notified by email when the analysis had successfully completed. Notification of errors and pipeline completion, meant users could allow the scripts to run in the background without the need for constant checking. On the completion of a run results files were placed in a new folder with readme.txt files explaining the results.

7.4.2.1. Pipeline problems and redevelopment

A number of problems were encountered during the development of the Perl script. One of the largest problems was the server resource available at the AHT. Files produced by massively parallel sequencing are large files of several gigabytes in size. Files produced by Illumina GAII machines in the first two sequencing experiments could be handled without problems. Data from the initial experiment were analysed using the Maq easyrun script which completed within a few hours, although the reference sequence used was just the 340 kb reference sequence rather than the entire genome reference. Use of a partial genome reference sequence is not recommended however, because it may result in the best alignments to the genome not being found, and potentially weaker alignments being accepted. The second sequencing experiment was a significantly larger dataset consisting of five paired files of probe-based target-enriched sequencing data. At this stage Maq analysis had been superseded by an early version of the NGS sequence analysis pipeline. Separately initiated pipelines had to be set up for each file pair, which could be performed in parallel using the split screen feature of the Linux command prompt. Although this approach makes the analysis more convenient, one disadvantage is that memory resources are split, which was particularly problematic with the resources available at the AHT. It was found that the five file pairs could be analysed in parallel if aligning to a partial genome reference, such as a chromosome or a targeted region, but

there was insufficient capacity to align all five paired datasets against the entire genome using the parallel approach. Individual analysis of a single file pair against a whole genome reference sequence took approximately 24 hours.

A more significant challenge for in-house data analysis came from sequencing data derived from the Illumina HiSeq 2000, which generated approximately five fold more data. The data analysis pipeline had been enhanced by including a variant effect predictor module and a structural variant analysis module, improving the output but increasing the run time of the script significantly. Analysis of a single file pair took approximately three days when aligning the data to the canine genome sequence. The server was not solely used for sequence data analysis, and if other users needed to perform tasks (such as GWAS analyses) the system would be extremely slow or would crash completely. For the first two sequencing experiments data were provided as raw unaligned FASTQ files only. However, in later sequencing runs data were provided in both unaligned FASTQ and aligned bam file format. To reduce running time the pipeline was changed to allow handling of the aligned .bam files as the input files. In addition the per-base depth of coverage module was disabled as it provided little useful information and contributed considerably to the run time, allowing a single pair of files to be analysed in approximately one day. This was a significant improvement, but meant that all ten file pairs still took ten days to analyse.

Attempts were made to develop a separate data analysis pipeline for analysis of mRNA-seq datasets, but unfortunately the widely used programs Tophat and Cufflinks required more memory than was available on the AHT Linux server (Trapnell et al., 2012). To use the mRNA-seq data to investigate candidate genes for canine ataxia the raw alignments to the canine genome therefore had to be used. This is suboptimal because exon boundaries are not considered for standard alignments to the genome and therefore many misalignments were apparent across the exon-intron boundaries. Because of the number of misalignments, data had to be analysed manually as computational methods of calling variants would have suggested a large number of splice site mutations. Analysis of gene expression by counting the reads aligned against the genes in the canine genome also had to be outsourced to the Wellcome Trust Centre for Human Genetics, Oxford.

The original NGS analysis pipeline was developed in the summer of 2010. Since then new analysis programs have been written by external institutes and old modules have been updated leaving the pipeline outdated. The ever increasing amounts of data produced by massively parallel sequencing techniques meant that a faster and more streamlined

pipeline was required, especially if large-scale experiments such as whole genome sequencing were to be considered. The purchase of a new 64-bit server by the AHT provided the opportunity to redevelop the pipeline including the most up-to-date modules. This redevelopment is currently being carried out by Dr Mike Boursnell based on the original Perl script. Most modules function in the same way as the original script with some notable improvements. The updated script allows a list of data files to be specified at the start of the script, so file pairs can be analysed consecutively without the need to initiate separate analysis runs in split-screens. Because of the vastly increased speed of the new server, consecutive rather than parallel analysis does not significantly increase the run time. Variant calls are made by considering a set of alignment .bam files rather than being called independently. This method increases the dataset and improves the accuracy of variant calls. Additionally a single more manageable variant calls file is produced containing a table of variant calls across all the specified samples. Thirdly variant annotation occurs after a separate variant handling stage in Excel, vastly increasing processing speed. The changes highlight the constant need to keep the pipeline up to date as sequencing technology evolves. As sequencing technology further develops, even larger datasets with long read lengths are likely. New sequencing applications may become available and it is therefore crucial to modernise procedures accordingly to enable appropriate handling of data.

7.4.3. Comparison of target enrichment approaches

Two target enrichment methods were employed during the investigations. These were long range PCR and probe-based target enrichment. There are advantages and disadvantages to each technique, which makes each of them more suitable for particular applications. Long range PCR is a very standard technique that can be performed in most basic laboratories, without the need to purchase a proprietary kit. The major advantage of a PCR based enrichment is that, barring regions of extreme GC richness (>80%), a complete template of contiguous PCR can be created for the target region of interest. This results in the ability to achieve near 100% sequence coverage of target regions (see Table 7.2).

Table 7.2 Summary of the four target enriched sequencing experiments

Enrichment attempt	Method	Platform	Dataset size (Gb)	Average target enrichment efficiency (%)	Bases achieving 10x coverage (%)	PCR duplicates (%)
1	LR PCR	GAII	0.05	88.3	98.9	Unknown
2	Probe capture	GAIIx	3.47	85.3	79.0	13.6
3	Probe capture	HiSeq2000	19.64	77.4	79.0	37.2
4	Probe capture	MiSeq	2.07	66.1	56.8	2.2

Long range PCR is a rapid technique. Primers can be designed, ordered and received within a couple of days, and PCRs can be set up and analysed in a day. Potentially, with an in-house sequencing facility, experiments could be designed, carried out and analysed within a week. Long range PCR is not a completely robust technique however. A pure, high quality source of genomic DNA is required, and amplification success is roughly inversely proportional to amplicon length. As a consequence there is a compromise between amplification success rate and the number of primer pairs/amplicons that need to be designed. Using short amplicons (eg 2-3 kb) may result in a high frequency of successful PCRs, but is an expensive strategy requiring a large number of primer pairs to be ordered. A strategy using long amplicons (10 to 20 kb) is initially inexpensive, but is likely to result in a higher failure rate, requiring primer redesign. Having to perform several rounds of primer redesign is both time consuming and expensive. The strategy of using longer amplicons was adopted for sequencing of the 340 kb *ITPR1* target region. This strategy was largely successful, with eventual contiguous coverage of the 340 kb target region requiring 60 PCRs with an average amplicon length of 6.3 kb. However, two particularly difficult regions were identified which required several rounds of primer design and PCR before successful amplification could be achieved. If a PCR failed a single pair of new primers was designed to cover the same region. If amplification failed again then the target region would be split into two, primers designed and possible combinations of the available primers used for PCR. For the particular regions in question, a fully contiguous sequence of overlapping amplicon was eventually achieved, with the smallest amplicon being 121 bp. Several rounds of primer design were required to get to this stage, which added several weeks to the study period. To reduce study timeframe these difficult regions could have been excluded from investigation after one or two rounds of primer design, if both case and control group PCRs failed to produce product bands. However, this would lead to an increased risk of the causal mutation not being identified. Amplification of regions larger than 340 kb would be time consuming and challenging, although once primers have been purchased and successful amplification has been achieved, large numbers of individuals could be processed at a relatively low cost. For this reason long range PCR is more suitable for processing multiple samples for small target regions.

The advent of probe-based enrichment systems enabled target regions of several megabases to be investigated for the first time and the launch of a solution based system made this technology accessible for use in the standard laboratory environment. Although a proprietary kit is required, the protocol is streamlined for processing of up to 12 samples in parallel. Cost per sample is high, and unlike for long range PCR, if processing of more

than 10 samples is required additional kits (or a larger kit size) must be purchased. The experiments in this thesis have shown that although probe-based target enrichment is highly effective, experimental design is hindered by a high frequency of repeat sequences in the canine genome. Because probes designed to non-unique sequences in the genome would lead to a very high level of non-specific capture, and as a result lower coverage of target regions, repeat masking is applied during probe design. For this reason base coverage by probes was only 56.7 and 64.8% for the two target enrichment experiments described. Ultra-deep sequencing using the GAIIx and HiSeq 2000 resulted in >10x coverage for 79% of target regions however, due to capture of regions flanking probes sequences (read depth around a probe is approximately normally distributed). The smaller dataset produced by the MiSeq however resulted in a smaller percentage of bases achieving 10x coverage (56.6%). The level of PCR duplicates observed for probe-based capture is proportional to the number of reads, and inversely proportional to the target capture size. It is also influenced by library preparation method. Including a gel size selection stage reduces the number of unique sequences and is therefore likely to increase the number of PCR duplicates. The number of PCR duplicates in a dataset is important because they are removed during analysis to avoid variant calling bias. Sequencing of probe-based capture libraries on the HiSeq produced the highest level of PCR duplicates at 37%. This suggests that the amount of useful data is approaching a limit for probe-based capture experiments and that production of additional reads may not significantly improve the quality of the dataset. As massively parallel sequencing technology continues to advance it will therefore be essential to carefully consider platform choice and potentially adopt a greater level of multiplexing to maximise sequencing experiment efficiency.

Identification of an intronic causal mutation within a repetitive region of the dog genome highlights the potential limitations of using probe-based enrichment techniques which result in incomplete capture of target regions. Although the intronic GAA repeat expansion was identified using target enrichment of the disease-associated interval, exact bait positioning facilitated mutation identification. Because the mutation locus and surrounding region contained repetitive elements, a 500 bp region containing the repeat expansion was excluded from capture-probe (bait) design. In controls sufficient flanking sequence was captured by the baits to result in sequence coverage all the way through the repetitive region. For cases however, because the repeat expansion increased the size of the masked repetitive region, no DNA fragments were captured that were of sufficient size to span across the GAA repeat. This resulted in a region of zero coverage in cases 3' of the GAA repeat sequence, as no probes 3' of the GAA repeat were located close enough to

capture this region. The gap in coverage in cases, but not controls provided a discernable difference that could be followed up, allowing mutation identification. The variation may also have been difficult to identify using a whole-genome sequencing approach as complete coverage would be achieved for both cases and controls. Singleton reads would be highlighted in the region, but potentially also present in control individuals with long normal GAA repeat alleles.

7.4.4. Genome-wide mRNA-seq

In this thesis the versatility of the mRNA-seq method was explored by using the approach to fulfil a variety of functions. Although mRNA-seq is most commonly used as a method of high-throughput gene expression level analysis or to facilitate genome annotation, it was used here as an alternative method of targeted exome enrichment. Although there is now a kit available for targeted canine exome enrichment, use of the mRNA-seq does hold some advantages. Firstly mRNA-seq enrichment has an extremely simple workflow. mRNA enrichment from total RNA can be achieved by using a polyA capture using bound oligo(dt) sequences. Oligo(dt) probes are not a new technology and products are widely available either in column or bead format, making them inexpensive in comparison to probe-based exome enrichment. Efficiency of mRNA enrichment from total RNA using oligo (dt) probes is near 100%, in comparison to exome enrichment methodology where closer to 70% target enrichment is expected, meaning more on-target sequence reads and therefore higher read depth when using mRNA-seq. Furthermore mRNA-seq is a precise method of exome capture. Unlike exome enrichment, no flanking intronic sequences are captured, which may equate to a large proportion of the sequencing dataset, although it is dependent on target fragment size. For disease studies mRNA-seq selectively enriches the most highly expressed genes for a particular tissue. High levels of expression for particular genes may indicate their importance in the function of a particular tissue, and therefore mutations in these genes may be more likely to disrupt function. The combination of high enrichment efficiency, precise exon capture and greater enrichment of highly expressed genes, enabled a disease-causing mutation to be identified using a 1.39 Gb dataset. Even if the causal mutation had not been identified with the candidate gene approach used, the mRNA-seq dataset could have been revisited subsequent to a GWAS study once a disease-associated genomic region had been identified. As with other methods there are risks associated with the use of mRNA-seq as a method of candidate gene sequencing, the obvious risk being that if the mutation is not in a candidate gene then the causal mutation will not be identified. Although ataxia is well studied in humans with many causal genes identified, there are some remaining disease loci for which neither the causal gene or mutation have been identified. Mutations within these loci

would be far more difficult to pinpoint, and also to prove as causal, although potentially more scientifically significant than identification of causal mutations in previously associated genes. Although mutations are most likely to be in the highly expressed genes for a particular tissue under investigation, there were three ataxia associated genes for which insufficient gene coverage was achieved to allow the entire sequence to be determined. Although this was a small proportion of the candidate genes investigated it still shows that genes expressed at a lower level can play a pivotal role in the function of a tissue if disrupted. RNA is a notoriously difficult material to work with, and both storage and extraction method are critical to experiment success. Although brain tissue contains a relatively high lipid content, it is a good tissue to work with for the extraction of RNA. The RNA content of the tissue is high, the RNase levels are low and the tissue is soft making disruption and homogenisation prior to RNA extraction straightforward. Tissues such as bone, which is extremely hard and contains little RNA, or pancreas, which contains an extremely high level of RNases, would be more challenging tissues to work with. Kits are available to perform mRNA-seq experiments from low quantities of fragmented RNA, and even using RNA from FFPE tissue using rRNA removal techniques rather than polyA enrichment. The advantage and risks of using mRNA-seq as a method of candidate gene sequencing imply that these things must be considered before undertaking an investigation, as the approach is not suitable for all situations.

Disease characterisation is a very important factor which is likely to influence the success of the experiment. The NCCD Beagle case was well characterised before the start of the investigation. The phenotype was constant, severe and early onset which enabled the diagnosis to be easily made. Histopathological examinations were undertaken which clearly indicated that degeneration of the cerebellum was leading to manifestation of clinical signs and therefore it was highly likely that genes expressed in the cerebellum tissue played an important role in pathogenesis. Pedigree information was available for the known cases, suggesting an autosomal recessive mode of inheritance for the disease. Affected individuals were present in two separate litters produced from the same mating, indicating that the condition was likely to be inherited. All the evidence suggested a single major effect gene mutation was the cause, which was highly likely to be expressed in the cerebellum. A tissue resource was also available as the affected dog investigated was euthanised, and many suitable candidate genes were available. In summary mRNA-seq is a suitable method of candidate gene sequencing, but the chances of success of the technique are very dependent on type and characterisation of the condition under investigation, how well similar conditions have been characterised in human studies and

the availability of a suitable tissue resource which has been shown to be involved in disease pathogenesis.

For the study of LOA in the PRT the mRNA-seq data were used for multiple purposes. Although a target enrichment approach had been used to amplify the LOA disease-associated region, the mRNA-seq dataset from the LOA cases cerebellum provided an excellent method of checking gene annotation. The data could also be used to double check sequence data across coding regions of the LOA disease-associated interval, which had not been sequenced to a high level of coverage or had been missed due to positioning near to a repetitive element or high-GC content of the sequence. The effects of variants potentially affecting splicing could also be investigated. Cerebellar RNA from a control sample was also sequenced, allowing expression levels to be investigated. Sequence read alignments could be assessed visually by browsing in IGV, but regrettably computational analysis of comparative gene expression had to be outsourced to the Wellcome Trust Centre for Human Genetics, Oxford, because the AHT Linux server did not have sufficient memory to allow the analysis programs to run. Although a gene expression analysis project would not normally be carried out using just one case and one control, the dataset proved highly informative. No significant differences in gene expression were observed across the disease-associated region for the PRT study and there appeared to be quite few major differences in gene expression across the genome. This may indicate that a relatively small collection of tissues samples would be required to perform a full genome-wide study of gene expression, depending on the condition under investigation and the tissue type.

Unlike targeted resequencing approaches which generate data that are likely only to be used for a specific project, genome-wide mRNA-seq data could potentially be used for projects in the future. The dataset has already been used on several occasions to check levels of gene expression and accuracy of genome annotation, even for projects where the site of pathogenesis is not the cerebellum. Individuals used in mRNA-seq could be used as additional controls in the future, and if a larger dataset is required it would still be possible to return to the original libraries to perform additional sequencing runs and acquire more data. To allow in-house data analysis and to avoid outsourcing, development of an mRNA-seq data analysis pipeline would be a future priority.

7.4.5. Limitations of massively parallel sequencing technology

Massively parallel sequencing techniques have revolutionised the way in which regions of the genome can be investigated in disease studies; and genome sequencing projects that

once would have taken years can now be completed in a few days. The technology is not without its limitations however. The limitations associated with target enrichment techniques and mRNA-seq have already been discussed. One particular difficulty that was encountered for all the projects was the ability to identify SINEs within sequencing datasets. Repetitive elements, in particular SINEs are especially abundant in the dog genome (Wang and Kirkness, 2005). Evidence for this is provided by only being able to achieve 56% to 65% bait coverage for target enrichment projects due to repeat masking, leading to reduced region coverage. During the development of the sequence alignment pipeline a program for identification of SINEs could not be identified. This meant that the only method of SINE identification was to view the sequence read alignments manually. As SINEs are flanked by repeat sequences, some elements could be identified because of a doubling of read depth which could be seen in the coverage histogram in IGV, but this was not consistent between SINEs. Therefore to ensure all SINEs were identified sequence read alignments had to be scanned. This is a highly laborious process as alignments around SINEs often look spurious making them difficult to distinguish from poor sequence read alignments. This was achievable for small target regions of a few megabases, but it would be difficult to manually scan larger amounts of data such as whole exome capture or whole genome sequencing.

Another limitation is cost. Some small scale sequencing projects that would require more sequencing data than could be quickly and cost effectively achieved by Sanger sequencing, would potentially be sequenced to a extremely high read depth using a massively parallel approach. Using a high level of multiplexing is potentially a solution to this, although may increase study timeframes as additional projects and libraries may need to be completed before the sequencing run can commence.

Data analysis is a particular limitation for smaller institutions such as the AHT. Benchtop sequencing platforms such as the MiSeq have made massively parallel sequencing extremely accessible. Firstly the increasing amount of data generated by massively parallel sequencing means that significant disk space is required just for storage purposes, although the advent of cloud storage technologies may be more widely used in the future. A huge number of free software packages are available for data analysis, but software choice is largely made on recommendation rather than suitability. Some of the best supported packages are under constant redevelopment making it challenging to keep analysis pipelines up to date.

As massively parallel sequencing technology and handling software continues to develop many of the difficulties that are currently experienced may become less problematic in the future, although the ever increasing dataset sizes may still present a challenge. New technology holds the promise of even greater read lengths and data quality in the future.

7.5. Developing massively parallel sequencing technologies

Although current platforms such as the Illumina MiSeq and Life Technologies Ion Proton are under constant development, much of the focus regarding new technologies is on single molecule sequencing methodologies. One strand-sequencing platform that is currently available is the Pacific Bioscience RS system. The Pacific Bioscience system works through anchored single DNA polymerase molecules at the bottom of holes of just tens of nanometres in diameter. Upon binding of a single stranded DNA molecule, fluorescent nucleotides are detected as they are held by the polymerase before incorporation into the elongating strand (Eid et al., 2009). The fluorophores are attached to the pyrophosphate group which is naturally cleaved as part of the elongation process allowing detection to occur in real-time by capturing a video rather than still images. As the nanopore is smaller than the wavelength of visible light, the light generated by the laser cannot pass all the way through the pore, resulting in low background fluorescence of unincorporated nucleotides and a high signal to noise ratio. The major advantage of strand sequencing is that much longer read lengths can be achieved with single molecule sequencing technology. The Pacific Bioscience RS system is capable of producing read lengths in excess of 10,000 bp with an average read length of 3,000 bp. Long read lengths can facilitate *de novo* assembly and alignment of reads to a genome. Assembly across repetitive regions of the genome are less challenging because reads flank into unique regions, helping genome assembly which may be especially difficult for some small microbial genomes containing many mobile repetitive elements. Large insertions and structural variants may also be much easier to identify and interpret with long read technology. Also no PCR amplification is required in the sequencing process minimising bias. Although useful for some applications, the Pacific Bioscience RS system is not currently a popular platform, partly due the large footprint of the machine, but mainly due to the low read accuracy in comparison to other next generation sequencing technologies. Much of the hope for improved single DNA molecule sequencing is now being pinned on products currently being developed by Oxford Nanopore. The technology being developed by Oxford Nanopore uses protein nanopores as biological sensors. Two systems are currently under development: an exonuclease based sequencing system which is being developed in strategic alliance with Illumina (Clarke et al., 2009) and a strand sequencing based system (Lieberman et al., 2010). Both systems work by flowing a continuous

current through a nanopore and measuring the characteristic change in current that occurs as molecules pass through the pore. In the case of a single stranded DNA molecule the patterns of current disruption created can be used to determine all four nucleotide bases as they pass through the nanopore. In exonuclease sequencing an exonuclease is bound near the entrance of the nanopore. Single stranded DNA molecules are fed into the exonuclease and nucleotides cleaved one by one. Nucleotides then flow through the nanopore causing a disruption of current that can be measured. In the strand sequencing method a DNA polymerase is bound to the nanopore which unzips double stranded DNA and feeds a single strand into the nanopore. The disruption in current as the strand passes through the nanopore can be used to determine the nucleotide sequence. Both strands of the DNA can be sequenced, by ligating a hairpin loop on the end of the double stranded DNA. Once the first unzipped strand has been sequenced the hairpin is fed through the nanopore and sequencing continues for the reverse strand. Two products are currently under development, the GridION system and the MinION, which are based on strand sequencing technology. The GridION system is an expandable sequencing system consisting of “nodes” which can be used singularly or in racked multi-node systems depending on data requirements. DNA samples prepared for sequencing are loaded onto cassettes that contain all the required reagents, which are inserted into the nodes of the system for sequencing. The MinION prototype is a miniaturised sequencing system which is approximately the size of a smart phone. The MinION is a stand alone single use sequencing platform which only requires USB connection to a laptop computer for operation. MinION units are projected to cost less than $1000, so have huge potential for use in a clinical setting. Nanopore technology is potentially highly versatile and nanopores could be adapted for analysis for several substrates including DNA, RNA, miRNA and proteins. The versatility and affordability of developing nanopore systems, along with high accuracy levels and minimal sample preparation would make the technology a very exciting proposition; although currently there are no sample datasets available and no products have been launched so it is not known how long it will take for this promise to be realised. Oxford Nanopore is not the only company aiming to make sequencing using nanopore technology a reality. Genia have developed a nanopore technology which utilises principles of both Oxford Nanopore and Pacific Bioscience to form a novel sequencing method. The Genia system uses a DNA polymerase attached to a membrane bound nanopore, which has a constant current flowing through it. The substrate for the DNA polymerase is nucleotides which are molecularly tagged on the pyrophosphate group. When the polymerase binds a DNA molecule and nucleotides are sequentially incorporated the tagged pyrophosphate groups are cleaved and pass through the nanopore. The unique current disruption signatures created by each tag as it passes

through the nanopore can be used to determine the nucleotide sequence of the template (Kumar et al., 2012). If this technology comes to fruition, Genia claim that they are aiming toward a $100 genome.

7.5.1. Applications of new sequencing techniques in the mapping of dog diseases

With the growth of massively parallel sequencing technologies showing no signs of slowing down, and price per base sequencing cost becoming increasingly less expensive, there are a number of new applications that may become applicable for use in the study of genetic disease in the dog. The use of genome-wide mRNA-seq has been used as a method to identify causal mutations in genetic diseases, and assuming that a more automated pipeline of variant analysis can be developed it is highly likely that more genome-wide sequencing studies will be undertaken in the future, potentially leading to the investigation of more individual cases in a “personal genomics” approach. Applying a genome sequencing approach to answer clinical questions could be possible with a high level of automation and if costs become sufficiently low. Cancer panels are available for targeted enrichment of genes known to cause cancer in humans so it is feasible that as our knowledge of genomics and disease genetics increases we may in future be able to answer clinical questions and perhaps one day tailor therapeutics based on genetic information. A vast number of disease-associated mutations have been identified in the dog so it may be possible that a sequencing assay will be designed in the future to test for all known mutations if there is a suitable demand for such a service.

Although the tools for GWAS such as genome-wide SNP arrays have not advanced at the same rate as for massively parallel sequencing, there is potential for advances in this field through the use of massively parallel sequencing approaches. A genotyping-by-sequencing (GBS) approach using restriction enzyme digestion of genomic DNA has been developed for use in maize and barley GWAS studies (Elshire et al., 2011). This methodology may become more applicable as sequencing costs are further reduced. Another potential methodology for use in GWAS is genotyping-by-exome sequencing. Although use of this method would rely on a reduction in cost of probe-based capture and would be a bioinformatics challenge, the potential advantage would be that if a significant association was found then sequencing data would be immediately available for investigation of genes in the disease-associated interval. Use of genotyping-by-genome sequencing may also become a possibility if the $100 genome becomes a reality, although would require a finely tuned data analysis pipeline to reduce complexity, make genotyping calls and perform association analysis.

Using a GBS approach there may soon be the opportunity to move away from the use of microsatellite markers in DNA profiling (fingerprinting) systems and move to a higher throughput and automatable approach using SNPs. Microsatellites are still routinely used in both human and canine genetics, largely because of large historical databases of profiles that have built up over the years making it difficult to move to a new system, especially for forensic science purposes. It has been suggested recently however that it may be possible to assay for both microsatellite sequences and SNPs in a parallel approach using next generation sequencing technology (Bornman et al., 2012). Linking this back to the dog it may one day be possible to test for multiple genetic diseases and produce a DNA profile in a single experiment at very little cost.

7.6. Study limitations and future work

Although five novel mutations have been associated with five genetic diseases in the dog, no functional work has been performed to prove that mutations are truly causal, and to potentially investigate the reasons why the mutations have a pathogenic effect. This is especially significant for mutations in *BCAN*, *FAM83H* and *CAPN1* which have not previously been associated with similar human conditions. Although knock-out mice have been created for *BCAN* and *CAPN1*, they were reported to be clinically normal, so perhaps it would be worth revisiting these models to reassess the potential clinical signs. No knock-out mouse exists for the *FAM83H* gene so future development of a murine resource may help improve understanding of the clinical signs seen in the dog and also amelogenesis inperfecta in human patients. Although the *CAPN1* mutation is predicted to result in replacement of the highly conserved catalytic cysteine residue, loss of enzymatic activity would only be confirmed by using a specific assay. This may involve synthesis of a mutant protein and development of a custom assay. In the study of SCA in the IS the effects of the GAA repeat expansion on *ITPR1* expression levels could not be investigated due to lack of a suitable tissue resource.

7.7. Concluding remarks

In this thesis the dog model has been successfully used to demonstrate the advances in disease mapping and sequencing techniques over the last five years. The unique population structure of the purebred dog has enabled large case-control cohorts for disease studies to be collected over short timeframes due to the high frequencies of particular diseases within certain breed populations. The efficiency of the dog model has been shown using GWAS approaches with a minimal sample cohort or in parallel with other diseases, demonstrating the advancement from microsatellite approaches through to genome-wide SNP array methodology. The long disease-associated intervals seen in

GWAS results demonstrate the high levels of LD seen in the dog, facilitating identification of disease loci. Investigation of genes in disease-associated intervals has shown the transition from Sanger sequencing through to massively parallel approaches, and the potential advantages and limitation have been described. A novel approach of candidate gene sequencing using genome-wide mRNA-seq has been established, which shows how individual clinical cases could be investigated using massively parallel sequencing in the future, and raises the prospect of an era of personal genomics as sequencing technology continues to advance. In this thesis five new disease-associated mutations have been identified, included three mutations in genes that have not been associated with similar diseases in humans or studies of other model organisms, showing how the dog model can be used to improve our understanding of genetics and provide new candidate genes for disease studies in humans.

Advances in massively parallel sequencing show no signs of slowing down with the development of new technologies. Extremely long read length with increased accuracy, higher levels of throughput and the potential to reduce cost are now on the horizon. It may soon be feasible to perform a GWAS by sequencing and with a streamlined bioinformatics pipeline go from sample cohort to identification of disease-associated mutation in a single process. A personal genomics era may also be around the corner, but as data output increases, data handling and storage may become increasingly difficult and potentially limiting. The prospects for the future of genetics remain extremely exciting, but as complexity increases, making the technology accessible presents a huge challenge.

Appendices

Appendix 1 Reagents and recipes

Nucleon Reagent A (pH 8)

6.304 g	Trizma hydrochloride
438.12 g	Sucrose
4.066 g	Magnesium chloride
40 ml	Triton
3.6 l	H_2O

Adjusted to pH 8 with 2 M NaOH.
Autoclaved.

Nucleon Reagent B (pH 8)

63g	Trizma hydrochloride
22.3g	EDTA
8.8g	NaCl
800 ml	H_2O

Adjusted to pH 8 with 2 M NaOH.
Autoclaved.
10 g Sodium dodecyl sulphate (SDS) added and volume adjusted to 1 l with ultrapure water.

Lysogeny broth agar (pH 7.5)

10 g	Bacto-tryptone
5 g	Yeast extract
10 g	NaCl
800 µl	H_2O

Adjusted to pH 7.5 with 2M NaOH.
15 g of agar added and volume adjusted to 1 l with ultrapure water.
Autoclaved and cooled to ~50°C.

1 ml	Ampicillin (50 mg/ml)
1 ml	Tetracycline (12.5 mg/ml)
1 ml	Blue-White Select (40 mg/ml X-Gal, 40 mg/ml IPTG) (Sigma-Aldrich)

TE - 50x stock solution

20 ml	Tris-HCl (1 M, pH 7.5)
400 µl	EDTA (0.5 M, pH 8.0)
19.6 ml	H_2O

TAE - 50x stock solution

700 ml	H_2O
242 g	Trisma-base
57.1 ml	Glacial acetic acid
37.2 g	Na_2EDTA

Adjusted to pH 8 with 2 M NaOH.

SBDD buffer

160 ml	1 M Trisma-base pH 9.0
3 ml	1 M $MgCl_2$
50 ml	Tetramethylenesulphone
290 ml	ddH_2O

Laemmli 2x sample buffer

4%	SDS
20%	Glycerol
125 mM	Tris, pH 6 .8
0.02%	Bromophenol blue
200 mM	DTT

SDS-PAGE running buffer

25 mM	Tris base
190 mM	Glycine
0.1%	SDS

Blotting buffer (western blotting)

50 mM	Tris base
380 mM	Glycine
0.1%	SDS
20%	Methanol

Appendix 2 Primer sequences

Illumina Genome Analyser library preparation primers

Primer Name	Sequence (5' to 3')
Multiplexing Adapters 1	5'Phosphate-GATCGGAAGAGCACACGTCT
Multiplexing Adapters 2	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
Multiplexing PCR Primer 1.0	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
Multiplexing PCR Primer 2.0	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
PCR Index primer 1	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTC
PCR Index primer 2	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTC
PCR Index primer 3	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTC
PCR Index primer 4	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTC
PCR Index primer 5	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTC
PCR Index primer 6	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTC
PCR Index primer 7	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTC
PCR Index primer 8	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTC
PCR Index primer 9	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTC
PCR Index primer 10	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTC
PCR Index primer 11	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTC
PCR Index primer 12	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTC

Genotyping primers for the episodic falling candidate gene study

Gene	CFA	Position	Forward primer sequence	Reverse primer sequence	Amplicon size (bp)
CACNL1A3	7	5.29	TTTTAGTCAAAGAGCTTGAAACG	CCTTATCTCACAAGGACATTCAGA	281
CACNL1A3	7	5.34	TCACAAACAACATAGTGCCAAT	TCTTTGAGGAAGTTTGTTTGCT	273
ATP2A1	6	21.26	GGTAAGTTTGGAGGGAGATGC	TGATCAGAGGCTGAAGGAGAA	215
ATP2A1	6	21.56	CAATGGGAAAGTGTCCTGCT	GTGGTATGGCTCCCTGCTC	256
CLCN1	16	9.26	GGTATTGTTTGCTCCTGTCCA	TCACATGCATGAGGTTCATTT	190
CLCN2	16	9.47	GAACAGTGCCAGCTAACAAGG	AGCAGCCTTCTCATGTGTTGT	87
GLRA1	4	60.64	AGCCTTGAAGCCAGTCACAC	TGGCTAGGAAGCAACTGGAT	147
GLRA1	4	60.68	CCCAGCCATGTGGAGTCTAT	TGTAACCGAGGACGGTTCTT	187
GLRA3	25	28.14	AACTGAGGATCTGGGCATTG	GCCCAGTTGGTAAGCATTTG	191
GLRA3	25	28.25	AGTTCTGTTCAGGGCAGGAAT	AACAGGTTTTGTCTGGTGGTG	132
GLRB	15	57.00	CCAAAGTTCGTTCCTGTAAAGT	TTATCCCAGTGACCAAGTGTG	140
GLRB	15	57.36	TGAGGCTGTTAAACGTCCCTA	GAGCAAATGGATTAGGTCTGC	108
GLRB	15	57.49	AACTTCCTTGTGATTGTCATGG	CCAAATTGGTACCGTTTGAAAG	193
GLRB	15	57.57	AAACAGTGTCTTCCCCATCAA	TAGGATATGAGAGGCCCAGGT	274
GPHN	8	43.92	GAAAAGGAATGATGTTGCAGGT	ATGGTCTTCAGGTTTTGCTCA	91
GPHN	8	43.98	CGGGAACCACATTTCTCATTA	ATTTCCTTCAGACCCTTTCCA	292
KCNE3	21	26.87	CCTCACACCTCACAGCACATA	CCAGCTCCAATGTCACCTCTA	233
KCNE3	21	27.01	ACAAATGGCAAACTTTCATCCT	GAACTGCCATGTGATCTACTAATTC	259
SCN4A	9	14.81	CCGCCCTAGAGGAAGATTTA	ATCTGGTCACTGGTGTGCAA	165
SCN4A	9	15.01	CTCTGCAGCCCACTTGTGTA	AACCCCAGTGGTATTGCTTG	188
SCN4A	9	15.36	TACCAGCCGGAGTAGTTCTCA	CTCCTAGGCAGGAGCCATTAC	280
SLC32A1	24	30.09	CTTGGTCCCTTCCTGTACCTC	TTTCATTCTCCACTCCTGTGC	233
SLC32A1	24	30.15	CCTGTGACCAGGATTCACTGT	CTGACTTCACGCCATACACCT	291
SLC6A5	21	45.757	GACAGGCCTCCTCTGTGAATA	TCTCCTCTTCACCTTCCATCA	164
SLC6A5	21	45.760	GCTTTCGACTCATCCTTACCC	CAGGGAGTATTTTCCCCTGTC	229
SLC6A5	21	45.77	ACTGTCCGCCAGTGTGTATG	GCAAGATGCATTGAAAGGTG	163
SLC6A5	21	45.92	CCAAGTCTGTGACGGAGAATC	TTCCTGATGATCTTGCCTTTG	172
SLC6A9	15	18.87	CTTTGTGCCACCAGTTTCTGT	TCTGAACCTGTGACGACTTGA	227
SLC6A9	15	19.01	CATTGGCTTCCTCAGTTTCAG	TGGGCAGTTCAAGAATGTACC	189
SLC6A9	15	19.012	CTGAAGCCTGTGGTTGGAAT	TCTGCTTCAGTCCCCTCTGT	220
SLC6A9	15	19.02	AACCTTGCCTAGGGACACCT	AATATGCAGGGAGCACATCC	215
SLC6A9	15	19.31	GACTGCCAACTCCAACCCTAT	GTGTCCAGGATTGACTCTCCA	284
SLC6A9	15	19.43	TCCTGCAACTAAGCCTTGAGA	CTGCATTCATAGACTGCCTTG	150

SLURP1 PCR and sequencing primers

Forward Name	Sequence	Reverse Name	Sequence	Amplicon size (bp)
SLURP1_5U_F	GATGTGGGCAGTGAACAAAGT	SLURP1_5U_R	GGAAGTGGCTGTGAGTCAGC	556
SLURP1_X1_F	ACTGCCTGAGGGTCTGTGAGT	SLURP1_X1_R	TTCCTCTCTGAGCCTTGGTCT	468
SLURP1_X2_F	CTCAGAGAGGAACGCAGTGTC	SLURP1_X2_R	CTCCAGTGATACCCTGTGACC	566
SLURP1_X3_F	GGTCACAGGGTATCACTGGAG	SLURP1_X3_R	CCTTCCATGAAATGAGTCCAA	541
SLURP1_3U_F	CTGGCAGAGAATAAAGCAAGC	SLURP1_3U_R	CCAAGCTGGCATGTCCTTAG	511

BCAN and *FAM83H* cDNA PCR and sequencing primers

Forward Name	Forward Sequence	Reverse Name	Reverse Sequence	Amplicon Size (bp
BCAN_cDNA_1	AGTGGAGAAAGGGGTTTTGTG	BCAN_cDNA_2	CACTCAGCACCAGGGACAC	582
BCAN_cDNA_3	GGTCAAGTGGACCTTCCTGTC	BCAN_cDNA_4	CTGGGTCCACCACTCCATAGT	500
BCAN_cDNA_5	TTCTCTACCGGGAAGGCTCT	BCAN_cDNA_6	GGAAGCAGTACACGTTGAAGC	585
BCAN_cDNA_7	AGTGTGCGCTATCCCATTGT	BCAN_cDNA_8	GATGGCCTAGGCTGTAGGACT	626
BCAN_cDNA_9	GTCCTCCGAAGAGGAAGACAA	BCAN_cDNA_10	CTGAACCCGCACTGAGGT	547
BCAN_cDNA_11	ACAGGGAGCTCTGAGGATAGC	BCAN_cDNA_12	GCCATCTGACCACAAGAAGTC	531
BCAN_cDNA_13	GGAGGAACAGGACTTCATCAAC	BCAN_cDNA_14	CGTGGTCACTTCCTATGATGG	628
BCAN_cDNA_15	GTAGACACGGTGCTTCGCTAC	BCAN_cDNA_16	GGCTGGTTTTACTGGTTCTCC	577
FAM83H_F3	GGTACCTGCCACCTCACTACA	FAM83H_R3	GTCAGCCATGTCCAGGAAGT	530
FAM83H_F4	CTACTGGCCCATGAACTCAGA	FAM83H_R4	CTCCTCATCGAAGCTGGAGAC	538
FAM83H_F5	TCTGCCCAGCAGGTGGTG	FAM83H_R5	GGTCGAGGAAAGAGGGGAATC	586
FAM83H_F6	GTGGATTTCCTGCGCGTG	FAM83H_R6	AAGGCGTCCATCTCCAGGT	599
FAM83H_F7	GGAGCTACAGCTTCATGTGGT	FAM83H_R7	TCTGAAAGCGCAGGTCGT	594
FAM83H_F8	AAGCGGCACAGCTTCGCA	FAM83H_R8	CATGGCAGCCGCTCAGGTAG	556
FAM83H_F9	TGGACTACGTGCCGTCCAG	FAM83H_R9	CGCTGAAGATCAGGGAGGAG	543
FAM83H_F10	CGAGGCATACGAGGACGAC	FAM83H_R10	AAAGGAGTCGCGCAGGTC	571
FAM83H_F11	CCGAGCTCCTGGAGAAGTACA	FAM83H_R11	GACTCCTCGGCGAAGGTAAG	553
FAM83H_F12	AGGCACCTCACCCTGAGC	FAM83H_R12	GATCTGCTCCAGAATGGCTTT	500
FAM83H_F13	CAGCTGCTGAGCCCCAAG	FAM83H_R13	ATGAACTTGCCCACCTTGCT	599
FAM83H_F13	CAGCTGCTGAGCCCCAAG	FAM83H_R14	GTTCTGCTGCCTGGTGTGAAG	723
FAM5END_1F	AGCCACCTGACTTGCTCAGTA	FAM5END_1R	ACTGCCAGGCGGTAATACTCT	566
FAM5END_2F	CGCCCTTTCTACACTGTGTCT	FAM5END_2R	GTCTGAGTTCATGGGCCAGTA	420

PCR primers for fine mapping of the IS spinocerebellar ataxia disease-associated region

Marker type	*Canfam2* position chr20 (Mb)	Primer pair name	Forward sequence	Reverse sequence	Amplicon size (bp)
Micro	chr20:13.32	CFA20_13.17	GGGGTTCAATCTCAAACCATT	GTATACATCGGGCCAAAGACA	233
Micro	chr20:13.68	CFA20_13.53	GGTCACAAAACTGCTTTGCTC	GAGCTTACCATGTGCCATATTC	196
Micro	chr20:13.85	CFA20_13.70	GCCCATTGTTCACACATAACC	TTTCAAGGGATGCTGAGATGT	230
Micro	chr20:14.33	CFA20_14.18	TCCTCCCAGATAGCCTCTGAT	TTTTGGAGTAGGCATCACGAC	189
Micro	chr20:14.42	CFA20_14.27	CCTGCTGGCCTACTCTGAAA	GGAGTTCTCCAGAGAAACAGGA	264
Micro	chr20:14.49	CFA20_14.34	TTTGGCTATGCCTGAACCTT	CTCAATGGCCACACTATGGA	184
Micro	chr20:14.53	CFA20_14.38	CCACCCACTAGAATGTAAACTCC	GAGCCAGAACTATTCCAGATGC	116
Micro	chr20:14.59	CFA20_14.44	CGATGTCTCCTGACCTCATTG	CCTGAGTTGTACCAGGGCATA	216
Micro	chr20:15.18	CFA20_15.03	GGCTCCATATTCAGTGGGAAT	TCCTGGTACTGATGAGGCACT	146
Micro	chr20:15.28	CFA20_15.13	TGAAACAGAGAAACAGGAATGG	TAACAGGGTGAGTTGCACTGG	121
Micro	chr20:15.42	CFA20_15.27	GGGCACAACTTATTGCAGACT	TCAGACCCATTGGCCATAGTA	244
Micro	chr20:15.52	CFA20_15.37	AGCTCAATTACCCAACAATGC	TTAAACCAAAGCGTCTTCAGC	225
Micro	chr20:15.54	CFA20_15.39	CATCAGGCTTCCTGTATGGAG	TCTCATGACAACCCTGAGAGG	151
Micro	chr20:15.74	CFA20_15.59	GGAGCTCCTCACTGGGTCTTA	TCTTATGGTCCAAACAGGTTGA	268
Micro	chr20:15.75	CFA20_15.60	CTGGCCCATTAGTGCATAGTC	GGTTCACTCTAGTCATTGTTTGC	188
Micro	chr20:15.89	CFA20_15.74	AGGCTTCATCATGTGAAATGG	CTGATTTGGTTGGAATTGGTG	168
Micro	chr20:15.97	CFA20_15.82	AGCAGAGATTGTTTGCTGCTT	GCCCTTTACTTACTGGCCATC	166
Micro	chr20:16.05	CFA20_15.90	TGTAGGCCACACAACCATGTA	CCCAGATGTTAATAGCCGTGA	259
Micro	chr20:16.21	CFA20_16.06	TGGAGGGTAAGAAGTCCCAAC	TCCCTCATCTTGACAACTGGT	228
Micro	chr20:16.57	CFA20_16.42	AGGCCTCAGTGTTTCACCATA	GTCCAAAATGCATTGTTACTCC	150
Micro	chr20:16.91	CFA20_16.76	AAACAGAGGGTCAGGCTCATT	CCCATAGTGTGATCCTGGAGT	208
Micro	chr20:16.92	CFA20_16.77	ATAAATGGCAACACTTCCTTCC	GCCAAAACAGAGAAGGAACCTA	137
Micro	chr20:16.95	CFA20_16.80	GTTGTGTGAAGGCCAGAAGAG	GTGGCCAAAAATCCCTAAGAG	238
Micro	chr20:16.97	CFA20_16.82	GGGGGTTTCTTTATTTTGGTG	ACCCAAACATACTAACTTGAAAAAGG	279
Micro	chr20:17.01	CFA20_16.86	TTTTCATGGGCCTTAGTGAAG	TTCCTTATGATGCTGGTGGTC	299
Micro	chr20:17.04	CFA20_16.89	GGATCGAGTTCCGTTTCG	TGGTTCTTGTGAGTGTCATGG	272
SNP	chr20:17.14	DOG_SNP_9	TAGCAAGGGAAAAGCCATCTT	GTCAGACAGCTCAATGCATCA	402
SNP	chr20:17.15	DOG_SNP_8	ATACCCCAAAGAGTGCTGGAC	TCAGGAGGGATCAATCTAGGA	425
Micro	chr20:17.16	CFA20_17.01	AATGAGCTCAGTTGGGGAAG	TTAGCACGGTTGTTTTCCTTG	305
SNP	chr20:17.19	DOG_SNP_5	CGCTATGGTGTCAGACCACTT	AACACAGGCCTGCTTACACCT	467
SNP	chr20:17.21	DOG_SNP_3	TTCACCAGGCCACAGTATAGG	GAAACCAGGAATTGGATGCTT	500
SNP	chr20:17.22	DOG_SNP_2	TGGGGACTCTTCTTCACCAC	GGCATTTCAAGGTCACAGGT	544
SNP	chr20:17.23	DOG_SNP_1	TGAAAGGACACATGCAAAGC	GATGCATAAGCCATGGTGTG	358
Micro	chr20:17.24	CFA20_17.09	TGGTTAATGTCTGCCTTCAGC	ACCTCCAACTTGGTTCCTGAC	220
Micro	chr20:17.31	CFA20_17.16	CCTAACCAGAGAGAGGCCAAT	AAGATAATCCCTGCGGTAAGA	115
Micro	chr20:17.39	CFA20_17.24	CAAGGCCTGTGTTACCCTGA	TGTAGGTGCCCCACTTCTTT	289
Micro	chr20:18.22	CFA20_18.07	CAACTCAGCAGTTGGGTTAGC	GGTGGACAAGATTTGTTTTGC	126

Primers for non-quantitative end-point PCR checking for gene expression of genes in the IS spinocerebellar ataxia disease-associated region

Forward name	Sequence	Reverse name	Sequence	Amplicon size (bp)
*BHLHE40*_cd1f	GACGTTCCGCTACTGCAGAC	*BHLHE40*_cd1r	TCTGAGACCACACGGTGAAG	594
SUMF1_cd1f	CACCAAGATGGTCCTCATCC	SUMF1_cd1r	GCACAGCGATACCTGTAGCA	700
SAP18_cd1f	ACGAGTTTTGCAACGGAAAC	SAP18_cd1r	AAGGAGGTGGTGCCTGATTT	253
SETMAR_cd1f	CTTGGAGAACGTGCCTGTG	SETMAR_cd1r	CATGCCCAGTTCCTTCTGAT	871
XP_cd1f	AACCTGCAAGAGGTTGGAGA	XP_cd1r	CTGGGATGGACCAGTGAGAT	509
LRRN1_cd1f	GCCCTCTGCACCCTACTTCT	LRRN1_cd1r	TTCCAATCACAGGGTTTTCC	893

Primers for PCR and exon resequencing of SCA disease-associated interval genes

Forward name	Forward sequence	Reverse name	Reverse sequence	AmpliconSize (bp)
BHLHE40_U1_F	CCAGGAGACGGGAACTTACTT	BHLHE40_U1_R	AGCCAAGTGAATGAGAAGTGG	505
BHLHE40_U2_F	AACTGAAGCAGCACCTCAAAG	BHLHE40_U2_R	GCGGTGTGTCTTACCTTGCTA	514
BHLHE40_1_F	GGGTAGAACACGTAGCTCCAA	BHLHE40_1_R	CAACTTACCCAGGTAGGTCTCC	520
BHLHE40_2_F	GAGACCTACCTGGGTAAGTTGG	BHLHE40_2_R	GCACTCGTTAATCCGGTCAC	575
BHLHE40_3_F	CTTCTGTGAAACGGCTTGAAC	BHLHE40_3_R	CGGTTCTCACTTCATTTGCAT	461
BHLHE40_4_F	GTCAGAGCTTTCTCCTGATGC	BHLHE40_4_R	CAGCATTCCAAACGAAAAGTC	576
BHLHE40_5.1_F	CCCGATCGTATTACTTTTGGA	BHLHE40_5.1_R	ATTCTCCTCCGTAGCCACTGT	501
BHLHE40_5.2_F	CTTCAAGGAGAAACCCAGCTC	BHLHE40_5.2_R	AGTCTCTGAGGCATGAGCAAG	543
BHLHE40_5.3_F	TGCTGGAGAAGTGCTGGTATC	BHLHE40_5.3_R	AAGCCCAGAGATCTGCTACAA	521
BHLHE40_5.4_F	ACCAAAGACTAGGGGACCTTG	BHLHE40_5.4_R	GCTATGCCAGTGTCTGCTACC	585
BHLHE40_5.5_F	CCCCCTCAGATACATCCAAAT	BHLHE40_5.5_R	TCTAGCTCTGCTCGTTGAAGG	594
BHLHE40_5.6_F	TGGTGCCTATTCCTAGGTCTC	BHLHE40_5.6_R	AAGGCAATGACTTTATACCAAAA	455
BHLHE40_5.7_F	CTTGTTTTCCGACTCATCCAG	BHLHE40_5.7_R	GTATCACACAACTGGGGCATT	448
ITPR1.U4_F	CTCGCGTTAGTCCACCTCCT	ITPR1.U4_R	GGACTACGAGTCCCAGAATCC	550
ITPR1.U3_F	TCTGCCTGAGCTACCTGGAT	ITPR1.U3_R	GAGCTCGTTCCGTTAGGATG	507
ITPR1.U2_F	GTTCTGCTGAAGCGTTTCCT	ITPR1.U2_R	AATGCCATTCCAGTCCAGAA	478
ITPR1.U_F	GAAGGCTACCCATTCTGGACT	ITPR1.U_R	AAACTGCCACCTTCTTTAGCC	524
ITPR1.ex1_F	TTTGCGTGCCTAGTCTCACTT	ITPR1.ex1_R	ACTGCCCTCTGGATCTCAAAT	490
ITPR1.ex2_F	TCATTTAAGTGCCTGCCTGTC	ITPR1.ex2_R	TGACAAGCAAGTCAGCATTTG	642
ITPR1.ex3_F	CCATCATGGATGTGGTTAAGG	ITPR1.ex3_R	TCGTGATGAGACCCAAAAATC	648
ITPR1.ex4_F	AAGGGCCTTGTCTTGTCATTT	ITPR1.ex4_R	TCCGAGAAAGACACTGGCTAA	438
ITPR1.ex5_F	TATGAACTGGCGACATTGGTT	ITPR1.ex5_R	AATTGCTCTTATTGGCCCTGT	581
ITPR1.ex6_F	AACATCCATCTTGTCCACAGC	ITPR1.ex6_R	TGTTTCCCAAACACTGGTTTC	558
ITPR1.ex7_F	CTTCAGGCAATGTTCTGCTTC	ITPR1.ex7_R	GTAAAAGAGGCCCTTCACTGG	662
ITPR1.ex8_F	CTGCTTCTTGGGCAGTATGAG	ITPR1.ex8_R	CACATCCAGCATCCTTCAGAT	563
ITPR1.ex10_F	AGGGATCCCGAGTTGTAGAAT	ITPR1.ex10_R	AAAGCAGAAGGAAAAGCAAGG	535
ITPR1.ex11_F	CAGAGGGAAGGGACAAGTTTT	ITPR1.ex11_R	GGGCATTCCAAGAGATACCAT	450
ITPR1.ex12_F	ATGGTTCTTTTCCTGCCAACT	ITPR1.ex12_R	GCTTCAAGAAAGGCAGTGATG	384
ITPR1.ex13_F	TCAGTATGCAGACGTTTTCTCC	ITPR1.ex13_R	CCAACAAAACACAGCTCCCTA	364
ITPR1.ex14_F	AGCTCCAGCTTTTTGGATAGC	ITPR1.ex14_R	TCAGAGTAACACTGGCCCCTA	538
ITPR1.ex15_F	GGAAGAATGGATCTGCTTGTG	ITPR1.ex15_R	AGTTGATCAAGCCTTCCAGGT	363
ITPR1.ex16_F	CCACCACCTGTGAGGTTGTAT	ITPR1.ex16_R	GAAGGGAACTTCTCTGGCATC	454
ITPR1.ex17_F	TTCTCATCAGGGACTGTTGCT	ITPR1.ex17_R	AGCTAGGATTCCACCACAGGT	436
ITPR1.ex18_F	TCCATCCACTGCTTCCTTATG	ITPR1.ex18_R	AGTGCAGACATCCACTGAACC	496
ITPR1.ex19_F	TGCAAGGCAGATAGAATCACC	ITPR1.ex19_R	GACTCCCTGGAACAATCAACA	507
ITPR1.ex20_F	TCGATGTTCAGGTGTTCAATG	ITPR1.ex20_R	CATGCCATCTTTCACATTCTTC	574
ITPR1.ex21_F	GGCGCCTCAAATGATTACATA	ITPR1.ex21_R	GATAGGATTCACCCTGGGTACTT	556
ITPR1.ex22_F	TCAGAAATGGCAAGCAAGGTA	ITPR1.ex22_R	AGTGGAGAGTTGGCGTCTTCT	562
ITPR1.ex23_F	TGAGTCCCCAGAGTTTTGATG	ITPR1.ex23_R	AAATGGCACTGAACAATGAGC	502
ITPR1.ex24_F	TTGTCAGGGTGCTAAAGTTTG	ITPR1.ex24_R	CCAAGATGGAAGGAAACAAGA	600
ITPR1.ex25_F	ACCTGGCTTCACAGCTTTACA	ITPR1.ex25_R	GCACTTGGGGAAACCTACTTC	427
ITPR1.ex26_F	AGGAGCTGGGAATCTGGATAA	ITPR1.ex26_R	TGGTAACAAGCAGCTGAACCT	554
ITPR1.ex27_F	CTTTCGGCACTTCAGTCAGAG	ITPR1.ex27_R	GTCCAAGGTGACACGGAAGTA	527
ITPR1.ex28_F	CAAGTGACTCAGGCTGAAACC	ITPR1.ex28_R	GCAAGCTTCTTTCTCCCTGTT	346
ITPR1.ex29_F	GAGGTGGGCGAGAACTCTTAC	ITPR1.ex29_R	CTTCATGTCTTCGCCAATGAT	386
ITPR1.ex30_F	GTCTGCTGGCCTAGCAATATG	ITPR1.ex30_R	ACCAGCAACTCTTTGGTGCTA	512
ITPR1.ex31_F	TGGCAGAAAGTGAGAAAGGAG	ITPR1.ex31_R	CTTAGGGTTGGGGAAGTCAAC	500
ITPR1.ex32_F	TGATACATGAGTGCTGGTCTGA	ITPR1.ex32_R	AGGTCTCTGCACCATGAGGT	584
ITPR1.ex33_F	ATTGCCCTACATGGTTGTTCA	ITPR1.ex33_R	GGCAAGCTGAAATTATGTGGA	469
ITPR1.ex34_F	ACACGTGGCTAGGAACTTGAA	ITPR1.ex34_R	TATGAATGCCTTGCAGAATGG	517
ITPR1.ex34B_F	CTTTCTTGGGTCATGGTGAGA	ITPR1.ex34B_R	CACACGAAATTGCCTTATGCT	392
ITPR1.ex35_F	CAGGGACCTGAGGTAATTTGC	ITPR1.ex35_R	ACTTTCTGAGAGCAGGGCTTT	566
ITPR1.ex36_F	CATGCAAGTGACATGTTTTGG	ITPR1.ex36_R	CTATGGCCAAACTTGGGTTTT	409
ITPR1.ex37_F	TTGCCTAGCTTGTGGTCATTT	ITPR1.ex37_R	GGCACAGAAGATGATCGGTTA	483
ITPR1.ex38_F	CCCTTTTGCTGTTACTGACCA	ITPR1.ex38_R	GGCAAGCAGCCTATAAGACAA	551
ITPR1.ex38b_F	CAGATGTCACCGTGAGCTGTA	ITPR1.ex38b_R	GGACAAATGGCAAACAAACTC	401

Primers for PCR and exon resequencing of SCA disease-associated interval genes (continued)

ITPR1.ex39_F	AGCCATGTCACCTCATCTTTG	ITPR1.ex39_R	TCAGAGAGGCACATCTCACCT	404
ITPR1.ex40_F	AATTGTGAGGATTTGCCCTCT	ITPR1.ex40_R	TCAACACCAACAAAAGCTGTC	466
ITPR1.ex41_F	AACTGAGTACCCCCGTGTTTT	ITPR1.ex41_R	AATTACCCAGAGGGTTGCTGT	460
ITPR1.ex42_F	ACCTTCCTTGTGGGGTTTTTA	ITPR1.ex42_R	TCCATCATAGAAGAGCCTTGG	431
ITPR1.ex43_F	CAGGCTCTGATCTCTGTGAGG	ITPR1.ex43_R	ACCAATGCATGGTAGTTCAGC	415
ITPR1.ex44_F	TTGGAGTGTTTGGCTAACCTT	ITPR1.ex44_R	TTGTCCCTGTGGAGATGGTAG	551
ITPR1.ex45_F	CATCATGGAGGCTGTCTCTTC	ITPR1.ex45_R	CTAGGCCAACAGGCAATGTTA	491
ITPR1.ex46_F	TACTGGGAGTGGCAAGAATTG	ITPR1.ex46_R	ATTTGCAGGGGAAGACTGAAT	414
ITPR1.ex47_F	TCTGCCATAGGATGAAACGTC	ITPR1.ex47_R	TCTCAAGAGGGGTAAGGAGGA	408
ITPR1.ex49_F	CTTGGTCCAGCTACAGCAAAT	ITPR1.ex49_R	AAAGGAAAAGCAAGTCGAAGC	446
ITPR1.ex50_F	CATGTCCCTCATAGCCACTGT	ITPR1.ex50_R	CTGTGCTTGGACATCATTCCT	506
ITPR1.ex51_F	TGAATGCTGAGCCTATGACCT	ITPR1.ex51_R	CGCTTACCAGCTACAAAGGTG	462
ITPR1.ex52_F	TGAGCCTATGAGTCAACCTCAA	ITPR1.ex52_R	AGGCTCATGGAGACCATTTCT	561
ITPR1.ex53_F	GTGTGACGGACTACGTCTGGT	ITPR1.ex53_R	GCAATTTTCTGCACTGAAAGC	435
ITPR1.ex54_F	CTTAGATCTGGGCACTTGGTG	ITPR1.ex54_R	CTCTAGATCCCGGAAATGAGG	457
ITPR1.ex55_F	CCCATCTCGGTTTTCACTATG	ITPR1.ex55_R	CGTACCATACTGCCCTTTGAA	550
ITPR1.ex56_F	TGTGCAGACTTGGTGTAAACG	ITPR1.ex56_R	TACTGGCCTTCCTTCTGAACA	526
ITPR1.ex57_F	AGGTGACTCCTGCTTTTTGG	ITPR1.ex57_R	AGTCACACCCTCCGTCTAACA	595
ITPR1.ex58_F	ATGTTAGGCCCATCGAGTCTT	ITPR1.ex58_R	ACGCCAAGACACATTCATTTC	532
ITPR1.ex59_F	AGACCCAAGTCCCTCTCACTC	ITPR1.ex59_R	GAGCCTCATTGTTTCCTCCTT	554
ITPR1.ex60_F	GCCTGATGAAAGAGGTCACAA	ITPR1.ex60_R	AAAGTGCAGTGAAGGGGAACT	454
ITPR1.ex61_F	TGCCTAGGAAGATGTCCCTTT	ITPR1.ex61_R	GGTTCAAGTGCAAATCAGGTG	538
ITPR1.ex61.2_F	AGTTCTGATTCACCCACAAAGA	ITPR1.ex61.2_R	GAGTTTAAATGATCAGGTCAGAAGC	621
LRRN1_1_F	TGTGCATTTGATGATTTTACCC	LRRN1_1_R	AGCTCAACTCCTGACACTCCA	481
LRRN1_2_F	AGGCACATCTCCAGACTTTGA	LRRN1_2_R	GCCAGCCCTACCTCCTTAATA	645
LRRN1_3_F	GATTGCAATGATCTCCGCTTA	LRRN1_3_R	CAAGTTGAGGGACTTTGACCA	604
LRRN1_4_F	AGTTATCGATAGCCGCTGGTT	LRRN1_4_R	TTTTGGTAAACGGCATTCAAG	525
LRRN1_5_F	GGAACTTGTTTCTGTGGATCG	LRRN1_5_R	TGGGCAACACAAGTGTATCTTC	612
LRRN1_6_F	TACTGGGTCACTCCTCTTGGA	LRRN1_6_R	CTAATGACGGCAAACATGGAT	560
LRRN1_7_F	CAGGGTCCCAGTAGATGTTCA	LRRN1_7_R	GAAATATCCACCCGTCCTCTC	574
LRRN1_8_F	TTAACCTCTGGGAAGGTGACA	LRRN1_8_R	CAAAAGGTTCAATGCTGCTTC	547
SAP18_1_F	GCAGGTACAAATCCTCCATCA	SAP18_1_R	CCAAACTTGTCAGTCCTGCTC	398
SAP18_2_F	AGAGGCCAGAAGAAGGATCAG	SAP18_2_R	CTCGATAGCCAGGCCTTTTAG	562
SAP18_3_F	TACACTTGGATGGATGCAACA	SAP18_3_R	GCAAAACTGCTTTGTCATGGT	430
SETMAR_U1_F	CGGCGTCTGTAAGACAGAAAG	SETMAR_U1_R	GAGCAGGGTTCTGAAGTGCT	556
SETMAR_U2_F	TTCTGGGAAACCGAGAACAC	SETMAR_U2_R	CCTCAGGCTCCTCCTCAGAC	572
SETMAR_1_F	TCGGGAAAGGTGAGAGAAAA	SETMAR_1_R	AGAGGCGCCCACAGAACTAC	448
SETMAR_2_F	GGTCAGGTTCCCTGTTCAGT	SETMAR_2_R	CCAGCCTTTCTTATCCGTCTT	483
SETMAR_2.2_F	GTCCTATGCCAGTGCAGTGAT	SETMAR_2.2_R	TGATTTGGCACCACAATAACA	522
SETMAR_3_F	GCCAAAGATATTTTGCCAGAAG	SETMAR_3_R	AAACGTAGTCATTTAGCCACATCT	531
SUMF1_U_F	GGTTCTCTCATGTTCCCAGGT	SUMF1_U_R	GAGTTGATTTGGGAGGTAGCC	569
SUMF1_1_F	GGGTTCACAGTTGGCTAGAAGT	SUMF1_1_R	ATCCCGACCCTGTCCAAG	545
SUMF1_1.2_F	CTGGGTCTCGTCTTCCTGCT	SUMF1_1.2_R	TGGGGTGTGGTTAGAATCTCC	364
SUMF1_2_F	GCTGGGTGTTTACTGTGTGCT	SUMF1_2_R	AGGCTTGAATCCTAGCTCCAC	497
SUMF1_3_F	AGCTGAGGATACCTGGTGGTT	SUMF1_3_R	AGGAGGGTGGCATTTGTAGTT	466
SUMF1_4_F	CATGCTGGTGCTTATGATATGG	SUMF1_4_R	GGCACAGTGCTTGGGTATTTA	480
SUMF1_5_F	CTGTGGTGGAAGGTGACAGAT	SUMF1_5_R	CAAAAGGCTCTGAGGAGGACT	551
SUMF1_6_F	TGAGCTAGTTGTGCGGAAGTT	SUMF1_6_R	TGAATGCAGGAAGAGACTGGT	401
SUMF1_7_F	CATTTTGACTGTCGGGATGAA	SUMF1_7_R	TGCTCACAGACATTTCTGCAC	445
SUMF1_8_F	ATGCTGCCAAGCAGCTTATTA	SUMF1_8_R	AGTGGATGGGACATTTTAGGG	460
SUMF1_9_F	TGGTGAACTGTGAATGTCCAA	SUMF1_9_R	TAAAGCACAGCGATCATCTCC	514
XP53_U_F	TGAAAGGGAACTTGGAGATCA	XP53_U_R	GACCAAGGTCTAGTGCGACAG	513
XP53_1.1_F	TCCCTTGAGCAAAAGCTGTAA	XP53_1.1_R	GTCTTCTTTGACACCGACGAG	466
XP53_1.2_F	GAAGAGGTTCGCAACTGTGAC	XP53_1.2_R	TGTGAAATTCAGATCGCAGTG	514
XP53_1.3_F	TTGAGTGGAAACCCTCTCCTT	XP53_1.3_R	CAATTCCCAAGTCCTCATCAA	368
XP53_2_F	TTAGAGGAGGATGGGATTGCT	XP53_2_R	GCATATCTGCCACATCAGTCA	461
XP53_3_F	GTCTTGATTTCCTGCCTTTCC	XP53_3_R	GAGCTGCATAATGAGCCAGAC	359

Primer for *ITPR1* cDNA sequencing

Forward name	Forward sequence	Reverse name	Reverse sequence	Amplicon size (bp)
ITPR1_cd1_f	AGTTCTGGCTTTGATCCTGGT	ITPR1_cd1_r	AATGCTGACTTGGAGCACTGT	1538
ITPR1_cd2_f	GACCGTGAACACCAGTGACTT	ITPR1_cd2_r	CCACCCTACACACGTCAGAGT	1877
ITPR1_cd3_f	CTCATGGTCGAAATGTCCAGT	ITPR1_cd3_r	TCCAGCTCATCCTTGGTCTTA	1966
ITPR1_cd4_f	CCGACATCCTAATTGAGACCA	ITPR1_cd4_r	GACGTCCTCTCCCGAATTAAC	2081
ITPR1_cd5_f	ATGCTGCAGACTTGGAAAAGA	ITPR1_cd5_r	TTGGAAGCTCCGCTACTATCA	2197
ITPR1_cd6_f	TCTGGGACTCGTAGTCCTTCC	ITPR1_cd6_r	GCAGAGCTGGAAGCCTCTTAT	861

Internal primers for sequencing

ITPR1_cd1_in1f	GCGGAGTAGGAGATGTGCTC
ITPR1_cd1_in2r	TCCAAACCTTTAGAGTAGCAATCA
ITPR1_cd2_in1r	CGAAGAATTGTGGAGGCAAT
ITPR1_cd2_in2r	GGGCACAGGAAACACAATCT
ITPR1_cd2_in3r	GATATTGTGCCCCACATTCC
ITPR1_cd3_in1r	GTAGTGGTCATCGGGGTCAG
ITPR1_cd3_in2r	CTTGTCAAGGTGGCACTGAA
ITPR1_cd3_in3r	GGTCATCATTTCTCGGAGAGTC
ITPR1_cd3_in4r	ACTTGGCTGTCAAGGTCCAC
ITPR1_cd4_in1f	CTGCCGTCTCATGCTTCATA
ITPR1_cd4_in2f	AGCAAAATGGCAAAAGGAGA
ITPR1_cd4_in3f	TCCCAGACTTCGGAAACATC
ITPR1_cd4_in4f	TTCAGTCAGAGGCAGGAGGT
ITPR1_cd4_in5f	GTGCCTCTGTGAGGAAGAGC
ITPR1_cd5_in1r	AGAGCCACACCTCCTCTTCA
ITPR1_cd5_in2r	GCCGGCAGATATGTCTGAAT
ITPR1_cd5_in3r	TTTCTCCAGCTTCCCAGCTA
ITPR1_cd5_in4r	ACTCCAGGCATTCTTCCTCA
ITPR1_cd3_altf	CGAAAGTGGTGGCTTCATTT
ITPR1_cd3_altr	GTGCTGTATGGTGGTGTTGC
ITPR1_cd3_alt2f	GCGGCTTCTAGAGACTACCG
ITPR1_cd3_alt2r	TGAAACACTCGGTCACTGGA

Long range PCR primers across *ITPR1*

See Appendix 10 for PCR products on agarose gel

Gel Lane	Forward primer genomic position (kb)	Forward primer sequence	Reverse primer genomic position (kb)	Reverse primer sequence	Annealing	Amplicon size (bp)	Conditions
2	16077.9	CAACTGAGCCTGAGAAAGCTGAAG	16068.6	CTTTGGAGAAATCCATCACCACTG	64	9300	
58	16069.0	TGGGGACGTTAGTTAAAGTCTCTGC	16067.2	TCAGCAGAACATCCACATGGTTAC	64	1833	GC
59	16067.3	GATTCTGGGACTCGTAGTCCTTCC	16065.2	CTGGTCCTCCACCATGTTATTCTG	64	2067	GC
3	16065.5	GCCTGAATGTCAAGGAATCTGATG	16060.7	CGGGGTAGGTGATGAGTAAAGACC	64	4778	
4	16061.0	GTCATTCAGCTTTCATCACCCATC	16050.9	ACTCTGGCCATGATGAACTGTCTC	64	10131	
5	16052.5	TTAGTTCCTGACGCCAGTGCTTAG	16043.6	CTCTTCTCGAAACAGGAGGAGGAG	64	8980	
6	16044.6	ACTGCAAAGTTGTTGGCTGACTTC	16033.8	AGTGCCATCCTGAGACTGTGATTC	64	10857	
7	16035.3	ATGAGCAAGGCAGCATTCTGATAG	16024.1	CTGCTGATTCCATTCCTGTAGGTG	64	11184	
8	16025.6	CTTGCCCTTAAGGAATTTGCAATC	16014.2	CAGTGGGTTAGATGGGTACTGTGC	64	11454	
9	16014.8	TGACCTGGATTGCTGAGAAGTAGC	16005.1	GCAAATCAGGAATCCCTACAGGAC	64	9701	
10	16006.4	CCTGTATTGATTTCGCTGGAGATG	15995.3	CTGCAGTACCAATCTACCCTGGTG	64	11154	
11	15996.9	GCCGTCTGAGCAGGTAACCTAAC	15986.1	ATGGCTAATAAACTGTGGGGATGG	64	10769	
12	15987.9	TTCTGGACTAGCAGGTTGTGGATG	15978	CACCTCAAAGGACTGCATGAAGAG	64	9889	
13	15978.5	CCTGGTGGTACCTAGACATGGTTG	15976.1	TTTTAGGGACAAGGCATACAATGG	64	2394	
14	15976.3	ATCTCACAGAGACTTCCCCTGACC	15974.6	CTGTGACCCAACCGTAAATGAAAG	64	1699	
62	15975.2	TCAGTCTCTGCATCCAGTGTTGAC	15969.6	GGTTTAACAACCCAACCCTCTGAC	64	5604	GC
15	15970.1	AGTGCAGTGGACGATTAAGACAGC	15959.9	AATTCTACATCCCTGGCACCTCTC	64	10216	
16	15960.3	CCTCTCCACACATTGGATTAGGTG	15950.5	TCAGCAATGAGTGAGTGACGTAGC	64	9864	
17	15951.0	TTCCTAAAAGGGCCTTGTCTTGTC	15940.2	CTTCATTGTTACCCTGGTGAGTGG	64	10804	
18	15941.4	GAAAATAGTGCGTGAGTCCAATGC	15930	ATAAGCAGGGAAATGAGGCAAGAG	64	11494	
19	15930.4	TGGGAGGCTATTTCTTCTTTGGAG	15920.7	GCTCTCCCAATGTTCATGTAATGC	64	9713	
22	15921.5	AAGCTTTTAGGTCAGATGGCCAAG	15911.4	GGGTGGGAGGAGTCATTTTGTAAC	64	10121	
23	15912.9	GAACATCATGTCTGATCCCCAAAG	15901	TCTGAGCTCTTGGTCATTGTTTCC	64	11959	
24	15902.4	CTGAGGGTGATTGTTGGTAACAGG	15893.7	TTCACCAAGAACTTCCTTGAGCAC	64	8769	
25	15894.3	TCCAGGCAGTTTAAGCACTACAGG	15889.3	AGTGGACAAAGTGCTGGACAACTC	64	5045	
26	15889.4	AACAATTTCCAGCTCTGCAGTGAG	15888.4	GACCAGCACTCATGTATCAAGTGG	64	976	
27	15888.6	CCCTATACAGTCCTGGGGACAATG	15885.5	TCAAACTACCAGGGCAGAGTTGAG	64	3221	
28	15886.2	CTCAGACACCTTCCTGAGCTTGAG	15875.6	TCCCATCAAACGGTGGTCTATATG	64	10603	
29	15876.0	AGAGATTGGCTGGCCTCAGTAAAG	15872.2	CGGTGCATTCACTTGTAGGGATAG	64	3825	
30	15872.4	AATGGACCTCTGTCCTTCATCTCC	15870.7	CTGCTATTAACCTTTGCCCATTCC	64	1659	
31	15870.9	TGTAGCTTGGTGGCAACTTTATGG	15869.5	AAGACGGAAATGACCTCACACTTG	64	1403	
32	15869.8	GGGGTGTTGCATTAAACCATTTC	15868.6	TTGTGGGTGGACTTCAAAACATTC	64	1193	
66	15868.7	TGCCTAAGGTCGTCACACAG	15868.3	TTCTTCTAGCAGCCCTGAGC	58	472	HST+
70	15868.3	GCTCAGGGCTGCTAGAAGAA	15868.2	TCATTGTTGGTCCTCTGCAT	58	124	HST+
33	15868.2	ATGCAGAGGACCAACAATGA	15866.4	AACACGTGGACCAAGAGAACAGTG	58	1822	
34	15866.6	GAGTTTGTTTGCCATTTGTCCTTG	15863.1	GGCAATTACTGCTGAAGGTCACAC	64	3509	
35	15863.4	AGGTGCTTCTGCAAGCAAAGTATG	15858.8	AGAAATTTCACCACGACAGCTCAG	64	4619	
36	15859.4	AATCGTGGAGCTGTCTTTCACAAG	15854.9	ATGGTCAGTTTGCAATCAAGCAAG	64	4547	
37	15855.9	TAAGTGGCTGGAACCTAGGAAAGC	15846.8	TCCTGGAACCACAGGTCTCTACTG	64	9174	
38	15847.3	AGGGATGTTGGAGATGAGAAGGTC	15836.77	AGTAGCCCCTCAGTGAAAATGGAG	64	10569	
68	15836.9	GTTGGCAAAGCTGATATTTGAACC	15835.3	CTCCAGGCATAATCAGCACTCCTT	58	1631	
71	15835.3	AAGGAGTGCTGATTATGCCTGGAG	15835.1	TTCTCAGAACACTTCTGGATGATGC	58	191	HST+
69	15835.2	TTTAGAATACCGAAGAACCCATTT	15834.7	TTCTTTGCTGACGCACTTGAGAAC	58	505	
44	15835.0	CCTGAGCAAGACCTAACCCTGAAG	15832.9	TTCTCACCTCTCACAGGCTTATCG	64	2192	
45	15833.1	GCAACATAAGCCCACAAACTGTTC	15828.5	TTGTCGACCTCTAAACACCTCCTG	64	4559	
46	15829.5	TATTGAATGTTTAGGGGCCATCTG	15819.5	GTTCCCATCAAGTACCACAGCATC	64	9935	
47	15820.1	TGAGGAGGTTAATGAGCCACTCTG	15809.2	TGGGGATTCTTTAGAGCAGTAGCC	64	10926	

Long Range PCR primers across *ITPR1* (continued)

48	15810.0	CACAAGCACCCATGTTTAAGTTCC	15804.7	GTAACTGGGTATTCCCTGGCTTTG	64	5404	
49	15805.2	TTGTTCAGGATGGAGACTGGAGAC	15800.1	CTAATGAACTACACCAGGGCATCG	64	5055	
63	15800.3	TGGTTATCACACCCGTCCAAATAG	15798.7	ATGTCCACGAAGACTTTGATGCTC	64	1622	GC
50	15798.9	TCCATAGAGCTGCTTGAGTTCCAC	15796.3	ATGCAAGTCATAACCTTCCTGCTG	64	2541	
51	15796.8	CATGCAGGGTTTTAGATGTTCCAG	15790.7	AGCAATTTTCTGCACTGAAAGCAC	64	6162	
52	15791.4	TGAGTGCTTCTCTCAGGGAATTTG	15780.6	AATCCTCAGCCCAGACCAGTTTAG	64	10740	
64	15781.1	TGGTGGGAAATTGCAGGATAATAAC	15777.6	ACAATGTGCATGTTCTGGGTTAGG	64	3500	GC
65	15777.9	CTGCAGATGGTACCAGGCATTAAG	15773.5	TACTTAACCCTGAGCCTGCCTTTC	64	4400	GC
53	15774.0	AGAGTTTCGATTGTGATCCAGCAG	15766.8	TGAGAAGAAGTGGGTTGGTGAAAG	64	7200	GC
54	15767.2	ACACACCTTGGTAATCGACTCTGG	15758.2	TTGAGATTCTTTCCCTCACCTTCC	64	8911	
55	15758.6	AAGATGAAGAAAGTTCCCGAGGTG	15748.3	AACATTTCTGACTGCCTGTCTTGC	64	10437	
56	15750.5	GTGTATGGGTGGCAAAGAACAGAC	15742.6	GTTTGAATTTGCACCTTTGTGGAG	64	7893	
57	15743.3	TCCCTTGTCCTTAGCTTCCAACTC	15738.3	TGATATCCAGAAATGGGATTGGTG	64	5057	

ITPR1 polymorphism sequencing primers

Genomic region	Feature	Forward primer	Reverse primer	Size (bp)
15888739-15889265	SINE insertion	TAGGATTCCGGTGGAAGGTAG	CCCAGGACTGTATAGGGCTTC	533
15990820-15991263	SINE insertion	CTACCTCTCTTTTGGCATGGAT	CTTTGCAATGCTTGCTAATCAC	445
15883508-15883914	SNP	CAACCTCAGAAGGAGAAGCAG	TTCTGCAGGCTGTACGAATG	407
15930447-15931004	SNP	CACCAATTACCAGTGCCAGTT	TCAGAGCTCCAACATATGGAA	555
15901797-15902333	possible insertion	ACTGAGGACAGGGATTGGTCT	CCAGCAAAGTGCTTTACGTATG	549
15821684-15822358	SNP	GGGAGCACAAGTAGGAAGGAC	TGAAGGACAGGTCAGAAATGG	675
15919562-15920003	2 bp deletion	TGTCATGGAGAGCAAAGAAGC	TGATCTGTCCTCACTCCTGGT	442
16015255-16015700	SINE deletion	CAAAAGGCCAATGTGTGTTCT	GCCCTGCTGAGTCTTTAAGGT	425
16011285-16011680	SINE insertion	TTCTCAGTATTCATTGGGCAGA	CACTGCATTCTCCAACACTCA	407
16010817-16011268	SINE deletion	GCACCAGAGTCCCATAATCAA	GAAAAGATACACACCCACATGC	538
15807374-15807822	SNP	CAGTACTTGGCACATGGAACA	AGGAAGCTCAGCATTCTTTCC	451
16005179-16005716	9 BP deletion	GAAGGTCACATTGAGGATTCAG	TCAGGAATCCCTACAGGACAA	538
15955174-15955544	SINE insertion	AACGTTGTGGTGTACTGAATGC	TGGTACAACTCTAATGCGAGGA	376
15747434-15747873	SNP	AGCAAGGATTTGTGCTTCCTT	GATAACACCAGGCGACTTTGA	444
15952316-15952708	SNP	GTCCAAATAGCAAAGGGCAAT	CCAAGTCAAACACCCCACTTA	393

Primers for amplification across the *ITPR1* GAA triplet repeat expansion

Name	Forward primer	Reverse primer	Size (bp)
ISP_GAA_3	GGTGAGGAGCATGTTCTGGT	TGTCTCAGCGGTTGAATGTC	238

Primers for PCR amplification and exon sequencing of the *SPTBN2* gene

Exon	Forward primer	Sequence	Reverse primer	Sequence	Amplicon size (bp)
1	SPTBN2_1_F	GACGGAATTCTCCTGGTTGA	SPTBN2_1_R	GTGGGTGCCCCCTATAATG	452
(2)+3	SPTBN2_(2)+3_F	ACTTTTAGCTCCCAGCTTTGC	SPTBN2_(2)+3_R	TACTCTTGGCACTGGAATTGG	644
4	SPTBN2_4_F	GACCTCACTTCTGGGGTCTTC	SPTBN2_4_R	GGCTTGGGCAGATTAAGATTC	538
5	SPTBN2_5_F	GCTTTGGCTCTGTGAACTTTG	SPTBN2_5_R	TGGATATCACAGCAACCACAA	484
6+7	SPTBN2_6+7_F	TGACTTCTGACCCATTCTGCT	SPTBN2_6+7_R	CCAGCTCTTGTCTGGATGTTC	666
8	SPTBN2_8_F	GGGCACCACACTACATAAAGG	SPTBN2_8_R	AAAGGAAACCCCTTACCCATT	425
9+10	SPTBN2_9+10_F	ATTCGGGGTGAGTGATTCAAG	SPTBN2_9+10_R	GATTGCTTCTGGCACTTTGAG	693
11	SPTBN2_11_F	ATTTTCTGGCTTTTGCACTGA	SPTBN2_11_R	GGGGCGTGTTGTACAAGATAA	446
12	SPTBN2_12_F	TCCAGTCCTCACCTAGAAGCA	SPTBN2_12_R	AACCCTTTCTCCTGTTCAAGC	552
13	SPTBN2_13_F	GAGGCTTGAACAGGAGAAAGG	SPTBN2_13_R	GCCCTGTTTGCTCTGTGTATC	616
14	SPTBN2_14_F	GACAGAACTGGGGTGTCACTG	SPTBN2_14_R	CCCTAATGCTGTCACCTACCA	526
15.1	SPTBN2_15.1_F	TGGTAGGTGACAGCATTAGGG	SPTBN2_15.1_R	TCTGCCTGGAACTGGTAGAGA	581
15.2	SPTBN2_15.2_F	CTTGAAGCTTACGCTGGAACA	SPTBN2_15.2_R	TTGTAGCAAGAGCAGGCAGTT	685
16	SPTBN2_16_F	CTCCCCACTTGACCCATAGAT	SPTBN2_16_R	GCTCATTCTCCCTTGACTCCT	518
17.1	SPTBN2_17.1_F	AGGAGTCAAGGGAGAATGAGC	SPTBN2_17.1_R	ATCTAAGCTGCGCAGGAAGTC	576
17.2	SPTBN2_17.2_F	GGCTTGGAGAGGTACAAGCTG	SPTBN2_17.2_R	TATGTTCTCCCAGTCCACCAG	600
18	SPTBN2_18_F	GTTTGGGCTGTTAAGGTCCTC	SPTBN2_18_R	GCACTATCCAGCTGCTTCATC	499
19	SPTBN2_19_F	GTGGTCAGAATGGGACTAAGC	SPTBN2_19_R	GCATATGCAGTGAAGCATAAGG	460
20	SPTBN2_20_F	GCTCTTGGTTCCTCCATGTC	SPTBN2_20_R	GGAGGTTTGGCACTAAAGACC	513
21	SPTBN2_21_F	GGTCTCTGTGGCTTCTTTGTG	SPTBN2_21_R	GACGTGGCCTTGAGAAATACC	658
22	SPTBN2_22_F	ACTGAGGTAGGCCATGAGGAT	SPTBN2_22_R	ACTTTCAGTGCCAAAGGGAAG	503
23	SPTBN2_23_F	TTCCCTTTGGCACTGAAAGTA	SPTBN2_23_R	CCTAAGGTTGTTAGGCCTTCG	499
24	SPTBN2_24_F	GTGGGTTTCCATTGTAGCAGA	SPTBN2_24_R	TGCTTCTTTACCTCTGCTTGG	574
25	SPTBN2_25_F	TCCAAGAGATGTGGAAACACC	SPTBN2_25_R	TGTACCAAAGGCTCCACAGTC	471
26	SPTBN2_26_F	AGCCAAGACATGATTGACCAC	SPTBN2_26_R	TGACCTCCTTTCTGCACAGTT	627
27	SPTBN2_27_F	TGCTCTGCAGAAGAAACCTGT	SPTBN2_27_R	GCATCCTCCTAGCATTTCTGA	617
28	SPTBN2_28_F	CTCCTGCTTCTCTGATGGATG	SPTBN2_28_R	GGCCTCTATCTCTGCCTTGAT	529
29	SPTBN2_29_F	TACTGGACACCACGGACAAGT	SPTBN2_29_R	GGCAGAGACGTGAGTTAGCAC	578
30	SPTBN2_30_F	TCCAACTCTTCTCCAGAACCA	SPTBN2_30_R	CAGAGAAATGGCAGGTCAGAG	461
31	SPTBN2_31_F	GGCAGTAGCTTCCTCCATGTC	SPTBN2_31_R	TCCTGAAGCTGCAGTGACAAT	450
32+33	SPTBN2_32+33_F	TACCCATTGTCACTGCAGCTT	SPTBN2_32+33_R	GCTGCTCTGCTCAAGTCTCTG	657
34	SPTBN2_34_F	AGGAGGAAGAACGGAGGAAAC	SPTBN2_34_R	GGGAAAGTGAGGAATGGTGTC	624
35	SPTBN2_35_F	CCCAGAGGTGTGGGTGTTAAT	SPTBN2_35_R	CCAGAGGACCCTCCTCCTTAC	543
36	SPTBN2_36_F	CGAGCTCAGAAACAGGAACAG	SPTBN2_36_R	TGGATGTTGTCAGTGGTCAGG	501
37	SPTBN2_37_F	ACAGGAAGGGTTTGGGAATCT	SPTBN2_37_R	GAAGCTGGAAAGCCTACGAGA	477
38	SPTBN2_38_F	CCAGTGAGGTGAAGTAGTCCAG	SPTBN2_38_R	AAAGAAGTTGGGAGATGGGAGT	607

qPCR primers and probes

Primers used for qPCR assays of gene expression

All probes were 5' 6-FAM and 3' Iowa Black labelled, with internal ZEN labelling.

Assay name	Forward primer sequence	Probe sequence	Reverse primer sequence	Amplicon size (bp)
SPTBN2	TGGATGGTGAAGAGCAGAAC	TTCTCAGTGAACTTGGGTGGCTTCTC	TCCTTCAATTCCTATCGCACG	83
ACTB	CCAACCGTGAGAAGATGACC	CGAGACTTTCAACACCCCAGCCA	CGTACAGGGACAGCACAG	90
TBP	TCTGGCATATTTCCTCGCTG	ACTGTTCTTCACTCTTGGCTCCCG	TTCAGTTCTGGGAAGATGGTG	78
BCAN	CACCGTGTCTACTTCATAGCG	CACAGGACACCAGCCCCATCTT	GAGTGATGTACCCTGCAACTAC	133
FAM83H	CCACGTGAAGGAGAAGTTTCTG	TGTAGCTCCCGCTCATCACCAC	TTCTCGAAGGACCACATGAAG	81

Diagnostic testing primers

Primers for assaying the mutations causing episodic falling and congenital keratoconjunctivitis sicca and ichthyosiform dermatosis in the Cavalier King Charles Spaniel

Forward primers		Reverse primers		Expected product size (bp)	
Name	**Sequence**	**Name**	**Sequence**	**Wild-type**	**Mutant**
CKCSID_F	6Fam-CTTACACCCTGGCCCCGTA	CKCSID_R	GGTCGAGGAAAGAGGGGAAT	135	134
EF_F	6Fam-TGTGCTCAGGAGTCTGTCCA	EF_bridge_R	GCAACCTACTCAGACCTGGAA	15,831	115
EF_F	6Fam-TGTGCTCAGGAGTCTGTCCA	EF_normal_R	GGAGCAAATGCTCTGGAAGG	105	N/A

Primers for assaying two microsatellites markers linked to spinocerebellar ataxia in the IS

Forward name	Forward sequence	Reverse name	Reverse sequence	Amplicon size (bp)
Micro 1 F	CTGGCCCATTAGTGCATAGTC	Micro 1 R	GGTTCACTCTAGTCATTGTTTGC	188
Micro 2 F	TCTCCATCTTCATTTCAGCATT	Micro 2 R	TGCAGTATTTGTCTTTCCCTGT	230

TaqMan assay for the *CAPN1* non-synonymous SNP associated with spinocerebellar ataxia (late onset ataxia) in the Parson Russell Terrier. Reporter 1 was a major groove binding probe with 5' VIC and 3'NFQ labelling. Reporter 2 was a major groove binding probe with 5' FAM and 3'NFQ labelling. (VPS51 and Intergenic SNP assays designs also shown)

Assay	Forward primer	Reverse primer	Reporter 1	Reporter 2
CAPN1	GGGTGAGGGAGGCAATGG	GCAGAGCAGGCCTGATATGG	AGAAGCCAACAGTCCC	AGAAGCCAATAGTCCC
VPS51	CCTTCCGCCGACGCT	GCAGCCGCTGGACCT	AGGCGGAGGCCC	AGGCGAAGGCCC
PRT_54.15	GTCAGCATTTCCTTGCATC TTGTTA	GAGGCTGCCATAGAAAGATTT ACCT	TAGAGTTGGACTTTTTC	AGTTGGGCTTTTTC

Primers for assaying the mutation causing neonatal cerebellar cortical degeneration in the Beagle

Forward name	Forward sequence	Reverse name	Reverse sequence	Amplicon size (bp)
SPTBN2_29F	TACTGGACACCACGGACAAGT	SPTBN2_28R	GGCCTCTATCTCTGCCTTGAT	268

Appendix 3 ABI3130xl genetics analyser running parameters

Standard sequencing parameters on ABI3130xl genetics analysers

Parameter	Setting
Oven temperature (°C)	60
Polymer fill volume (nl)	4840
Current stability	5.0
Pre-run voltage (V)	15.0
Pre-run time (s)	180
Injection voltage (V)	1.2
Injection time (s)	5
Voltage number of steps	20
Voltage step interval (s)	15
Data delay time (s)	80
Run voltage (s)	13.2
Run time (s)	1200

Standard fragment analysis settings for genotyping on ABI3130xl genetics analysers

Parameter	Setting
Oven temperature (°C)	60
Polymer fill volume (nl)	6500
Current stability	5.0
Pre-run voltage (V)	15.0
Pre-run time (s)	180
Injection voltage (V)	3.0
Injection time (s)	10
Voltage number of steps	20
Voltage step interval (s)	15
Data delay time (s)	60
Run voltage (s)	15.0
Run time (s)	1000

Appendix 4 Non-standard PCR cycling parameters

Microsatellite amplification using tailed primers (Section 2.14.)

Initial denaturation

94℃ for 4 minutes

30 cycles of:

94℃ for 1 minute

57℃ for 1 minute

72℃ for 1 minute

8 cycles of:

94℃ for 1 minute

57℃ for 1 minute

72℃ for 1 minute

Final elongation

72℃ for 30 minutes

Hold at 4℃

Amplification of fluorescently labelled primers (Section 2.15.1.)

Initial denaturation

95℃ for 10 minutes

30 cycles of:

95℃ for 30 seconds

57℃ for 30 seconds

72℃ for 30 seconds

Final elongation

72℃ for 10 minutes

Hold at 4℃

Long range PCR (Section 2.16.1)

Initial denaturation

98℃ for 10 minutes

35 cycles of:

98℃ for 8 seconds

57℃ for 15 seconds

72℃ for 6 minutes

Hold at 4℃

Amplification of adapter ligated library fragments (Section 2.16.7)

Initial denaturation

98℃ for 30 seconds

18 cycles of:

98℃ for 30 seconds

65℃ for 30 seconds

72℃ for 30 seconds

Final elongation

72℃ for 5 minutes

Hold at 4℃

SureSelect pre-hybridisation amplification (Section 2.16.8.3)

Initial denaturation

98℃ for 10 seconds

9 cycles of:

98℃ for 10 seconds

65℃ for 15 seconds

72℃ for 15 seconds

Final elongation

72℃ for 5 minutes

Hold at 4℃

SureSelect post-capture amplification (Section 2.16.8.8)

Initial denaturation

98℃ for 10 seconds

13 cycles of:

98℃ for 10 seconds

57℃ for 30 seconds

72℃ for 30 seconds

Final elongation

72℃ for 5 minutes

Hold at 4℃

Quantification of sequencing libraries (section 2.16.10.)

Initial denaturation

95℃ for 3 minutes

35 cycles of:

95℃ for 30 seconds

60℃ for 45 seconds

Quantitative PCR (section 2.17.)

Initial denaturation

95℃ for 3 minutes

40 cycles of:

95℃ for 10 seconds

60℃ for 30 seconds

IS diagnostic DNA test (Section 2.19.1)

Initial denaturation

98℃ for 30 seconds

36 cycles of:

98℃ for 8 seconds

55℃ for 15 seconds

72℃ for 30 seconds

Final elongation

72℃ for 1 minute

Hold at 4℃

PRT diagnostic DNA test (Section 2.19.4)

Pre-PCR read

25℃ for 30 seconds

Initial denaturation

95℃ for 30 seconds

40 cycles of:

95℃ for 3 seconds

60℃ for 10 seconds

Post-PCR read

25℃ for 30 seconds

Amplification across the GAA triplet repeat region (Section 2.19.5)

Initial denaturation

93℃ for 3 minutes

35 cycles of:

93℃ for 15 seconds

60℃ for 30 seconds

68℃ for 3 minutes

Hold at 4℃

Appendix 5 Key commands of the NGS analysis Perl script

1 **Index the reference sequence (FASTA format)**

```
> bwa index -a is ref.fasta
```

2 **Finds suffix array (SA) coordinates of good hits for each read**

```
> bwa aln ref.fasta reads1.fastq > aln_sa1.sai
> bwa aln ref.fasta reads2.fastq > aln_sa2.sai
```

3 **Convert SA coordinates into chromosomal coordinates (PE data)**

```
bwa sampe ref.fasta aln_sa1.sai aln_sa2.sai reads1.fastq reads2.fastq > aligned.sam
```

4 **Create the fasta sequence dictionary file**

```
> java16 -jar /opt/picard/CreateSequenceDictionary.jar R=ref.fasta O=ref.dict
```

5 **Create the fasta index file**

```
> samtools faidx ref.fasta
```

6 **Sort the SAM file generated by BWA**

```
> java16 -Xmx2g -jar /opt/picard/SortSam.jar I=aligned.sam O=aligned_sorted.sam
SO=coordinate
```

7 **Convert the SAM file from BWA to a BAM file**

```
> java16 -Xmx2g -jar /opt/picard/SamFormatConverter.jar I=aligned_sorted.sam
O=aligned_sorted.bam
```

8 **Duplicate removal**

```
> samtools rmdup aligned_sorted.bam aligned_sorted2.bam
```

9 **Create an index for the BAM file**

```
> java16 -Xmx2g -jar /opt/picard/BuildBamIndex.jar I=aligned_sorted2.bam
```

10 **Count covariates**

```
> java16 -Xmx2g -jar /opt/gatk/GenomeAnalysisTK.jar -R ref.fasta -I
aligned_sorted2.bam -T CountCovariates -cov ReadGroupCovariate -cov
QualityScoreCovariate -cov CycleCovariate -cov DinucCovariate -recalFile
reads.csv --default_platform illumina
```

11 **Table recalibration**

```
> java16 -Xmx2g -jar /opt/gatk/GenomeAnalysisTK.jar -R ref.fasta -I
aligned_sorted2.bam -T TableRecalibration -outputBam recal.bam -recalFile
reads.csv -- default_platform Illumina
```

12 **Index the new BAM file**

```
> samtools index recal.bam
```

13 **Creating intervals**

```
> java16 -Xmx2g -jar /opt/gatk/GenomeAnalysisTK.jar -T RealignerTargetCreator -I
recal.bam -R $ref -o forRealigner.intervals
```

14 Realigning

```
> java16 -Xmx2g -jar /opt/gatk/GenomeAnalysisTK.jar -I recal.bam -R ref.fasta - T IndelRealigner -targetIntervals forRealigner.intervals --output cleaned.bam
```

15 Sort the cleaned BAM file

```
> java16 -Xmx2g -jar /opt/picard/SortSam.jar I=cleaned.bam O=cleaned_sorted.bam SO=coordinate VALIDATION_STRINGENCY=LENIENT
```

16 Create an index for the cleaned and sorted BAM file

```
> java16 -Xmx2g -jar /opt/picard/BuildBamIndex.jar I=cleaned_sorted.bam VALIDATION_STRINGENCY=LENIENT
```

17 IndelGenotyper

```
> java16 -Xmx2g -jar /opt/gatk/GenomeAnalysisTK.jar -T IndelGenotyperV2 -R ref.fasta -I cleaned_sorted.bam -O indels.raw.bed -o detailed.output.bed --verbose
```

18 Make SNP calls

```
> java16 $mem -jar /opt/gatk/GenomeAnalysisTK.jar -R $ref -T UnifiedGenotyper -I cleaned_sorted.bam -stand_emit_conf 10.0 -varout snps.raw.vcf -stand_call_conf 50.0 --platform SOLEXA
```

19 Pindel (Structural Variant analysis)

```
> java16 -Xmx2g -jar /opt/picard/SortSam.jar I=cleaned_sorted.bam O=queryname_sorted.bam SO=queryname VALIDATION_STRINGENCY=LENIENT

> perl /opt/pindel/bam2pindel.pl -i queryname_sorted.bam -o sv -s $name -pi (insert size)

> cat sv_chr1.txt sv_chr2.txt sv_chr3.txt sv_chr4.txt sv_chr5.txt sv_chr6.txt sv_chr7.txt sv_chr8.txt sv_chr9.txt sv_chr10.txt sv_chr11.txt sv_chr12.txt sv_chr13.txt sv_chr14.txt sv_chr15.txt sv_chr16.txt sv_chr17.txt sv_chr18.txt sv_chr19.txt sv_chr20.txt sv_chr21.txt sv_chr22.txt sv_chr23.txt sv_chr24.txt sv_chr25.txt sv_chr26.txt sv_chr27.txt sv_chr28.txt sv_chr29.txt sv_chr30.txt sv_chr31.txt sv_chr32.txt sv_chr33.txt sv_chr34.txt sv_chr35.txt sv_chr36.txt sv_chr37.txt sv_chr38.txt sv_chrX.txt sv_chrM.txt sv_chrUn.txt>all.txt

> pindel022 -f $ref -p all.txt -c ALL -o pindel

> pindel2vcf -p pindel_BP -r $ref -R canfam2 -d 2006
> pindel2vcf -p pindel_D -r $ref -R canfam2 -d 2006
> pindel2vcf -p pindel_INV -r $ref -R canfam2 -d 2006
> pindel2vcf -p pindel_LI -r $ref -R canfam2 -d 2006
> pindel2vcf -p pindel_SI -r $ref -R canfam2 -d 2006
> pindel2vcf -p pindel_TD -r $ref -R canfam2 -d 2006
```

20 Create GC bias metric histogram

```
> java16 -Xmx2g -jar /opt/picard/CollectGcBiasMetrics.jar REFERENCE_SEQUENCE=ref.fasta
  INPUT=cleaned_sorted.bam O=out1.junk CHART_OUTPUT=GC_bias.pdf
  VALIDATION_STRINGENCY=LENIENT
```

21 Create alignment summary metrics

```
> java16 -Xmx2g -jar /opt/picard/CollectAlignmentSummaryMetrics.jar
  I=aligned_sorted.bam O=Alignment_Summary.xls R=ref.fasta
  VALIDATION_STRINGENCY=LENIENT
```

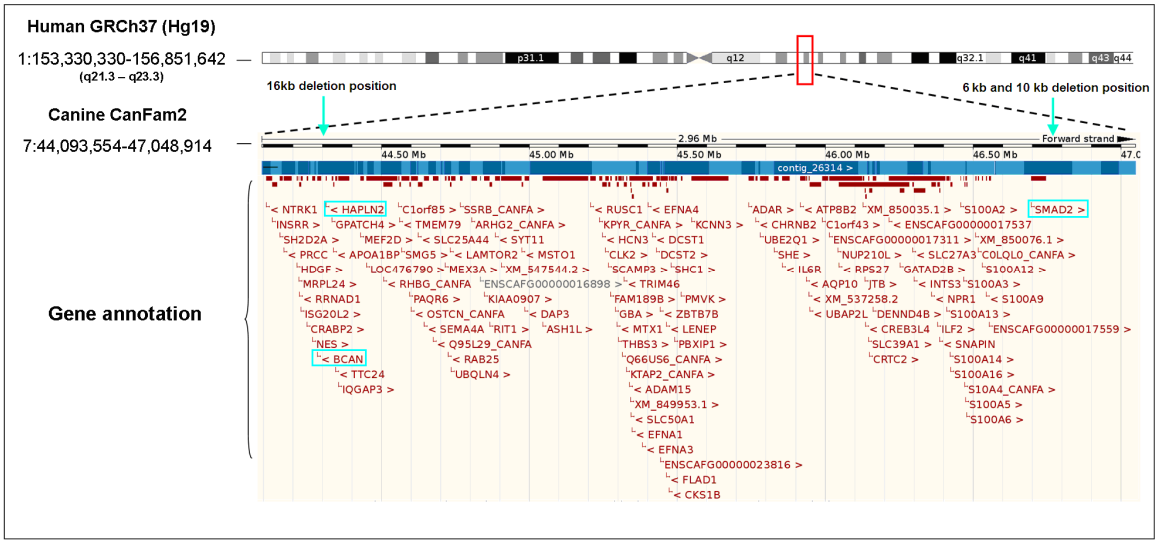
22 Create insert size histogram

```
> java16 -Xmx2g -jar /opt/picard/CollectInsertSizeMetrics.jar INPUT=aligned_sorted.bam
  O=out2.junk MINIMUM_PCT=0.05 HISTOGRAM_FILE=INSERT_SIZE.pdf
  VALIDATION_STRINGENCY=LENIENT
```

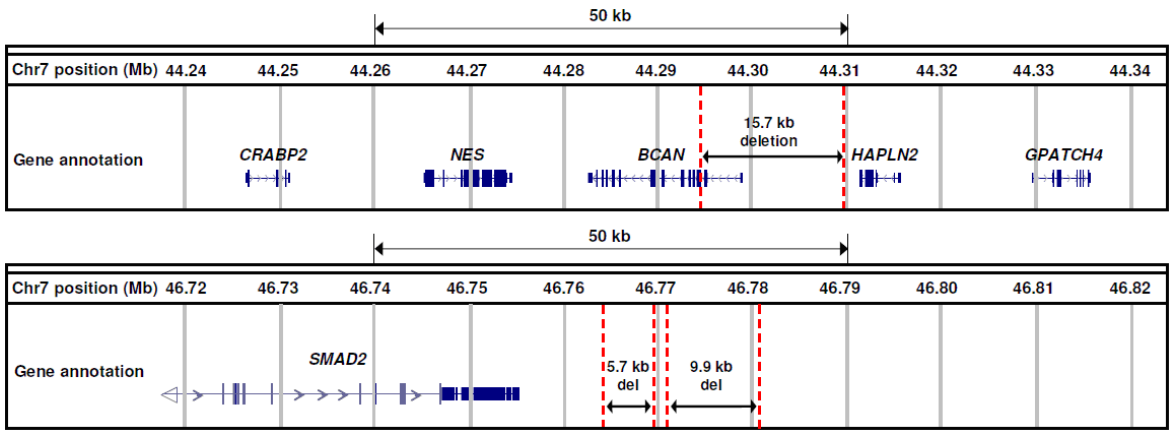
Appendix 6 CKCS critical regions, features and human synteny

EF disease-associated interval

Deletion positions are indicated by the blue arrows, with genes in close proximity highlighted with a blue box.

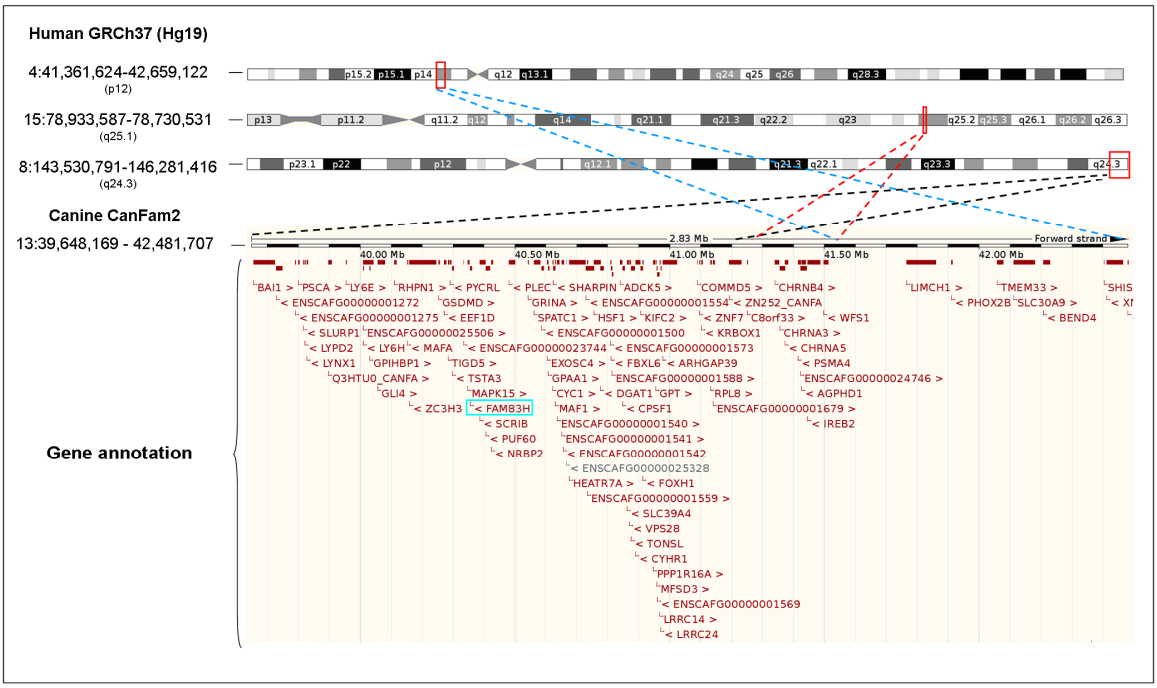


Schematic overview of the genes in close proximity to the three identified deletions.



CKCSID disease-associated interval

FAM83H is highlighted with a blue box.

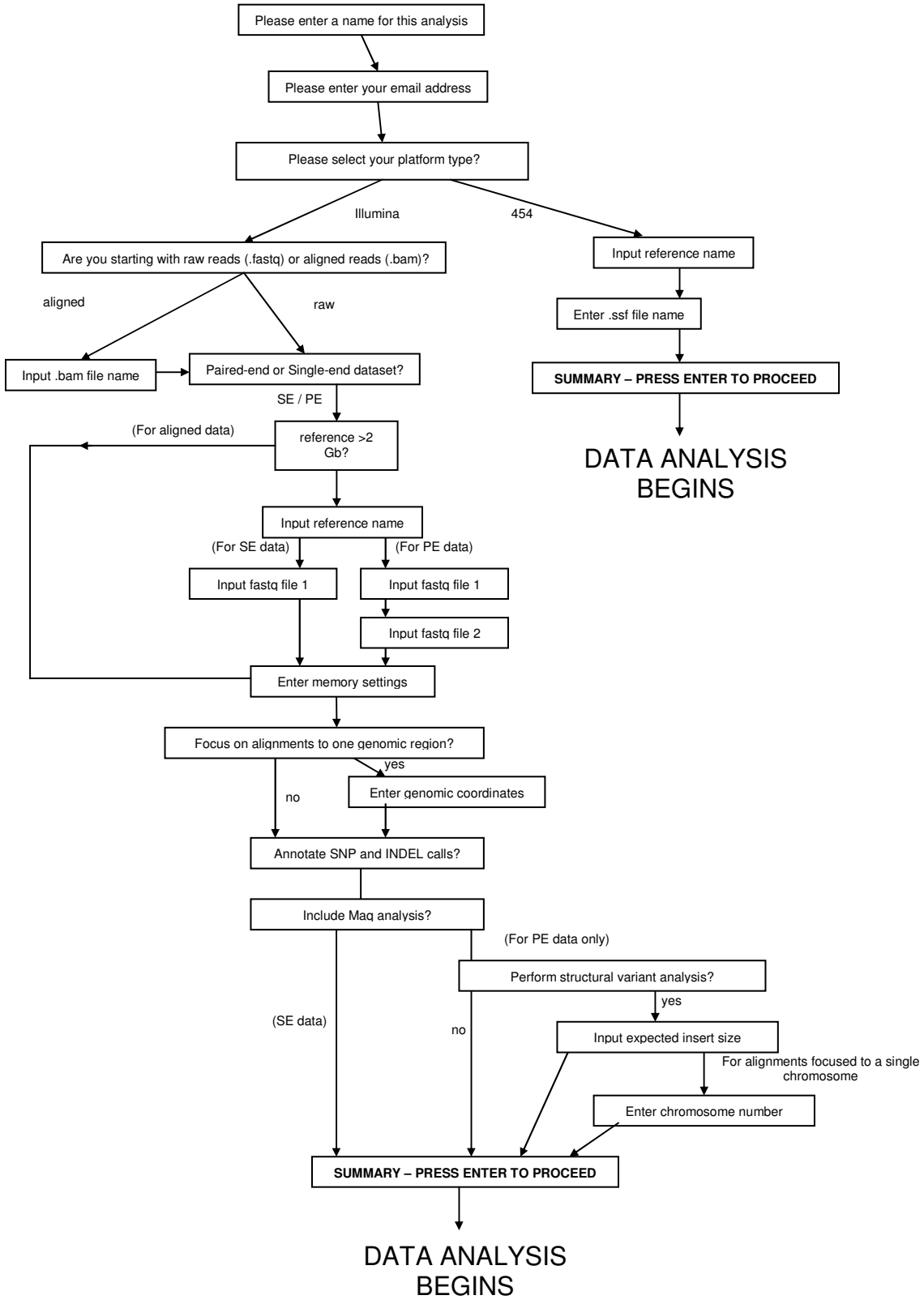


Appendix 7 Results files from the NGS analysis pipeline

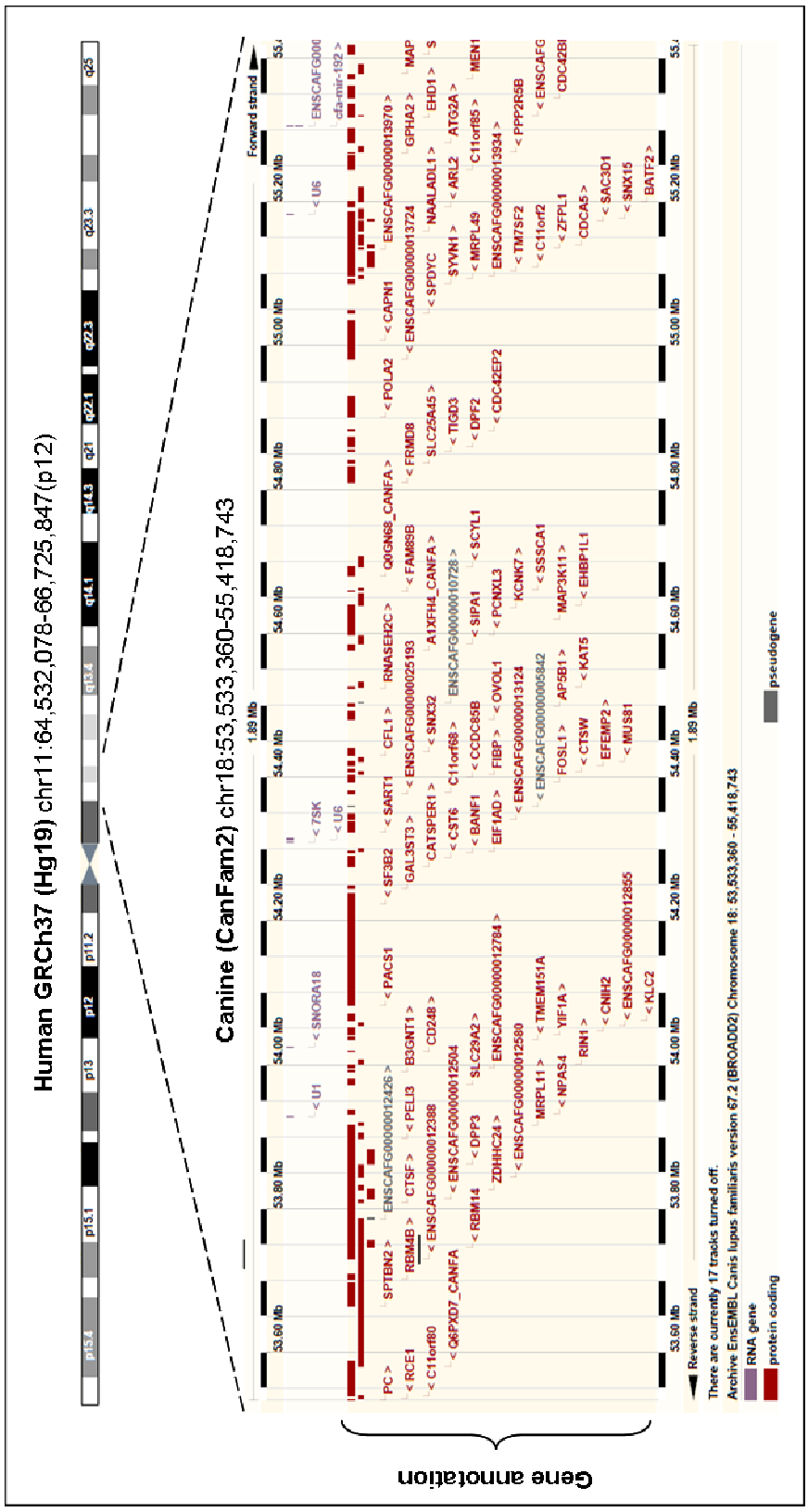
The following information is displayed in a readme.txt file included in the results folder created on completion of the NGS script.

```
========================
Summary of results files
========================

PDF FILES

GC_ Bias.pfd   Histogram of GC content of aligned reads
INSERT_SIZE.pdf        Insert size histogram

ALIGNMENT FILES

raw_align.bam  Raw alignments to the reference
best_align.bam Best alignments to the reference after processing by GATK

SNP AND INDEL CALLS

INDELS.bed             List of Indels
SNPS.vcf               List of SNPs
annotated_indels.xls   List of Indels annotated using the Ensembl database
annotated_snp.xls      List of SNPs annotated using the Ensembl database

STRUCTURAL VARIANT FILES

SV_Break_points.txt
SV_Deletions.txt
SV_Inversions.txt
SV_Long_insertions.txt
SV_Non_template_seq_in_deleletion.txt
SV_Tandom_Dup.txt

INFORMATION FILES

Depth_summary          Summary of reads depth across the target region
Log.rtf                Run log for the NGS pipeline
Duplicate_info.rtf     fraction of PCR duplicate reads in the dataset
Alignment_Summary.xls  Alignment summary information

NOTES

.bam files must have an associated .bai index file for loading into IGV
Alignment_Summary.xls can be used to calculate success of target enrichment
```

Appendix 8 NGS analysis user input workflow

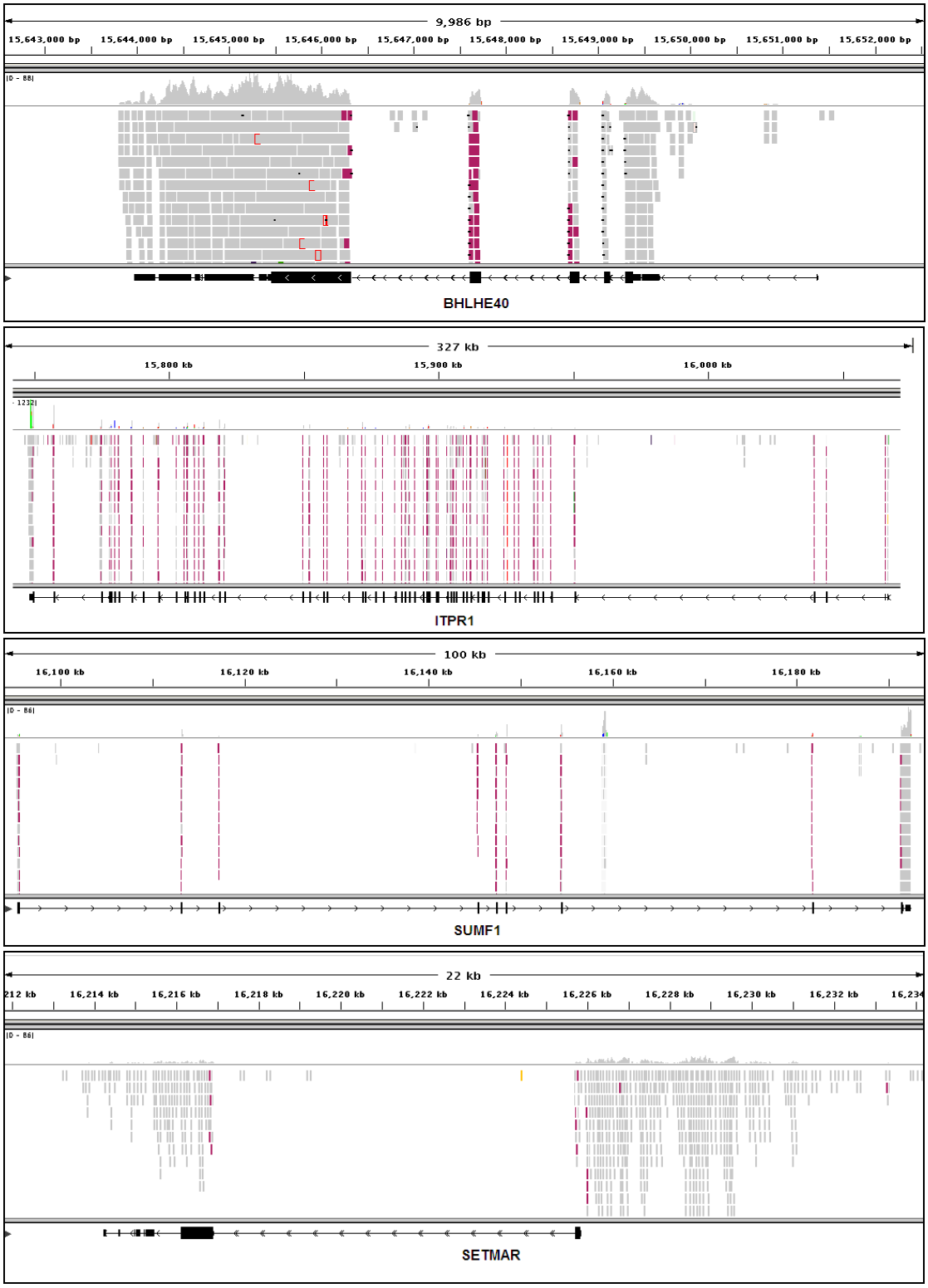


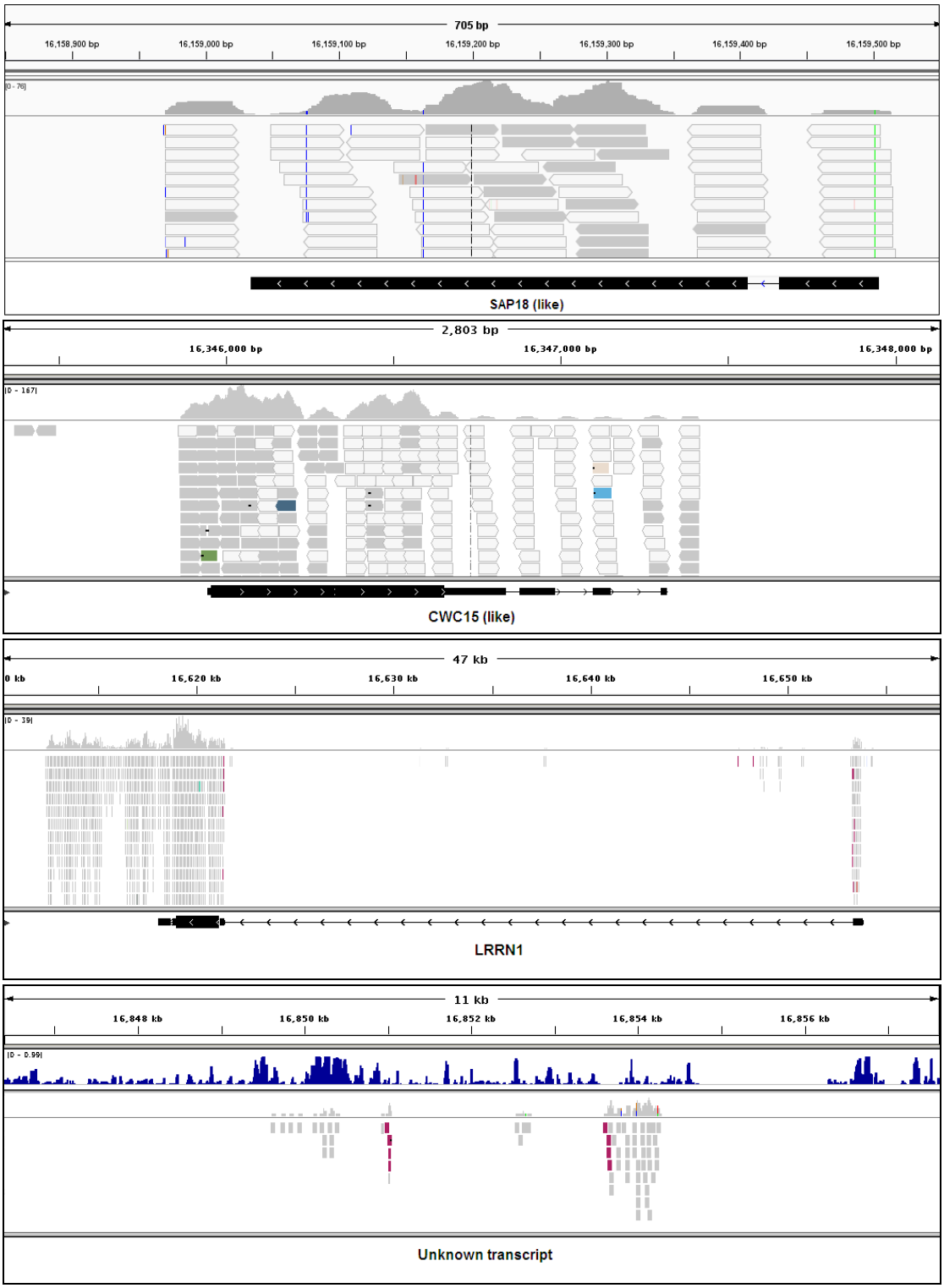
Appendix 9 Genes in the LOA interval and human syntenic region



Appendix 10 Expressed genes across the SCA critical region

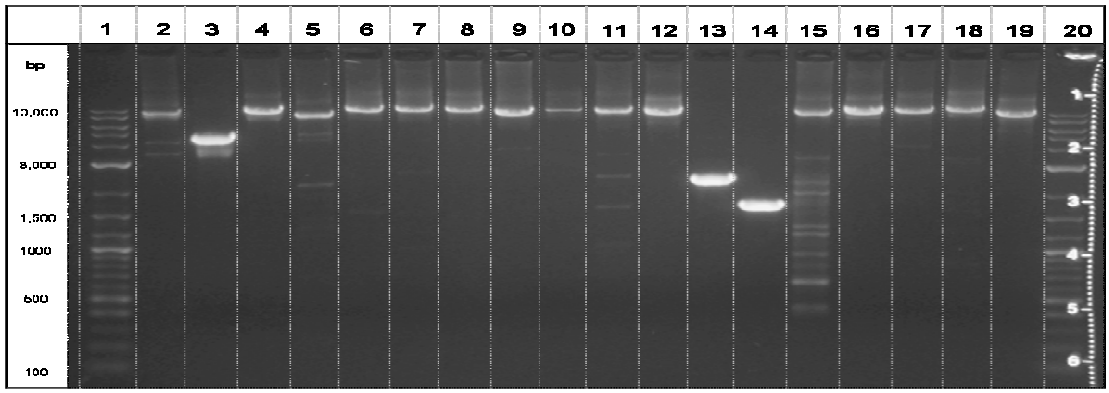
Reads are shown in grey. Read pairs spanning exons are shown in red. A histogram of read depth is shown in grey for all genes. For the unknown transcript a conservation histogram is shown in blue.

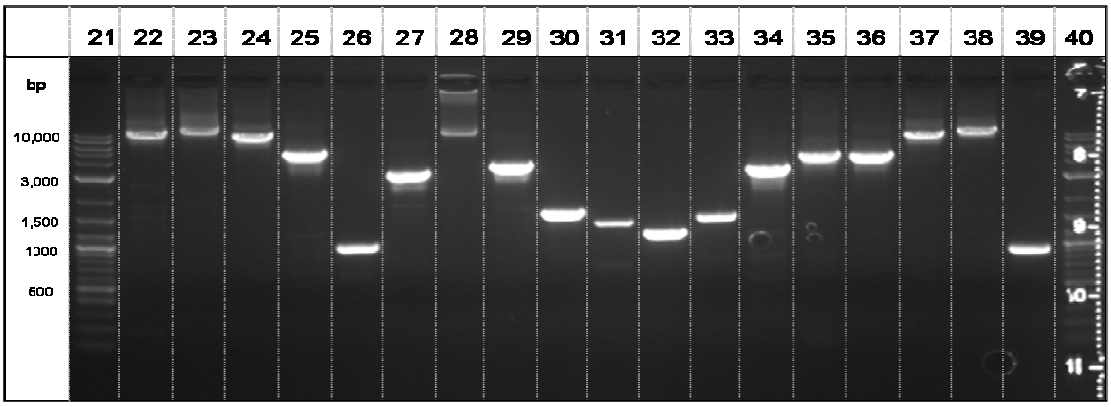


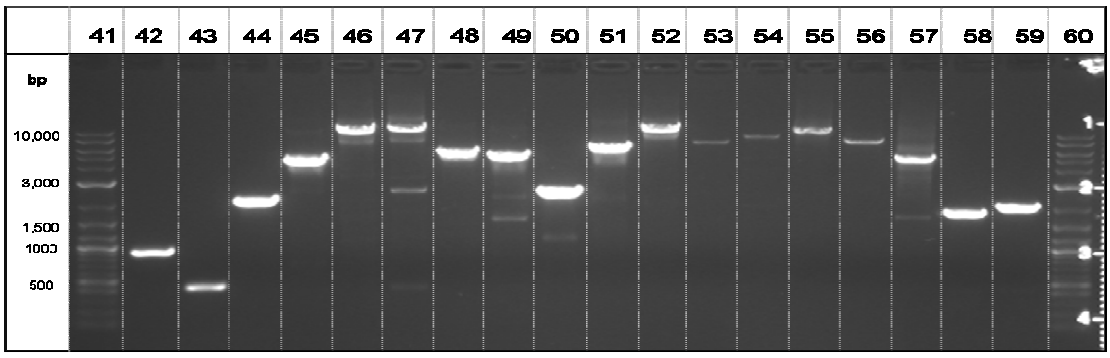

705 bp
16,158,900 bp
16,159,000 bp
16,159,100 bp
16,159,200 bp
16,159,300 bp
16,159,400 bp
16,159,500 bp
SAP18 (like)
2,803 bp
16,346,000 bp
16,347,000 bp
16,348,000 bp
CWC15 (like)
47 kb
0 kb
16,620 kb
16,630 kb
16,640 kb
16,650 kb
LRRN1
11 kb
16,848 kb
16,850 kb
16,852 kb
16,854 kb
16,856 kb
Unknown transcript

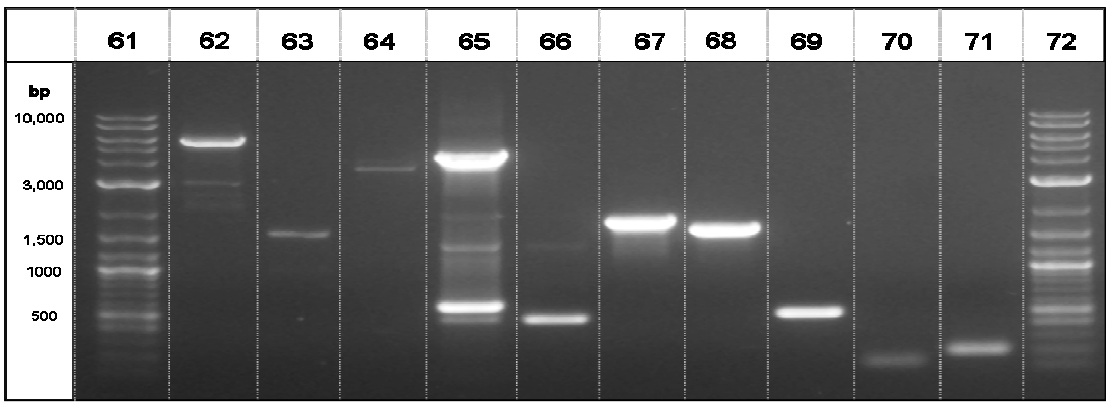
Appendix 11 Long range PCR products spanning *ITPR1*

Primer details, expected product sizes and lane contents are listed in Appendix 2.









Appendix 12 GAA repeat number calculations

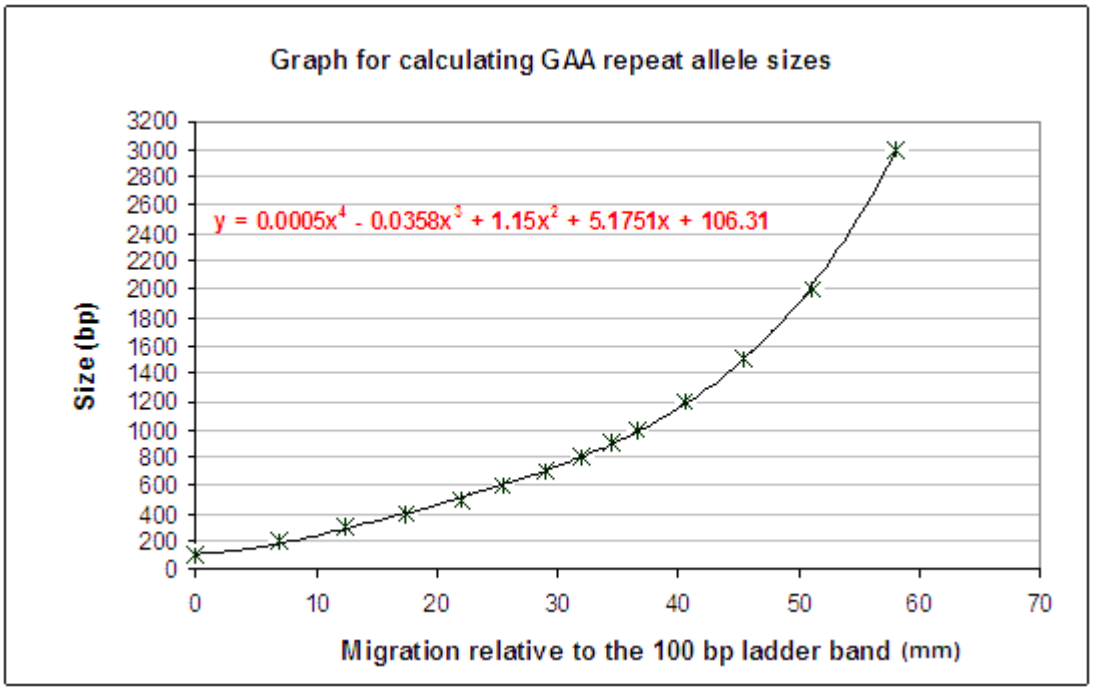
Migration distance measurements relative to the 100 bp ladder band.

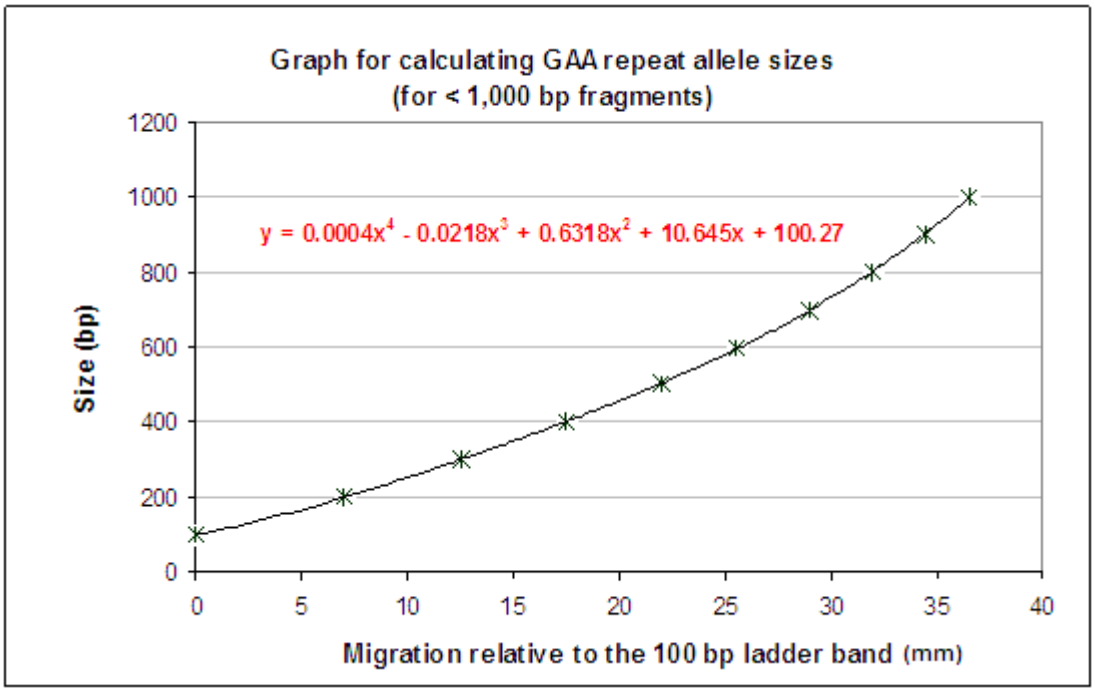
Band size (bp)	Relative migration (mm)
100	0
200	7
300	12.5
400	17.5
500	22
600	25.5
700	29
800	32
900	34.5
1000	36.5
1200	40.5
1500	45.5
2000	51
3000	58

Estimates of GAA repeat copy number based on PCR product migration.

Gel lane	ID	Allele	Migration (mm)	Calculated size (bp)	GAA copy number estimate
2	5357	1	45	1456	414
		2	48	1699	495
3	5397	1	47	1613	466
		2	49.5	1840	542
4	5404	1	46.5	1572	453
		2	48.5	1745	510
5	5405	1	9	234	7
		2	47.5	1655	480
6	5407	1	9	234	7
		2	47.5	1655	480
7	5436	1	9.5	243	10
		2	48	1699	495
8	PRT	1	9.5	243	10
		2	10.5	261	16
9	NTC	n/a	n/a	n/a	n/a
		n/a	n/a	n/a	n/a
10	6422	1	42.5	1286	357
		2	50.5	1942	576
11	6477	1	42	1256	347
		2	50	1890	559
12	6478	1	9.5	243	10
		2	52.5	2166	651
13	6479	1	11.5	280	22
		2	44.5	1420	402
14	6685	1	40.5	1169	318
		2	49.5	1840	542
15	8636	1	9.5	243	10
		2	49.5	1840	542
16	8637	1	45.5	1493	426
		2	49	1792	526

Graphs for production of a regression line equation for estimating unknown band sizes on agarose gel.





Please note the independent variable has been plotted on the y axis for optimal regression line calculation and equation purposes.

Appendix 13 mRNA-seq library fragment cloning results

150 bp insert library

Clone number	Target insert size (bp)	Index	Insert size (bp)	Gene	Region
1	150	2	140	*TTC7B*	Intronic
2	150	7		*SYT1*	3' UTR
5	150	4	184	*VDAC2*	3' UTR
11	150	5	79	*CACNA1D*	Exonic
13	150	5	136	*KPNA4*	3' UTR
15	150	2	121	-	Intergenic
16	150	7	120	*STAU1*	3' UTR
19	150	6	139	*IL16*	Exonic
21	150	7	215	*MEGF8*	3' UTR
24	150	7	155	*RABGAP1L*	3' UTR
26	150	5	162	*SFRS18*	3' UTR
30	150	4	149	*MYO18A*	Exonic
32	150	4	198	*THEM55A*	Exonic
35	150	5	141	*MTMR9*	3' UTR
36	150	?	140	*PSD3*	3' UTR
39	150	4	154	*SEPHS1*	Exonic
40	150	5	166	*ARL6IP1*	3' UTR
41	150	2	162	*RNF152*	3' UTR
43	150	4	129	*SLC7A1*	3' UTR
45	150	?	166	*RPL41*	Exonic
			150		

250 bp insert library

Clone number	Target insert size (bp)	Index	Insert size (bp)	Gene	Region
49	250	5	229	*TET3*	3' UTR
51	250	?	239	*FAM13b*	3' UTR
52	250	5	86	*FIBCD1*	Exonic 3' end
55	250	?	237	*ssx21p*	Exonic
57	250	4	260	*TPD52L2*	3' UTR
58	250	6	274	*SET*	Exon/ 3' UTR
62	250	4	258	*UBC*	Exonic
63	250	5	239	*FAM13B*	3' UTR
64	250	?	290	*SNAP25*	3' UTR
65	250	4	303	*IL16*	intronic
66	250	4	255	*REEP5*	3' UTR
67	250	4	252	*CMPK1*	3' UTR
68	250	5	269	*RMND1*	Exon/ 3' UTR
71	250	5	238	*MTCH1*	3' UTR
82	250	5	231	*ECD*	3' UTR
83	250	5	286	*(COX3)*	unknown
74	250	4	265	*ACVR1B*	3' UTR
75	250	4	297	*YEPL4*	3' UTR
78	250	5	268	*CHD9*	3' UTR
79	250	5	230	*(COX3)*	unknown
82	250	5	220	*PTTG1IP*	3' UTR
84	250	4	304	*unknown*	unknown
			251		

Appendix 14 Ataxia candidate genes

Summary of candidate genes investigated for sequence polymorphisms after mRNA-seq of a single Beagle NCCD case.

Ataxia type	Associated gene	Human Loci	GRCh37/hg19 region	CanFam2 syntenic region	Notes
SCA1	*ATXN1*	6p22.3	chr6:16,299,344 -16,761,721	chr35:18,460,465-18,865,962	Heterozygous SNPs identified - gene excluded
SCA2	*ATXN2*	12q24.12	chr12:111,890,019-112,037,480	chr26:12,091,459-12,206,061	No polymorphisms identified
SCA3	*ATXN3*	14q32.12	chr14:92,524,897-92,572,965	chr8:4,316,475-4,352,391	No polymorphisms identified
SCA4	-	16q22.1	chr16:66,700,000-70,800,000	chr5:79,600,000-85,600,000	n/a
SCA5	*SPTBN2*	11q13.2	chr11:66,452,720-66,488,870	chr18:53,664,195-53,696,100	8 bp deletion exon 29. Frameshift. Prediction: 27 Aberrant amino acids, 410 amino acid truncation (p.I1953Rfs*28).
SCA6	*CACNA1A*	19p13.2	chr19:13,317,256-13,617,274	chr20:51,822,254-52,037,590	Heterozygous SNPs identified - gene excluded
SCA7	*ATXN7*	3p14.1	chr3:63,850,233 -63,982,293	chr20:30,222,502-30,361,703	No polymorphisms identified
SCA8	*ATXN8OS*	13q21.33	chr13:70,681,345-70,713,885	chr22:27,500,593-27,536,451	Insufficient read depth to analyse
SCA9	-	-	-	-	n/a
SCA10	*ATXN10*	22q13.31	chr22:46,067,678-46,241,187	chr10:23,331,497-23,495,457	No polymorphisms identified
SCA11	*TTBK2*	15q15.2	chr15:43,036,542-43,213,007	chr30:12,667,929-12,836,871	Heterozygous SNPs identified - gene excluded
SCA12	*PPP2R2B*	5q32	chr5:145,969,068-146,461,033	chr2:43,854,239-44,295,609	No polymorphisms identified
SCA13	*KCNC3*	19q13.33	chr19:50,818,765-50,832,634	-	No canine orthologue
SCA14	*PRKCG*	19q13.42	chr19:54,385,467-54,410,901	chr1:106,377,729-106,388,206	No polymorphisms identified
SCA15	*ITPR1*	3p26.1	chr3:4,535,032-4,889,524	chr20:15,748,595-16,067,260	4 Exonic SNPs. 1 Non-synonymous SNP (p.E2491Q). Q residue conserved across mammalian species
SCA16 (see SCA15)	-	-	-	-	n/a
SCA17	*TBP*	6q27	chr6:170,863,471-170,881,946	chr12:75,485,739-75,496,000	No polymorphisms identified
SCA18	*IFRD1**	7q22-q32	chr7:112,063,199-112,117,258	chr14:59,850,326-59,898,828	No coding polymorphism identified
SCA19	-	1p21-q21	chr1:94,700,000 - 155,000,000	multiple chromosomes	n/a
SCA20	12 gene duplication	11q12	chr11:61,453,940-61,746,519	chr18:57,759,467-57,508,030	n/a
SCA21	-	7p21.3-p15.1	ch7:10,075,254-28,198,059	multiple chromosomes	n/a
SCA22	-	1p21-q21	chr1:94,700,000 - 155,000,000	multiple chromosomes	n/a
SCA23	*PDYN*	20p13	chr20:1,959,402 -1,974,931	chr24:22,050,982-22,054,521	Insufficient read depth to analyse

Ataxia candidate genes continued

SCA24 (see SCAR4)	-	-	-	-	n/a
SCA25	-	2p21-p13	chr2:41,800,000 -75,000,000	chr10,chr17	n/a
SCA26	-	19p13.3	chr19:998,644-4,392,667	chr20:58,226,442-60,850,479	n/a
SCA27	*FGF14*	13q33.1	chr13:102,373,205-103,054,124	chr22:54,328,225-54,929,598	No coding polymorphisms identified.
SCA28	*AFG3L2*	18p11.21	chr18:12,328,943-12,377,275	chr7:80,792,562-80,830,840	No polymorphisms identified
SCA29	-	heterogeneous	-	-	n/a
SCA30	-	4q34.3-q35.1	chr4:179,213,356-184,216,082	chr16:47,000,000-53,000,000	n/a
SCA31	*BEAN1*	16q21	chr16:66,460,816-66,527,432	chr5:85,780,302-85,806,244	1 non synonymous exonic SNP (p.R247Q). Conservation data suggests R or Q acceptable.
SCA32	-	7q32-q33	chr7:131125523-132115310	chr14 77,000,000-86,000,000	n/a
SCA33	*not characterised*	-	-	-	n/a
SCA34	-	6p12.3-q16.2	chr6:46,200,000 - 100,600,000	chr12	n/a
SCA35	*TGM6*	20p13	chr20:2,361,554-2,413,399	chr24:21,697,299-21,709,977	Insufficient read depth to analyse
SCA36	*NOP56*	20p13	chr20:2,633,178-2,639,039	chr24:21,559,606-21,563,511	No polymorphisms identified
SCAR1	*SETX*	9q34.13	chr9:135,136,827-135,230,372	chr9:55,244,347-55,323,850	Heterozygous SNPs identified - gene excluded
SCAR2	-	9q34-qter	chr9:137,919,616-138,285,463	chr9:53,000,000-55,000,000	n/a
SCAR3	-	6p23-p21	chr6:13,400,000 - 46,200,000	chr35,chr12	n/a
SCAR4	-	1p36	chr1:3,584,862-15,028,985	chr5,chr2	n/a
SCAR5	*ZNF592*	15q25.3	chr15:85,291,818-85,349,663	chr3:56,896,457-56,914,856	No coding polymorphism identified
SCAR6	-	20q11-q13	chr20:19,831,375-43,649,175	chr23,chr24	n/a
SCAR7	-	11p15	chr11:2754932-7292350	chr18,chr21	n/a
SCAR8	*SYNE1*	6q25.1-q25.2	chr6:152,442,819-152,958,534	chr1:45,428,916-45,877,003	
SCAR9	*ADCK3*	1q42.13	chr1:227,127,938-227,175,246	chr7:41,056,946-41,076,326	1 non-synonymous SNP (p.S328P.) P is conserved among species
SCAR10	*ANO10*	3p22.1	chr3:43,407,818-43,663,560	chr23:5,692,957-5,911,487	No polymorphisms identified
SCAR11	*SYT14*	1q32.2	chr1:210,111,538-210,337,633	chr7:11,556,473-11,746,853	One synonymous SNP identified
SCAR12	-	16q21-q23	chr16:65,067,301-82,980,450	chr5:71,732,917-87,083,276	n/a
Friedreich ataxia	*FXN*	9q21.11	chr9:71,650,479-71,715,094	chr1:91,355,667-91,378,319	One synonymous SNP identified

References

Acland, G. M., Aguirre, G. D., Ray, J., Zhang, Q., Aleman, T. S., Cideciyan, A. V., Pearce-Kelling, S. E., Anand, V., Zeng, Y., Maguire, A. M., Jacobson, S. G., Hauswirth, W. W. & Bennett, J. (2001) Gene therapy restores vision in a canine model of childhood blindness. *Nat Genet,* 28**,** 92-5.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. & Sunyaev, S. R. (2010) A method and server for predicting damaging missense mutations. *Nat Methods,* 7**,** 248-9.

Asan, N. F., Xu, Y., Jiang, H., Tyler-Smith, C., Xue, Y., Jiang, T., Wang, J., Wu, M., Liu, X., Tian, G., Yang, H. & Zhang, X. (2011) Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol,* 12**,** R95.

Astle, W. & Balding, D. J. (2009) Population structure and cryptic relatedness in genetic association studies. *Statistical Science,* 24**,** 451–471.

Azam, M., Andrabi, S. S., Sahr, K. E., Kamath, L., Kuliopulos, A. & Chishti, A. H. (2001) Disruption of the mouse mu-calpain gene reveals an essential role in platelet function. *Mol Cell Biol,* 21**,** 2213-20.

Banerjee, S. & Hasan, G. (2005) The InsP3 receptor: its role in neuronal physiology and neurodegeneration. *Bioessays,* 27**,** 1035-47.

Bannai, H., Fukatsu, K., Mizutani, A., Natsume, T., Iemura, S., Ikegami, T., Inoue, T. & Mikoshiba, K. (2004) An RNA-interacting protein, SYNCRIP (heterogeneous nuclear ribonuclear protein Q1/NSAP1) is a component of mRNA granule transported with inositol 1,4,5-trisphosphate receptor type 1 mRNA in neuronal dendrites. *J Biol Chem,* 279**,** 53427-34.

Barnett, K. C. (2006) Congenital keratoconjunctivitis sicca and ichthyosiform dermatosis in the cavalier King Charles spaniel. *J Small Anim Pract,* 47**,** 524-8.

Barrett, A. J. & Rawlings, N. D. (2001) Evolutionary lines of cysteine peptidases. *Biol Chem,* 382**,** 727-33.

Bauer, P., Schols, L. & Riess, O. (2006) Spectrin mutations in spinocerebellar ataxia (SCA). *Bioessays,* 28**,** 785-7.

Beardow, A. W. & Buchanan, J. W. (1993) Chronic mitral valve disease in cavalier King Charles spaniels: 95 cases (1987-1991). *J Am Vet Med Assoc,* 203**,** 1023-9.

Beck, C. L., Fahlke, C. & George, A. L., Jr. (1996) Molecular basis for decreased muscle chloride conductance in the myotonic goat. *Proc Natl Acad Sci U S A,* 93**,** 11248-52.

Becker, R. E. (1977) Myotonia congenita and syndromes associated with myotonia. *Top Hum Gen,* Vol. III. .

Bekku, Y., Rauch, U., Ninomiya, Y. & Oohashi, T. (2009) Brevican distinctively assembles extracellular components at the large diameter nodes of Ranvier in the CNS. *J Neurochem,* 108**,** 1266-76.

Bekku, Y., Vargova, L., Goto, Y., Vorisek, I., Dmytrenko, L., Narasaki, M., Ohtsuka, A., Fassler, R., Ninomiya, Y., Sykova, E. & Oohashi, T. (2010) Bral1: its role in diffusion barrier formation and conduction velocity in the CNS. *J Neurosci,* 30**,** 3113-23.

Bennett, J., Ashtari, M., Wellman, J., Marshall, K. A., Cyckowski, L. L., Chung, D. C., McCague, S., Pierce, E. A., Chen, Y., Bennicelli, J. L., Zhu, X., Ying, G. S., Sun, J., Wright, J. F., Auricchio, A., Simonelli, F., Shindler, K. S., Mingozzi, F., High, K. A. & Maguire, A. M. (2012) AAV2 gene therapy readministration in three adults with congenital blindness. *Sci Transl Med,* 4**,** 120ra15.

Bennett, V. & Baines, A. J. (2001) Spectrin and ankyrin-based pathways: metazoan inventions for integrating cells into tissues. *Physiol Rev,* 81**,** 1353-92.

Berti, P. J. & Storer, A. C. (1995) Alignment/phylogeny of the papain superfamily of cysteine proteases. *J Mol Biol,* 246**,** 273-83.

Bjorck, G., Dyrendahl, S., Olsson, S. E. (1957) Hereditary ataxia in Smooth-haired Fox Terriers. *Vet Rec,* 69**,** 87--92.

Bjorck, G., Mair, W., Olsson, S. E., Sourander, P. (1962) Hereditary ataxia in Fox Terriers. *Acta Neuropath,* 1**,** 45-48.

Boersma, A., Zonnevylle, H., Sanchez, M. A. & Diaz Espineira, M. (1995) Progressive ataxia in a rottweiler dog. *Vet Q,* 17**,** 108-9.

Bonfield, J. K., Smith, K. & Staden, R. (1995) A new DNA sequence assembly program. *Nucleic Acids Res,* 23**,** 4992-9.

Bornman, D. M., Hester, M. E., Schuetter, J. M., Kasoji, M. D., Minard-Smith, A., Barden, C. A., Nelson, S. C., Godbold, G. D., Baker, C. H., Yang, B., Walther, J. E., Tornes, I. E., Yan, P. S., Rodriguez, B., Bundschuh, R., Dickens, M. L., Young, B. A. & Faith, S. A. (2012) Short-read, high-throughput sequencing technology for STR genotyping. *Biotechniques*.

Brakebusch, C., Seidenbecher, C. I., Asztely, F., Rauch, U., Matthies, H., Meyer, H., Krug, M., Bockers, T. M., Zhou, X., Kreutz, M. R., Montag, D., Gundelfinger, E. D. & Fassler, R. (2002) Brevican-deficient mice display impaired hippocampal CA1 long-term potentiation but show no obvious deficits in learning and memory. *Mol Cell Biol,* 22**,** 7417-27.

Bralic, M., Stemberga, V. & Stifter, S. (2012) Introduction of calpain inhibitors in traumatic brain injury: a novel approach? *Med Hypotheses,* 79**,** 358-60.

Breen, M., Jouquand, S., Renier, C., Mellersh, C. S., Hitte, C., Holmes, N. G., Cheron, A., Suter, N., Vignaux, F., Bristow, A. E., Priat, C., McCann, E., Andre, C., Boundy, S., Gitsham, P., Thomas, R., Bridge, W. L., Spriggs, H. F., Ryder, E. J., Curson, A., Sampson, J., Ostrander, E. A., Binns, M. M. & Galibert, F. (2001) Chromosome-specific single-locus FISH probes allow anchorage of an 1800-marker integrated radiation-hybrid/linkage map of the domestic dog genome to all chromosomes. *Genome Res,* 11**,** 1784-95.

Breen, M., Reimann, N., Bosma, A. A., Landon, D., Zijlstra, C., Bartnitzke, S., Switonski, M., Long, S. E., de Haan, N. A. & Binns, M. M. (1998) *Standardisation of the chromosome nos. 22-38 of the dog (Canis familiaris) with the use of chromosome painting probes. Proceedings of the 13th European Colloquium on Cytogenetics of Domestic Animals. June 1-6. Hungarian Academy of Sciences, Budapest, Hungary.*

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S. R., Moon, K., Burcham, T., Pallas, M., DuBridge, R. B., Kirchner, J., Fearon, K., Mao, J. & Corcoran, K. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol,* 18**,** 630-4.

Brody, I. A. (1969) Muscle contracture induced by exercise. A syndrome attributable to decreased relaxing factor. *N Engl J Med,* 281**,** 187-92.

Campuzano, V., Montermini, L., Molto, M. D., Pianese, L., Cossee, M., Cavalcanti, F., Monros, E., Rodius, F., Duclos, F., Monticelli, A., Zara, F., Canizares, J., Koutnikova, H., Bidichandani, S. I., Gellera, C., Brice, A., Trouillas, P., De Michele, G., Filla, A., De Frutos, R., Palau, F., Patel, P. I., Di Donato, S., Mandel, J. L., Cocozza, S., Koenig, M. & Pandolfo, M. (1996) Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science,* 271**,** 1423-7.

Carmichael, S., Griffiths, I. R. & Harvey, M. J. (1983) Familial cerebellar ataxia with hydrocephalus in bull mastiffs. *Vet Rec,* 112**,** 354-8.

Chieffo, C., Stalis, I. H., Van Winkle, T. J., Haskins, M. E. & Patterson, D. F. (1994) Cerebellar Purkinje's cell degeneration and coat color dilution in a family of Rhodesian Ridgeback dogs. *J Vet Intern Med,* 8**,** 112-6.

Clark, R. G., Hartley, W. J., Burgess, G. S., Cameron, J. S. & Mitchell, G. (1982) Suspected inherited cerebellar neuroaxonal dystrophy in collie sheep dogs. *N Z Vet J,* 30**,** 102-3.

Clarke, J., Wu, H. C., Jayasinghe, L., Patel, A., Reid, S. & Bayley, H. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol,* 4**,** 265-70.

Clarkson, Y. L., Gillespie, T., Perkins, E. M., Lyndon, A. R. & Jackson, M. (2010) Beta-III spectrin mutation L253P associated with spinocerebellar ataxia type 5 interferes with binding to Arp1 and protein trafficking from the Golgi. *Hum Mol Genet,* 19**,** 3634-41.

Coates, J. R., O'Brien, D. P., Kline, K. L., Storts, R. W., Johnson, G. C., Shelton, G. D., Patterson, E. E. & Abbott, L. C. (2002) Neonatal cerebellar ataxia in Coton de Tulear dogs. *J Vet Intern Med,* 16**,** 680-9.

Copeland, H., Dukes-McEwan, J., Sargan, D., Kennedy, L., Starkey, M., Hendricks, A. & Callanan, S. (2008) LUPA - studying human diseases using dog genetics. *Vet Rec,* 163**,** 550.

Cossee, M., Schmitt, M., Campuzano, V., Reutenauer, L., Moutou, C., Mandel, J. L. & Koenig, M. (1997) Evolution of the Friedreich's ataxia trinucleotide repeat expansion: founder effect and premutations. *Proc Natl Acad Sci U S A,* 94**,** 7452-7.

Credille, K. M., Barnhart, K. F., Minor, J. S. & Dunstan, R. W. (2005) Mild recessive epidermolytic hyperkeratosis associated with a novel keratin 10 donor splice-site mutation in a family of Norfolk terrier dogs. *Br J Dermatol,* 153**,** 51-8.

Credille, K. M., Minor, J. S., Barnhart, K. F., Lee, E., Cox, M. L., Tucker, K. A., Diegel, K. L., Venta, P. J., Hohl, D., Huber, M. & Dunstan, R. W. (2009) Transglutaminase 1-deficient recessive lamellar ichthyosis associated with a LINE-1 insertion in Jack Russell terrier dogs. *Br J Dermatol,* 161**,** 265-72.

Cremers, F. P., van den Hurk, J. A. & den Hollander, A. I. (2002) Molecular genetics of Leber congenital amaurosis. *Hum Mol Genet,* 11**,** 1169-76.

Cunliffe, J. (2004) *Cavalier King Charles Spaniel*, Kennel Club Books.

Darke, P. G. & Kelly, D. F. (1976) Correspondence: Cerebellar ataxia in the kerry blue terrier. *Vet Rec,* 98**,** 307.

Deelman, L. E., Jonk, L. J. & Henning, R. H. (1998) The isolation and characterization of the promoter of the human type 1 inositol 1,4,5-trisphosphate receptor. *Gene,* 207**,** 219-25.

Demarchi, F. & Schneider, C. (2007) The calpain system as a modulator of stress/damage response. *Cell Cycle,* 6**,** 136-8.

Derrien, T., Vaysse, A., Andre, C. & Hitte, C. (2012) Annotation of the domestic dog genome sequence: finding the missing genes. *Mamm Genome,* 23**,** 124-31.

Ding, Z. L., Oskarsson, M., Ardalan, A., Angleby, H., Dahlgren, L. G., Tepeli, C., Kirkness, E., Savolainen, P. & Zhang, Y. P. (2011) Origins of domestic dog in Southern East Asia is supported by analysis of Y-chromosome DNA. *Heredity,* 108, 507-14.

Durr, A., Cossee, M., Agid, Y., Campuzano, V., Mignard, C., Penet, C., Mandel, J. L., Brice, A. & Koenig, M. (1996) Clinical and genetic abnormalities in patients with Friedreich's ataxia. *N Engl J Med,* 335**,** 1169-75.

Dutt, P., Croall, D. E., Arthur, J. S., Veyra, T. D., Williams, K., Elce, J. S. & Greer, P. A. (2006) m-Calpain is required for preimplantation embryonic development in mice. *BMC Dev Biol,* 6**,** 3.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J. & Turner, S. (2009) Real-time DNA sequencing from single polymerase molecules. *Science,* 323**,** 133-8.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S. & Mitchell, S. E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One,* 6**,** e19379.

Epplen, C., Epplen, J. T., Frank, G., Miterski, B., Santos, E. J. & Schols, L. (1997) Differential stability of the (GAA)n tract in the Friedreich ataxia (STM7) gene. *Hum Genet,* 99**,** 834-6.

Faddis, B. T., Hasbani, M. J. & Goldberg, M. P. (1997) Calpain activation contributes to dendritic remodeling after brief excitotoxic injury *in vitro*. *J Neurosci,* 17**,** 951-9.

Finnigan, D. F., Hanna, W. J., Poma, R. & Bendall, A. J. (2007) A novel mutation of the *CLCN1* gene associated with myotonia hereditaria in an Australian cattle dog. *J Vet Intern Med,* 21**,** 458-63.

Fischer, J., Bouadjar, B., Heilig, R., Huber, M., Lefevre, C., Jobard, F., Macari, F., Bakija-Konsuo, A., Ait-Belkacem, F., Weissenbach, J., Lathrop, M., Hohl, D. & Prud'homme, J. F. (2001) Mutations in the gene encoding SLURP-1 in Mal de Meleda. *Hum Mol Genet,* 10**,** 875-80.

Furuichi, T., Simon-Chazottes, D., Fujino, I., Yamada, N., Hasegawa, M., Miyawaki, A., Yoshikawa, S., Guenet, J. L. & Mikoshiba, K. (1993) Widespread expression of inositol 1,4,5-trisphosphate receptor type 1 gene (*Insp3r1*) in the mouse central nervous system. *Recept Chan,* 1**,** 11-24.

Gandini, G., Botteron, C., Brini, E., Fatzer, R., Diana, A. & Jaggy, A. (2005) Cerebellar cortical degeneration in three English bulldogs: clinical and neuropathological findings. *J Small Anim Pract,* 46**,** 291-4.

Gao, Y., Perkins, E. M., Clarkson, Y. L., Tobia, S., Lyndon, A. R., Jackson, M. & Rothstein, J. D. (2011) beta-III spectrin is critical for development of purkinje cell dendritic tree and spine morphogenesis. *J Neurosci,* 31**,** 16581-90.

Garosi, L. S., Platt, S. & Shelton, G. D. (2002) Hypertonicity in the Cavalier King Charles Spaniel. *J Vet Intern Med,* 16**,** 330.

George, A. L., Jr., Crackower, M. A., Abdalla, J. A., Hudson, A. J. & Ebers, G. C. (1993) Molecular basis of Thomsen's disease (autosomal dominant myotonia congenita). *Nat Genet,* 3**,** 305-10.

Gill, J. L., Tsai, K. L., Krey, C., Noorai, R. E., Vanbellinghen, J. F., Garosi, L. S., Shelton, G. D., Clark, L. A. & Harvey, R. J. (2011) A canine *BCAN* microdeletion associated with episodic falling syndrome. *Neurobiol Dis,* 45**,** 130-6.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D. B., Lander, E. S. & Nusbaum, C. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol,* 27**,** 182-9.

Grall, A., Guaguere, E., Planchais, S., Grond, S., Bourrat, E., Hausser, I., Hitte, C., Le Gallo, M., Derbois, C., Kim, G. J., Lagoutte, L., Degorce-Rubiales, F., Radner, F. P., Thomas, A., Kury, S., Bensignor, E., Fontaine, J., Pin, D., Zimmermann, R., Zechner, R., Lathrop, M., Galibert, F., Andre, C. & Fischer, J. (2012) *PNPLA1* mutations cause autosomal recessive congenital ichthyosis in golden retriever dogs and humans. *Nat Genet,* 44**,** 140-7.

Grantham, R. (1974) Amino acid difference formula to help explain protein evolution. *Science,* 185**,** 862-4.

Hanzlicek, D., Kathmann, I., Bley, T., Srenk, P., Botteron, C., Gaillard, C. & Jaggy, A. (2003) [Cerebellar cortical abiotrophy in American Staffordshire terriers: clinical and pathological description of 3 cases]. *Schweiz Arch Tierheilkd,* 145**,** 369-75.

Hara, K., Shiga, A., Nozaki, H., Mitsui, J., Takahashi, Y., Ishiguro, H., Yomono, H., Kurisaki, H., Goto, J., Ikeuchi, T., Tsuji, S., Nishizawa, M. & Onodera, O. (2008) Total deletion and a missense mutation of *ITPR1* in Japanese SCA15 families. *Neurology,* 71**,** 547-51.

Hartley, C., Barnett, K. C., Pettitt, L., Forman, O. P., Blott, S. & Mellersh, C. S. (2012) Congenital keratoconjunctivitis sicca and ichthyosiform dermatosis in Cavalier King Charles spaniel dogs-part II: candidate gene study. *Vet Ophthalmol*, 15, 327-32.

Hartley, C., Donaldson, D., Smith, K. C., Henley, W., Lewis, T. W., Blott, S., Mellersh, C. & Barnett, K. C. (2011) Congenital keratoconjunctivitis sicca and ichthyosiform dermatosis in 25 Cavalier King Charles spaniel dogs - part I: clinical signs, histopathology, and inheritance. *Vet Ophthalmol*. 5, 315-26

Hartley, W. J. & Palmer, A. C. (1973) Ataxia in Jack Russell terriers. *Acta Neuropathol,* 26**,** 71-4.

Hedan, B., Thomas, R., Motsinger-Reif, A., Abadie, J., Andre, C., Cullen, J. & Breen, M. (2011) Molecular cytogenetic characterization of canine histiocytic sarcoma: A spontaneous model for human histiocytic cancer identifies deletion of tumor suppressor genes and highlights influence of genetic background on tumor behavior. *BMC Cancer,* 11**,** 201.

Henikoff, S. & Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A,* 89**,** 10915-9.

Hernandez, D., Francois, P., Farinelli, L., Osteras, M. & Schrenzel, J. (2008) De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res,* 18**,** 802-9.

Herrtage, M. E. & Palmer, A. C. (1983) Episodic falling in the cavalier King Charles spaniel. *Vet Rec,* 112**,** 458-9.

Higgins, J. J., Nee, L. E., Vasconcelos, O., Ide, S. E., Lavedan, C., Goldfarb, L. G. & Polymeropoulos, M. H. (1996) Mutations in American families with spinocerebellar ataxia (SCA) type 3: SCA3 is allelic to Machado-Joseph disease. *Neurology,* 46**,** 208-13.

Higgins, R. J., LeCouteur, R. A., Kornegay, J. N. & Coates, J. R. (1998) Late-onset progressive spinocerebellar degeneration in Brittany Spaniel dogs. *Acta Neuropathol,* 96**,** 97-101.

Hillbertz, N. H. & Andersson, G. (2006) Autosomal dominant mutation causing the dorsal ridge predisposes for dermoid sinus in Rhodesian ridgeback dogs. *J Small Anim Pract,* 47**,** 184-8.

Hitte, C., Derrien, T., Andre, C., Ostrander, E. A. & Galibert, F. (2004) CRH_Server: an online comparative and radiation hybrid mapping server for the canine genome. *Bioinformatics,* 20**,** 3665-7.

Holleran, E. A., Ligon, L. A., Tokito, M., Stankewich, M. C., Morrow, J. S. & Holzbaur, E. L. (2001) beta III spectrin binds to the Arp1 subunit of dynactin. *J Biol Chem,* 276**,** 36598-605.

Holmes, S. E., O'Hearn, E. E., McInnis, M. G., Gorelick-Feldman, D. A., Kleiderlein, J. J., Callahan, C., Kwak, N. G., Ingersoll-Ashworth, R. G., Sherr, M., Sumner, A. J., Sharp, A. H., Ananth, U., Seltzer, W. K., Boss, M. A., Vieria-Saecker, A. M., Epplen, J. T., Riess, O., Ross, C. A. & Margolis, R. L. (1999) Expansion of a novel CAG trinucleotide repeat in the 5' region of *PPP2R2B* is associated with SCA12. *Nat Genet,* 23**,** 391-2.

Hosfield, C. M., Elce, J. S., Davies, P. L. & Jia, Z. (1999) Crystal structure of calpain reveals the structural basis for Ca(2+)-dependent protease activity and a novel mode of enzyme activation. *EMBO J,* 18**,** 6880-9.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I. & Clamp, M. (2002) The Ensembl genome database project. *Nucleic Acids Res,* 30**,** 38-41.

Ikeda, Y., Dick, K. A., Weatherspoon, M. R., Gincel, D., Armbrust, K. R., Dalton, J. C., Stevanin, G., Durr, A., Zuhlke, C., Burk, K., Clark, H. B., Brice, A., Rothstein, J. D., Schut, L. J., Day, J. W. & Ranum, L. P. (2006) Spectrin mutations cause spinocerebellar ataxia type 5. *Nat Genet,* 38**,** 184-90.

Iwaki, A., Kawano, Y., Miura, S., Shibata, H., Matsuse, D., Li, W., Furuya, H., Ohyagi, Y., Taniwaki, T., Kira, J. & Fukumaki, Y. (2008) Heterozygous deletion of *ITPR1*, but not *SUMF1*, in spinocerebellar ataxia type 16. *J Med Genet,* 45**,** 32-5.

Jackson, M., Song, W., Liu, M. Y., Jin, L., Dykes-Hoberg, M., Lin, C. I., Bowers, W. J., Federoff, H. J., Sternweis, P. C. & Rothstein, J. D. (2001) Modulation of the

neuronal glutamate transporter EAAT4 by two interacting proteins. *Nature,* 410**,** 89-93.

Kampfl, A., Posmantur, R., Nixon, R., Grynspan, F., Zhao, X., Liu, S. J., Newcomb, J. K., Clifton, G. L. & Hayes, R. L. (1996) mu-calpain activation and calpain-mediated cytoskeletal proteolysis following traumatic brain injury. *J Neurochem,* 67**,** 1575-83.

The Kennel Club http://www.thekennelclub.org.uk/ (Accesssed on 18th May 2012)

The Kennel Club http://www.thekennelclub.org.uk/download/7748/top20breedreg.pdf (Accesssed on 16th April 2013)

Kent, M., Glass, E. & deLahunta, A. (2000) Cerebellar cortical abiotrophy in a beagle. *J Small Anim Pract,* 41**,** 321-3.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. & Haussler, D. (2002) The human genome browser at UCSC. *Genome Res,* 12**,** 996-1006.

Kim, J. W., Lee, S. K., Lee, Z. H., Park, J. C., Lee, K. E., Lee, M. H., Park, J. T., Seo, B. M., Hu, J. C. & Simmer, J. P. (2008) *FAM83H* mutations in families with autosomal-dominant hypocalcified amelogenesis imperfecta. *Am J Hum Genet,* 82**,** 489-94.

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

Kirkwood, K. L., Homick, K., Dragon, M. B. & Bradford, P. G. (1997) Cloning and characterization of the type I inositol 1,4,5-trisphosphate receptor gene promoter. Regulation by 17beta-estradiol in osteoblasts. *J Biol Chem,* 272**,** 22425-31.

Koch, M. C., Steinmeyer, K., Lorenz, C., Ricker, K., Wolf, F., Otto, M., Zoll, B., Lehmann-Horn, F., Grzeschik, K. H. & Jentsch, T. J. (1992) The skeletal muscle chloride channel in dominant and recessive human myotonia. *Science,* 257**,** 797-800.

Koide, R., Kobayashi, S., Shimohata, T., Ikeuchi, T., Maruyama, M., Saito, M., Yamada, M., Takahashi, H. & Tsuji, S. (1999) A neurological disease caused by an expanded CAG trinucleotide repeat in the TATA-binding protein gene: a new polyglutamine disease? *Hum Mol Genet,* 8**,** 2047-53.

Koob, M. D., Moseley, M. L., Schut, L. J., Benzow, K. A., Bird, T. D., Day, J. W. & Ranum, L. P. (1999) An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). *Nat Genet,* 21**,** 379-84.

Kumar, S., Tao, C., Chien, M., Hellner, B., Balijepalli, A., Robertson, J. W., Li, Z., Russo, J. J., Reiner, J. E., Kasianowicz, J. J. & Ju, J. (2012) PEG-labeled nucleotides and nanopore detection for single molecule DNA sequencing by synthesis. *Sci Rep,* 2**,** 684.

Lander, E. & Kruglyak, L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet,* 11**,** 241-7.

Lathrop, G. M. & Lalouel, J. M. (1984) Easy calculations of lod scores and genetic risks on small computers. *Am J Hum Genet,* 36**,** 460-5.

Le Meur, G., Stieger, K., Smith, A. J., Weber, M., Deschamps, J. Y., Nivard, D., Mendes-Madeira, A., Provost, N., Pereon, Y., Cherel, Y., Ali, R. R., Hamel, C., Moullier, P. & Rolling, F. (2007) Restoration of vision in RPE65-deficient Briard dogs using an AAV serotype 4 vector that specifically targets the retinal pigmented epithelium. *Gene Ther,* 14**,** 292-303.

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

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

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

Lieberman, K. R., Cherf, G. M., Doody, M. J., Olasagasti, F., Kolodji, Y. & Akeson, M. (2010) Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase. *J Am Chem Soc,* 132**,** 17961-72.

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

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., Clamp, M., Chang, J. L., Kulbokas, E. J., 3rd, Zody, M. C., Mauceli, E., Xie, X., Breen, M., Wayne, R. K., Ostrander, E. A., Ponting, C. P., Galibert, F., Smith, D. R., DeJong, P. J., Kirkness, E., Alvarez, P., Biagi, T., Brockman, W., Butler, J., Chin, C. W., Cook, A., Cuff, J., Daly, M. J., DeCaprio, D., Gnerre, S., Grabherr, M., Kellis, M., Kleber, M., Bardeleben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K. P., Parker, H. G., Pollinger, J. P., Searle, S. M., Sutter, N. B., Thomas, R., Webber, C., Baldwin, J., Abebe, A., Abouelleil, A., Aftuck, L., Ait-Zahra, M., Aldredge, T., Allen, N., An, P., Anderson, S., Antoine, C., Arachchi, H., Aslam, A., Ayotte, L., Bachantsang, P., Barry, A., Bayul, T., Benamara, M., Berlin, A., Bessette, D., Blitshteyn, B., Bloom, T., Blye, J., Boguslavskiy, L., Bonnet, C., Boukhgalter, B., Brown, A., Cahill, P., Calixte, N., Camarata, J., Cheshatsang, Y., Chu, J., Citroen, M., Collymore, A., Cooke, P., Dawoe, T., Daza, R., Decktor, K., DeGray, S., Dhargay, N., Dooley, K., Dorje, P., Dorjee, K., Dorris, L., Duffey, N., Dupes, A., Egbiremolen, O., Elong, R., Falk, J., Farina, A., Faro, S., Ferguson, D., Ferreira, P., Fisher, S., FitzGerald, M., Foley, K., et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature,* 438**,** 803-19.

Lindblad, K., Savontaus, M. L., Stevanin, G., Holmberg, M., Digre, K., Zander, C., Ehrsson, H., David, G., Benomar, A., Nikoskelainen, E., Trottier, Y., Holmgren, G., Ptacek, L. J., Anttinen, A., Brice, A. & Schalling, M. (1996) An expanded CAG repeat sequence in spinocerebellar ataxia type 7. *Genome Res,* 6**,** 965-71.

Lise, S., Clarkson, Y., Perkins, E., Kwasniewska, A., Sadighi Akha, E., Schnekenberg, R. P., Suminaite, D., Hope, J., Baker, I., Gregory, L., Green, A., Allan, C., Lamble, S., Jayawant, S., Quaghebeur, G., Cader, M. Z., Hughes, S., Armstrong, R. J., Kanapin, A., Rimmer, A., Lunter, G., Mathieson, I., Cazier, J. B., Buck, D., Taylor, J. C., Bentley, D., McVean, G., Donnelly, P., Knight, S. J., Jackson, M., Ragoussis, J. & Nemeth, A. H. (2012) Recessive mutations in SPTBN2 implicate beta-III spectrin in both cognitive and motor development. *PLoS Genet,* 8**,** e1003074.

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J. & Turner, D. J. (2010) Target-enrichment strategies for next-generation sequencing. *Nat Methods,* 7**,** 111-8.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. & Rothberg, J. M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* 437**,** 376-80.

Matsumoto, M., Nakagawa, T., Inoue, T., Nagata, E., Tanaka, K., Takano, H., Minowa, O., Kuno, J., Sakakibara, S., Yamada, M., Yoneshima, H., Miyawaki, A., Fukuuchi, Y., Furuichi, T., Okano, H., Mikoshiba, K. & Noda, T. (1996) Ataxia and epileptic seizures in mice lacking type 1 inositol 1,4,5-trisphosphate receptor. *Nature,* 379**,** 168-71.

Matsuura, T., Yamagata, T., Burgess, D. L., Rasmussen, A., Grewal, R. P., Watase, K., Khajavi, M., McCall, A. E., Davis, C. F., Zu, L., Achari, M., Pulst, S. M., Alonso, E.,

Noebels, J. L., Nelson, D. L., Zoghbi, H. Y. & Ashizawa, T. (2000) Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. *Nat Genet,* 26**,** 191-4.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M. A. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res,* 20**,** 1297-303.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P. & Cunningham, F. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics,* 26**,** 2069-70.

Mellersh, C. (2008) Microsatellite-based candidate gene linkage analysis studies. *Methods Mol Biol,* 439**,** 75-86.

Mellersh, C. S., Hitte, C., Richman, M., Vignaux, F., Priat, C., Jouquand, S., Werner, P., Andre, C., DeRose, S., Patterson, D. F., Ostrander, E. A. & Galibert, F. (2000) An integrated linkage-radiation hybrid map of the canine genome. *Mamm Genome,* 11**,** 120-30.

Mellersh, C. S., Langston, A. A., Acland, G. M., Fleming, M. A., Ray, K., Wiegand, N. A., Francisco, L. V., Gibbs, M., Aguirre, G. D. & Ostrander, E. A. (1997) A linkage map of the canine genome. *Genomics,* 46**,** 326-36.

Mellersh, C. S., Pettitt, L., Forman, O. P., Vaudin, M. & Barnett, K. C. (2006) Identification of mutations in HSF4 in dogs of three different breeds with hereditary cataracts. *Vet Ophthalmol,* 9**,** 369-78.

Minouchi, O. (1928) The spermatogenesis of the dog, with special reference to meiosis. *Jap J Zoology,* 1**,** 255-268.

Miyadera, K., Kato, K., Boursnell, M., Mellersh, C. S. & Sargan, D. R. (2012) Genome-wide association study in RPGRIP1(-/-) dogs identifies a modifier locus that determines the onset of retinal degeneration. *Mamm Genome,* 23**,** 212-23.

Miyoshi, Y., Yamada, T., Tanimura, M., Taniwaki, T., Arakawa, K., Ohyagi, Y., Furuya, H., Yamamoto, K., Sakai, K., Sasazuki, T. & Kira, J. (2001) A novel autosomal dominant spinocerebellar ataxia (SCA16) linked to chromosome 8q22.1-24.1. *Neurology,* 57**,** 96-100.

National Health Service http://www.nhs.uk/conditions/ataxia (Accesssed on 16th April 2013).

Neff, M. W., Broman, K. W., Mellersh, C. S., Ray, K., Acland, G. M., Aguirre, G. D., Ziegle, J. S., Ostrander, E. A. & Rine, J. (1999) A second-generation genetic linkage map of the domestic dog, Canis familiaris. *Genetics,* 151**,** 803-20.

Ng, P. C. & Henikoff, S. (2001) Predicting deleterious amino acid substitutions. *Genome Res,* 11**,** 863-74.

Nibe, K., Kita, C., Morozumi, M., Awamura, Y., Tamura, S., Okuno, S., Kobayashi, T. & Uchida, K. (2007) Clinicopathological features of canine neuroaxonal dystrophy and cerebellar cortical abiotrophy in Papillon and Papillon-related dogs. *J Vet Med Sci,* 69**,** 1047-52.

Nishino, S., Ripley, B., Overeem, S., Lammers, G. J. & Mignot, E. (2000) Hypocretin (orexin) deficiency in human narcolepsy. *Lancet,* 355**,** 39-40.

Odermatt, A., Taschner, P. E., Khanna, V. K., Busch, H. F., Karpati, G., Jablecki, C. K., Breuning, M. H. & MacLennan, D. H. (1996) Mutations in the gene-encoding SERCA1, the fast-twitch skeletal muscle sarcoplasmic reticulum Ca^{2+} ATPase, are associated with Brody disease. *Nat Genet,* 14**,** 191-4.

Oetting, W. S., Lee, H. K., Flanders, D. J., Wiesner, G. L., Sellers, T. A. & King, R. A. (1995) Linkage analysis with multiplexed short tandem repeat polymorphisms using infrared fluorescence and M13 tailed primers. *Genomics,* 30**,** 450-8.

OMIA http://omia.angis.org.au/ (Accesssed on 18 May 2012)

Orr, H. T., Chung, M. Y., Banfi, S., Kwiatkowski, T. J., Jr., Servadio, A., Beaudet, A. L., McCall, A. E., Duvick, L. A., Ranum, L. P. & Zoghbi, H. Y. (1993) Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nat Genet,* 4**,** 221-6.

Ovodov, N. D., Crockford, S. J., Kuzmin, Y. V., Higham, T. F., Hodgins, G. W. & van der Plicht, J. (2011) A 33,000-year-old incipient dog from the Altai Mountains of Siberia: evidence of the earliest domestication disrupted by the Last Glacial Maximum. *PLoS One,* 6**,** e22821.

Pang, J. F., Kluetsch, C., Zou, X. J., Zhang, A. B., Luo, L. Y., Angleby, H., Ardalan, A., Ekstrom, C., Skollermo, A., Lundeberg, J., Matsumura, S., Leitner, T., Zhang, Y. P. & Savolainen, P. (2009) mtDNA data indicate a single origin for dogs south of Yangtze River, less than 16,300 years ago, from numerous wolves. *Mol Biol Evol,* 26**,** 2849-64.

Parker, H. G., Kukekova, A. V., Akey, D. T., Goldstein, O., Kirkness, E. F., Baysac, K. C., Mosher, D. S., Aguirre, G. D., Acland, G. M. & Ostrander, E. A. (2007) Breed relationships facilitate fine-mapping studies: a 7.8-kb deletion cosegregates with Collie eye anomaly across multiple dog breeds. *Genome Res,* 17**,** 1562-71.

Perkins, E. M., Clarkson, Y. L., Sabatier, N., Longhurst, D. M., Millward, C. P., Jack, J., Toraiwa, J., Watanabe, M., Rothstein, J. D., Lyndon, A. R., Wyllie, D. J., Dutia, M. B. & Jackson, M. (2010) Loss of beta-III spectrin leads to Purkinje cell dysfunction recapitulating the behavior and neuropathology of spinocerebellar ataxia type 5 in humans. *J Neurosci,* 30**,** 4857-67.

Peyron, C., Faraco, J., Rogers, W., Ripley, B., Overeem, S., Charnay, Y., Nevsimalova, S., Aldrich, M., Reynolds, D., Albin, R., Li, R., Hungs, M., Pedrazzoli, M., Padigaru, M., Kucherlapati, M., Fan, J., Maki, R., Lammers, G. J., Bouras, C., Kucherlapati, R., Nishino, S. & Mignot, E. (2000) A mutation in a case of early onset narcolepsy and a generalized absence of hypocretin peptides in human narcoleptic brains. *Nat Med,* 6**,** 991-7.

Pierson, T. M., Adams, D., Bonn, F., Martinelli, P., Cherukuri, P. F., Teer, J. K., Hansen, N. F., Cruz, P., Mullikin For The Nisc Comparative Sequencing Program, J. C., Blakesley, R. W., Golas, G., Kwan, J., Sandler, A., Fuentes Fajardo, K., Markello, T., Tifft, C., Blackstone, C., Rugarli, E. I., Langer, T., Gahl, W. A. & Toro, C. (2011) Whole-exome sequencing identifies homozygous *AFG3L2* mutations in a spastic ataxia-neuropathy syndrome linked to mitochondrial m-AAA proteases. *PLoS Genet,* 7**,** e1002325.

Pignol, B., Auvin, S., Carre, D., Marin, J. G. & Chabrier, P. E. (2006) Calpain inhibitors and antioxidants act synergistically to prevent cell necrosis: effects of the novel dual inhibitors (cysteine protease inhibitor and antioxidant) BN 82204 and its pro-drug BN 82270. *J Neurochem,* 98**,** 1217-28.

Porreca, G. J., Zhang, K., Li, J. B., Xie, B., Austin, D., Vassallo, S. L., LeProust, E. M., Peck, B. J., Emig, C. J., Dahl, F., Gao, Y., Church, G. M. & Shendure, J. (2007) Multiplex amplification of large sets of human exons. *Nat Methods,* 4**,** 931-6.

Priat, C., Hitte, C., Vignaux, F., Renier, C., Jiang, Z., Jouquand, S., Cheron, A., Andre, C. & Galibert, F. (1998) A whole-genome radiation hybrid map of the dog genome. *Genomics,* 54**,** 361-78.

PubMeb National Center for Biotechnology Information, U.S. http://www.ncbi.nlm.nih.gov/pubmed/ (Accesssed on 18th May 2010 2010)

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

Rhodes, T. H., Vite, C. H., Giger, U., Patterson, D. F., Fahlke, C. & George, A. L., Jr. (1999) A missense mutation in canine *C1C-1* causes recessive myotonia congenita in the dog. *FEBS Lett,* 456**,** 54-8.

Richard, G., Rouan, F., Willoughby, C. E., Brown, N., Chung, P., Ryynanen, M., Jabs, E. W., Bale, S. J., DiGiovanna, J. J., Uitto, J. & Russell, L. (2002) Missense mutations in *GJB2* encoding connexin-26 cause the ectodermal dysplasia keratitis-ichthyosis-deafness syndrome. *Am J Hum Genet,* 70**,** 1341-8.

Richard, I., Broux, O., Allamand, V., Fougerousse, F., Chiannilkulchai, N., Bourg, N., Brenguier, L., Devaud, C., Pasturaud, P., Roudaut, C. & et al. (1995) Mutations in

the proteolytic enzyme calpain 3 cause limb-girdle muscular dystrophy type 2A. *Cell,* 81**,** 27-40.

Riess, O., Schols, L., Bottger, H., Nolte, D., Vieira-Saecker, A. M., Schimming, C., Kreuz, F., Macek, M., Jr., Krebsova, A., Macek, M. S., Klockgether, T., Zuhlke, C. & Laccone, F. A. (1997) SCA6 is caused by moderate CAG expansion in the alpha1A-voltage-dependent calcium channel gene. *Hum Mol Genet,* 6**,** 1289-93.

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. (2011) Integrative genomics viewer. *Nat Biotechnol,* 29**,** 24-6.

Rohdin, C., Ludtke, L., Wohlsein, P. & Jaderlund, K. H. (2010) New aspects of hereditary ataxia in smooth-haired fox terriers. *Vet Rec,* 166**,** 557-60.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. & Nyren, P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem,* 242**,** 84-9.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., Clark, T. A., Huber, M., Branciforte, J. T., Stoner, I. B., Cawley, S. E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J. A., Namsaraev, E., McKernan, K. J., Williams, A., Roth, G. T. & Bustillo, J. (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature,* 475**,** 348-52.

Rozen, S. & Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol,* 132**,** 365-86.

Rusbridge, C. (2005) Neurological diseases of the Cavalier King Charles spaniel. *J Small Anim Pract,* 46**,** 265-72.

Saido, T. C., Yokota, M., Nagao, S., Yamaura, I., Tani, E., Tsuchiya, T., Suzuki, K. & Kawashima, S. (1993) Spatial resolution of fodrin proteolysis in postischemic brain. *J Biol Chem,* 268**,** 25239-43.

Sakaguchi, G., Orita, S., Naito, A., Maeda, M., Igarashi, H., Sasaki, T. & Takai, Y. (1998) A novel brain-specific isoform of beta spectrin: isolation and its interaction with Munc13. *Biochem Biophys Res Commun,* 248**,** 846-51.

Sakamoto, N., Chastain, P. D., Parniewski, P., Ohshima, K., Pandolfo, M., Griffith, J. D. & Wells, R. D. (1999) Sticky DNA: self-association properties of long GAA.TTC repeats in R.R.Y triplex structures from Friedreich's ataxia. *Mol Cell,* 3**,** 465-75.

Sambrook, J. (2001) Molecular Cloning: A Laboratory Manual
New York, Cold Spring Harbor Laboratory Press.

Sanchez-Carpintero Abad, R., Sanmarti Vilaplana, F. X. & Serratosa Fernandez, J. M. (2007) Genetic causes of epilepsy. *Neurologist,* 13**,** S47-51.

Sanpei, K., Takano, H., Igarashi, S., Sato, T., Oyake, M., Sasaki, H., Wakisaka, A., Tashiro, K., Ishida, Y., Ikeuchi, T., Koide, R., Saito, M., Sato, A., Tanaka, T., Hanyu, S., Takiyama, Y., Nishizawa, M., Shimizu, N., Nomura, Y., Segawa, M., Iwabuchi, K., Eguchi, I., Tanaka, H., Takahashi, H. & Tsuji, S. (1996) Identification of the spinocerebellar ataxia type 2 gene using a direct identification of repeat expansion and cloning technique, DIRECT. *Nat Genet,* 14**,** 277-84.

Sargan, D. R. (2004) IDID: inherited diseases in dogs: web-based information for canine inherited disease genetics. *Mamm Genome,* 15**,** 503-6.

Schwarz, J. M., Rodelsperger, C., Schuelke, M. & Seelow, D. (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods,* 7**,** 575-6.

Seppala, E. H., Koskinen, L. L., Gullov, C. H., Jokinen, P., Karlskov-Mortensen, P., Bergamasco, L., Baranowska Korberg, I., Cizinauskas, S., Oberbauer, A. M., Berendt, M., Fredholm, M. & Lohi, H. (2012) Identification of a novel idiopathic epilepsy locus in Belgian shepherd dogs. *PLoS One,* 7**,** e33549.

Shelton, G. D. (2004) Muscle pain, cramps and hypertonicity. *Vet Clin North Am Small Anim Pract,* 34**,** 1483-96.

Shelton, G. D. & Engvall, E. (2002) Muscular dystrophies and other inherited myopathies. *Vet Clin North Am Small Anim Pract,* 32**,** 103-24.

Shendure, J. & Ji, H. (2008) Next-generation DNA sequencing. *Nat Biotechnol,* 26**,** 1135-45.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D. & Church, G. M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science,* 309**,** 1728-32.

Skinner, B. A., Greist, M. C. & Norins, A. L. (1981) The keratitis, ichthyosis, and deafness (KID) syndrome. *Arch Dermatol,* 117**,** 285-9.

Sorimachi, H., Ishiura, S. & Suzuki, K. (1997) Structure and physiological function of calpains. *Biochem J,* 328**,** 721-32.

Stankewich, M. C., Gwynn, B., Ardito, T., Ji, L., Kim, J., Robledo, R. F., Lux, S. E., Peters, L. L. & Morrow, J. S. (2010) Targeted deletion of betaIII spectrin impairs synaptogenesis and generates ataxic and seizure phenotypes. *Proc Natl Acad Sci U S A,* 107**,** 6022-7.

Steinberg, H. S., Troncoso, J. C., Cork, L. C. & Price, D. L. (1981) Clinical features of inherited cerebellar degeneration in Gordon setters. *J Am Vet Med Assoc,* 179**,** 886-90.

Steinlein, O. K. (2008) Genetics and epilepsy. *Dialogues Clin Neurosci,* 10**,** 29-38.

Storey, E., Gardner, R. J., Knight, M. A., Kennerson, M. L., Tuck, R. R., Forrest, S. M. & Nicholson, G. A. (2001) A new autosomal dominant pure cerebellar ataxia. *Neurology,* 57**,** 1913-5.

Street, V. A., Bosma, M. M., Demas, V. P., Regan, M. R., Lin, D. D., Robinson, L. C., Agnew, W. S. & Tempel, B. L. (1997) The type 1 inositol 1,4,5-trisphosphate receptor gene is altered in the opisthotonos mouse. *J Neurosci,* 17**,** 635-45.

Sutter, N. B., Eberle, M. A., Parker, H. G., Pullar, B. J., Kirkness, E. F., Kruglyak, L. & Ostrander, E. A. (2004) Extensive and breed-specific linkage disequilibrium in *Canis familiaris. Genome Res,* 14**,** 2388-96.

Switonski, M., Reimann, N., Bosma, A. A., Long, S., Bartnitzke, S., Pienkowska, A., Moreno-Milan, M. M. & Fischer, P. (1996) Report on the progress of standardization of the G-banded canine (*Canis familiaris*) karyotype. Committee for the Standardized Karyotype of the Dog (*Canis familiaris*). *Chromosome Res,* 4**,** 306-9.

Tan, Y., Dourdin, N., Wu, C., De Veyra, T., Elce, J. S. & Greer, P. A. (2006) Conditional disruption of ubiquitous calpains in the mouse. *Genesis,* 44**,** 297-303.

Thannickal, T. C., Moore, R. Y., Nienhuis, R., Ramanathan, L., Gulyani, S., Aldrich, M., Cornford, M. & Siegel, J. M. (2000) Reduced number of hypocretin neurons in human narcolepsy. *Neuron,* 27**,** 469-74.

Thomsen, J. (1876) Tonische kraempfe in willkuerlich beweglichen muskeln in folge von ererbter psychischer disposition: Ataxia muscularis? *Arch. Psychiat. Nervenkr,* 6**,** 702-718.

Tijssen, M. A., Schoemaker, H. C., Edelbroek, P. J., Roos, R. A., Cohen, A. F. & van Dijk, J. G. (1997) The effects of clonazepam and vigabatrin in hyperekplexia. *J Neurol Sci,* 149**,** 63-7.

Tonelli, A., D'Angelo, M. G., Salati, R., Villa, L., Germinasi, C., Frattini, T., Meola, G., Turconi, A. C., Bresolin, N. & Bassi, M. T. (2006) Early onset, non fluctuating spinocerebellar ataxia and a novel missense mutation in *CACNA1A* gene. *J Neurol Sci,* 241**,** 13-7.

Touma, E., Kato, S., Fukui, K. & Koike, T. (2007) Calpain-mediated cleavage of collapsin response mediator protein(CRMP)-2 during neurite degeneration in mice. *Eur J Neurosci,* 26**,** 3368-81.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L. & Pachter, L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc,* 7**,** 562-78.

Urzua, B., Ortega-Pinto, A., Morales-Bozo, I., Rojas-Alcayaga, G. & Cifuentes, V. (2011) Defining a new candidate gene for amelogenesis imperfecta: from molecular genetics to biochemistry. *Biochem Genet,* 49**,** 104-21.

van de Leemput, J., Chandran, J., Knight, M. A., Holtzclaw, L. A., Scholz, S., Cookson, M. R., Houlden, H., Gwinn-Hardy, K., Fung, H. C., Lin, X., Hernandez, D., Simon-Sanchez, J., Wood, N. W., Giunti, P., Rafferty, I., Hardy, J., Storey, E., Gardner, R. J., Forrest, S. M., Fisher, E. M., Russell, J. T., Cai, H. & Singleton, A. B. (2007) Deletion at ITPR1 underlies ataxia in mice and spinocerebellar ataxia 15 in humans. *PLoS Genet,* 3**,** e108.

van der Merwe, L. L. & Lane, E. (2001) Diagnosis of cerebellar cortical degeneration in a Scottish terrier using magnetic resonance imaging. *J Small Anim Pract,* 42**,** 409-12.

van Steensel, M. A., Koedam, M. I., Swinkels, O. Q., Rietveld, F. & Steijlen, P. M. (2001) Woolly hair, premature loss of teeth, nail dystrophy, acral hyperkeratosis and facial abnormalities: possible new syndrome in a Dutch kindred. *Br J Dermatol,* 145**,** 157-61.

van Tongern, S. E., van Vonderen, I. K., van Nes, J. J. & van den Ingh, T. S. (2000) Cerebellar cortical abiotrophy in two Portuguese Podenco littermates. *Vet Q,* 22**,** 172-4.

Vanhaesebrouck, A., Franklin, R., Van Ham, L. & Bhatti, S. (2012) Hereditary ataxia, myokymia and neuromyotonia in Jack Russell terriers. *Vet Rec,* 171**,** 131-2.

Vignaux, F., Hitte, C., Priat, C., Chuat, J. C., Andre, C. & Galibert, F. (1999) Construction and optimization of a dog whole-genome radiation hybrid panel. *Mamm Genome,* 10**,** 888-94.

Vonholdt, B. M., Pollinger, J. P., Lohmueller, K. E., Han, E., Parker, H. G., Quignon, P., Degenhardt, J. D., Boyko, A. R., Earl, D. A., Auton, A., Reynolds, A., Bryc, K., Brisbin, A., Knowles, J. C., Mosher, D. S., Spady, T. C., Elkahloun, A., Geffen, E., Pilot, M., Jedrzejewski, W., Greco, C., Randi, E., Bannasch, D., Wilton, A., Shearman, J., Musiani, M., Cargill, M., Jones, P. G., Qian, Z., Huang, W., Ding, Z. L., Zhang, Y. P., Bustamante, C. D., Ostrander, E. A., Novembre, J. & Wayne, R. K. (2010) Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature,* 464**,** 898-902.

Wang, W. & Kirkness, E. F. (2005) Short interspersed elements (SINEs) are a major source of canine genomic diversity. *Genome Res,* 15**,** 1798-808.

Wessmann, A., Goedde, T., Fischer, A., Wohlsein, P., Hamann, H., Distl, O. & Tipold, A. (2004) Hereditary ataxia in the Jack Russell Terrier--clinical and genetic investigations. *J Vet Intern Med,* 18**,** 515-21.

Wheeler, S. & Rusbridge, C. (1996) Neurological syndrome in Italian spinones. *Vet Rec,* 138**,** 216.

Wright, J. A., Brownlie, S. E., Smyth, J. B., Jones, D. G. & Wotton, P. (1986) Muscle hypertonicity in the cavalier King Charles spaniel--myopathic features. *Vet Rec,* 118**,** 511-2.

Yamada, H., Watanabe, K., Shimonaka, M. & Yamaguchi, Y. (1994) Molecular cloning of brevican, a novel brain proteoglycan of the aggrecan/versican family. *J Biol Chem,* 269**,** 10119-26.

Yasuba, M., Okimoto, K., Iida, M. & Itakura, C. (1988) Cerebellar cortical degeneration in beagle dogs. *Vet Pathol,* 25**,** 315-7.

Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics,* 25**,** 2865-71.

Yuzbasiyan-Gurkan, V., Blanton, S. H., Cao, Y., Ferguson, P., Li, J., Venta, P. J. & Brewer, G. J. (1997) Linkage of a microsatellite marker to the canine copper toxicosis locus in Bedlington terriers. *Am J Vet Res,* 58**,** 23-7.

Zangerl, B., Goldstein, O., Philp, A. R., Lindauer, S. J., Pearce-Kelling, S. E., Mullins, R. F., Graphodatsky, A. S., Ripoll, D., Felix, J. S., Stone, E. M., Acland, G. M. & Aguirre, G. D. (2006) Identical mutation in a novel retinal gene causes progressive

rod-cone degeneration in dogs and retinitis pigmentosa in humans. *Genomics,* 88**,** 551-63.

Zhou, Q., Ben-Efraim, I., Bigcas, J. L., Junqueira, D., Wiedmer, T. & Sims, P. J. (2005) Phospholipid scramblase 1 binds to the promoter region of the inositol 1,4,5-triphosphate receptor type 1 gene to enhance its expression. *J Biol Chem,* 280**,** 35062-8.