Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics

Whole Genome Sequencing of a Small Pedigree for the Detection of a Rare Form of Retinopathy in Labrador retriever

Marta Gòdia Perelló

Examensarbete / Swedish University of Agricultural Sciences
Department of Animal Breeding and Genetics

Master's Thesis, 30 hp

477

Uppsala 2015

Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Whole Genome Sequencing of a Small Pedigree for the Detection of a Rare Form of Retinopathy in Labrador retriever

Marta Gòdia Perelló

Supervisors:

Tomas Bergström, SLU, Department of Animal Breeding and Genetics
Göran Andersson, SLU, Department of Animal Breeding and Genetics
Björn Ekesten, SLU, Department of Clinical Sciences

Examiner:

Sofia Mikko, SLU, Department of Animal Breeding and Genetics

Credits: 30 HEC
Course title: Degree project in Animal Science
Course code: EX0556
Level: Advanced, A2E

Place of publication: Uppsala
Year of publication: 2015
Cover picture: Björn Ekesten
Name of series: Examensarbete / Swedish University of Agricultural Sciences
Department of Animal Breeding and Genetics, 477
On-line publication: http://epsilon.slu.se

Key words: Whole genome sequencing, retinopathy, Labrador Retriever

Ackowledgments,

To my main supervisor, Tomas Bergström, for proposing me this great project, for his endless commitmet to our work and for giving me every help possible so as to make a master thesis an outstanding scientific work. I couldn't have had a better and nicer mentor.

To Göran Andersson, for sharing his expertice and all the scientific discussions, for his helpful guidance during the project, for his time and effort for cheching this thesis and his helpful comments and improvements.

To Björn Ekesten, for his great enthusiasm for our 'finding', for his encouragement and support during my work, for his clinical expertice and for providing all the material from the dogs that we needed.

Index

1. Introduction

1.2 Origin and domestication of the dog

Selective breeding in the dog has been carried out over centuries. It is generally believed that the closest living ancestor to the domestic dog (*Canis lupus familiaris)* is the gray wolf (*Canis lupus*) (Vila et al. 1997; Lindblad-Toh et al. 2005). Analysis of mitochondrial DNA and Single Nucleotide Polymorphism (SNP) suggests that the domestication process started around 15,000 years ago (Pang et al. 2009; Larson et al. 2012; Thalmann et al. 2013). However, some aspects of the domestication process of the dog remain unclear. It has been suggested that the domestication has been driven by selection on desirable traits associated with behavior (Ding et al. 2012), such as reduced fear and increased stress tolerance (Jensen 2014), adapatation to carbohydrate-rich food (Axelsson et al. 2013), as well as for size and shape (Larson et al. 2012).

Historical events have shaped the dog genome by two major bottlenecks. The first bottleneck occurred 15,000 years ago when dogs were domesticated from a small number of wolves (Lindblad-Toh et al. 2005). The second bottleneck took place approximately 200 years ago, when purebred dogs and breed standards were introduced based on certain desired morphological traits, such as skull shape, coat color, body size, *etc*., as well as behavioral traits like hunting, guarding, guiding, retrieving, pointing *etc*. (Ostrander and Wayne 2005). This has resulted in about 400 different breeds with unique characteristics (Wilcox and Walkowicz 1995). Additionally, genetic variation has been further reduced in some breeds due to breed popularity and breed-specific bottlenecks, resulting in singular pattern of linkage disequilibrium (LD) for each breed (Ostrander and Wayne 2005).

1.2 The domestic dog as a model

The genomic landscape of the dogs is characterized, by long haplotype blocks within breeds and between breeds LD is equivalent humans. This make the dog an attractive model for studies on inherited diseases and traits (Andersson 2001; Neff and Rine 2006). The domestic dog was already recognized as a model for mammalian evolution by Charles Darwin whom noted that the artificial selection has resulted in a remarkable phenotypical and behavioral variation between dog (Darwin 1871). The strict breeding practices have also resulted in an enrichment of unfavorable traits, and dog diseases have striking similarities with some human diseases, including cancer, heart diseases, epilepsy, autoimmune and metabolic diseases and allergies, among others (Patterson 2000; Ostrander and Wayne 2005; Karlsson and Lindblad-Toh 2008). Understanding the genetics underpinning of these disorders is more plausible in dogs than in humans. Firstly, certain diseases appear only in a restricted number of breeds, implying pre-breed ancestors from whose genetic characters are overrepresented in the current population. Moreover, disease presentation can be highly uniform if the animals belong to the same dog family since they will present resembling genetic background. The number of genetic traits and disorders estimated in dogs is about 250 according to the Online Mendelian Inheritance in Animals (OMIA) database (http://omia.angis.org.au/home/), and elucidation of the molecular nature of these diseases has been achieved in more than three-quarters of the total. Notwithstanding, there is still a relative gap of undetermined variants associated to rare or complex genetic diseases that needs to be gauged.

1.3 Advances on the molecular genetics tools

Since the discovery of the DNA structure (Watson and Crick 1953), the field of genetics have had a remarkable development for five decades. Prior to the current genomic technologies such as SNP-typing and whole genome sequencing (WGS), linkage analysis was the main tool used for mapping mendelian and complex diseases. The linkage analysis approach was used to identify a region of the genome associated to the disease of interest by evaluating the segregation of the trait with genetic markers in multiple family members (Katsanis and Katsanis 2013). Later, Quantitative trait loci mapping (QTL mapping) was used for studying quantitative traits that presented polygenic characteristics. In recent years, the focus has turned to genome-wide association studies (GWAS). This method uses Single Nucleotide Polymorphism markers (SNPs) and is suitable for studying both mendelian- and complex traits and diseases. It is produced more efficient mapping since unrelated affected and control organisms are used, surmounting the large linked regions due to a limited recombination within a pedigree family (Karlsson et al. 2007). In the case of dogs, since LD is extensive over large regions due to population isolation, just a few markers are required. Lindblad-Toh et al. (2005) performed simulations and proposed that as few as 20 cases and 20 controls were needed for mapping a recessive trait in dogs by GWAS, but in fact, with only 10 cases and 10 controls successful results were already obtained (Karlsson et al. 2007; Salmon Hillbertz et al. 2007). In recent years the genomics technologies has evolved further and one of the most important technological developments is different Next Generation Sequencing (NGS) technologies.

The pioneer method for DNA sequencing was firstly introduced by Maxam and Gillbert (Maxam and Gilbert 1977). Shortly after, Sanger reported a new method for determine the order of nucleotides in a sequence (Sanger et al. 1977), known as the 'first-generation' technology. Even though its limitations, it must be highlighted that Sanger's method accomplished one the biggest breakthrough in the genomic era, the sequence of the human genome (Lander et al. 2001; Venter et al. 2001). Afterwards, with the arrival of NGS, also known as Massive Parallel Sequencing (MPS), it was offered the possibility to produce large amounts of data with relatively low cost. In recent years, Whole Genome Sequencing (WGS) has become very attractive for its broader coverage and decreasing cost, changing the landscape of rare genetic diseases and offering the advantage to identify the causative genes in an accelerate time (van El et al. 2013). This promising technology has been successfully used for performing WGS to identify the causative mutations for rare diseases (Boycott et al. 2013).

1.4 Genomic resources in dog

Several genomic resources have been produced in *C. lupus familiaris.* The resources include canine meiotic linkage map (Lingaas et al. 1997; Mellersh et al. 1997; Neff et al. 1999; Werner et al. 1999), low- and high-resolution radiation hybrid (RH) maps (Vignaux et al. 1999; Breen et al. 2001; Guyon et al. 2003; Breen et al. 2004; Hitte et al. 2005), and BAC libraries (Li et al. 1999). Additionally, the first dog genome assembly became available in 2005, when the sequence of the female boxer Tasha was published (Lindblad-Toh et al. 2005). Presenting a genome size of 2.4 Gb and covering up ~ 99 % of euchromatic genome, it contains 38 autosomal chromosomes and the sex chromosomes. The dog genome-sequencing project also identified > 2.5 M SNPs, meaning 1 SNP per Kb, and the information was used to create a SNPs genotyping array for performing Genome-Wide Association Mapping (GWAM). The first SNP genotyping array contained 27,000 genes (Karlsson et al. 2007) and was produced by Affymetrix, the current Illumina´s CanineHD SNP chip array contains up to 170K SNPs evenly distributed in canine genome (Vaysse et al. 2011).

1.5 Genetics of retinal atrophies

More than 230 genes for inherited retinal diseases (RDs) have been identified in humans (https://sph.uth.edu/Retnet) (Figure 1). With the improvement of canine-specific genetic resources it has been possibly to identifiy over 24 mutations in 18 different genes from 58 different breeds (Miyadera et al. 2012). Whereas extensive and heterogenic clinical and molecular varieties are encompassed in human RDs, most of the forms of RDs in dogs have been found to be phenotypically and genetically uniform within a specific breed, due to breeding isolation and genetic homogeneity, or shared across different breeds, suggesting that the common ancestor population presented the mutation associated with the retinal disorder (Miyadera et al. 2012).

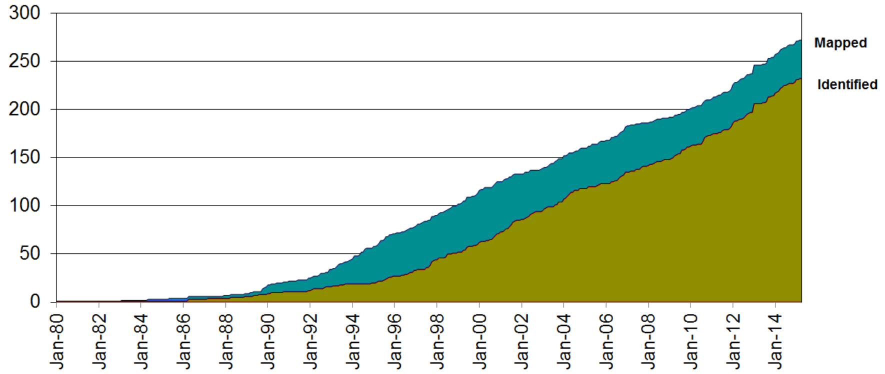


Figure 1. Graphic showing the accumulation of identified genes in human RD from 1980 to 2014. Image extracted and modified from RetNet (https://sph.uth.edu/Retnet/).

The dog structure of the retina is to a certain extent as in humans, composed of different layers of specialized cells. The most external cell layer of the neuroretina contains the cone and rod photoreceptors, which are in contact and nourished by the retinal pigment epithelium (RPE; the outermost cell layer of the entire retina). Below there is the inner nuclear layer (INL) containing the nuclei of the secondary and some tertiary neurons and the Müller glia. The innermost layer of cells, known as ganglion cell layer (GCL) receives the signals orginating from the photoreceptors and transmits the signals to the brain (Miyadera et al. 2012).

One group of well-known and serious hereditary RDs in dogs is progressive retinal atrophy (PRA), most frequently characterized by an initial degradation of rod photoreceptors and initially resulting in night blindness and later in both day and night blindness. Although sometimes only one form of RD is associated to a single variant and segregates uniformly in a breed, clinical studies in dogs have shown high heterogeneity within and across breeds. Age of onset, rate of progression, visual-behavioral abnormalities and allelic heterogeneity may vary (Miyadera et al. 2012). The *prcd* gene is associated to one form of PRA reported in more than 22 breeds (Optigen, http://www.optigen.com/) including the Labrador retriever.

Interestingly, two Swedish Labrador retriever littermates that had been tested normal for the known mutation in the *PRCD*-gene, were examined and diagnosed with a PRA-like retinopathy. Thus, Labrador retrievers are carrying a yet another form of retinopathy with unknown genetic reason.

Aim

The aim of this thesis it to perform a WGS approach to uncover a presumed autosomal recessive form of retinopathy in Labrador retrievers using the Illumina NextSeq® 500 sequencing platform.

2. Materials and methods

2.1 Animal samples

Blood samples were collected from four members of a Labrador retriever family. The samples included the parents, assumed to be carriers for retinopathy, and two offsprings affected by retinopathy (Figure 2).

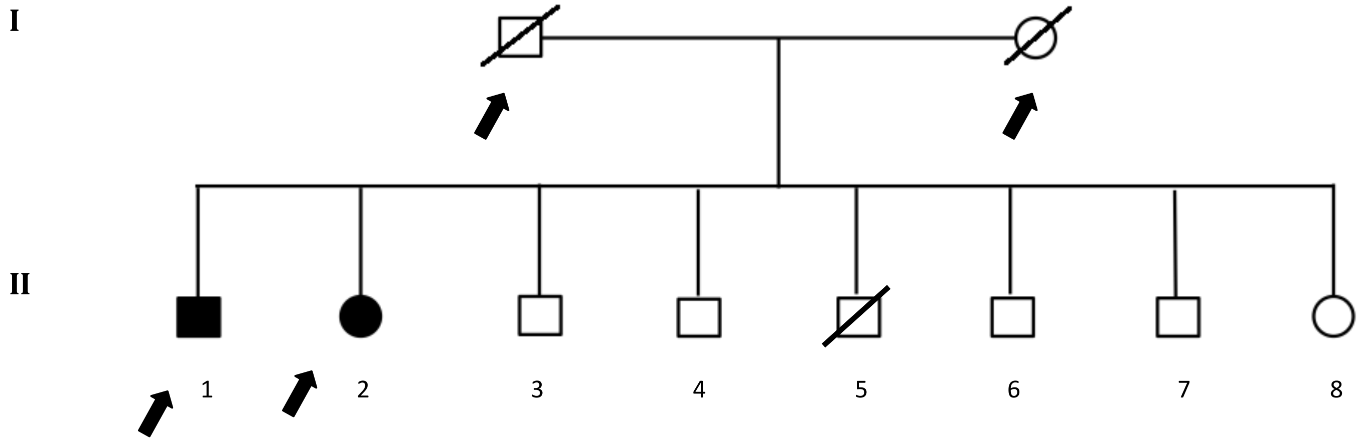


Figure 2. Family pedigree of the Labrador retriever family studied. Open symbols indicate healthy animals whilst solid symbols indicate affected animals. Crossed symbols indicate deceased animal. The four dogs that were whole genome sequenced were denoted as Sire I:1 and Dam I:2 and the littermates Offspring II:1 and Offspring II:2.

2.2 Clinical assessment

Two family members: offspring II:1 and II:2 were clinically evaluated by Swedish veterinary ophthalmologists.

2.3 DNA extraction and quantification

Blood samples were collected in EDTA tubes and genomic DNA was isolated using the QIAsymphony® SP automated system (Qiagen, Hilden, Germany) with the QIAsymphony DSP DNA Midi Kit (Appendix A. Protocol 1). For the unaffected sire, genomic DNA was extracted from a pre-treated sperm sample (Appendix A. Protocol 2) on the QIAsymphony® SP instrument with the QIAsymphony DSP DNA Midi Kit.

To determine DNA concentration of genomic DNA, 3 µl of the samples was analyzed on a Qubit® 2.0 Fluorometer (Life Technologies™, Stockholm, Sweden) using the dsDNA BR Assay.

2.4 Whole Genome Sequencing

Whole Genome Sequencing (WGS) was performed on DNA samples prepared from the four family members. For each of the four dogs, two different library sizes of 350 and 550 bp were prepared following Illumina Low Sample (LS) protocol TruSeq® DNA PCR-Free Library Prep (Illumina®, San Diego, CA). Briefly, genomic DNA from each of the four dogs was fragmented using the Covaris M220 (Covaris, Inc., Woburn, MA) to obtain a library insert size of 350 and 550 bp for run 1 and 2, respectively. Then, the fragmented DNA was end-repaired by removing 3' overhangs and filling the 5' overhangs. Afterwards, sample size selection to remove large and small DNA fragments was performed with Sample Purification Beads. Next, an 'A' nucleotide was added in the 3' ends of the purified fragments to prevent them from ligating each other. Finally, Illumina indexes were ligated for each of the samples (Appendix A. Table 1). The libraries were quantified using the KAPA Library Quantification Kit for Illumina® platforms (Kapa Biosystems, Inc., Wilmington, MA) on a StepOnePlus™ Real-Time PCR System (Life Technologies™) (Appendix A. Protocol 3). The size distributions of the libraries were analyzed on an Agilent 2100 Bioanalyzer Instruments (Agilent Technologies, Santa Clara, CA). The quantified libraries were then normalized and pooled into a single tube. Before loading for sequencing, the pooled library was denatured and diluted to a final concentration of 1.8 and 1.9 pM (for run 1 and 2, respectively). For quality sequencing control, the libraries were "spiked" with 1% PhiX DNA. Lastly, the combined library was loaded onto the Illumina NextSeq® 500 platform using the NextSeq® 500 High Output Kit (300 cycles) and NextSeq® 500 High Output v2 Kit (300 cycles) for run 1 and 2, respectively. The platform generated BLC base call per-cycle files as 100 and 150 bp pair-end reads for run 1 and 2, respectively (See Table 1 for a summary of the 2 Whole Genome Sequencing runs carried out).

Table 1. Whole Genome Sequence summary of the differences between the two runs performed.

	Insert size	Input DNA per sample	NextSeq®500 sequencing kit	Read length	Concentration of library loaded
Run 1	350 bp	1 µg	High Output Kit (300 cycles)	101 bp x 2	1.8 pM
Run 2	550 bp	2 µg	High Output Kit v2 (300 cycles)	150 bp x 2	1.9 pM

2.5 Reads, mapping and variant calling

BCL files were converted and demultiplexed into FastQ files using the Illumina software bclfastq2 (version v.2.15.0). Raw FastQ reads were checked using FastQC (version v.0.11.2) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Short reads (< 30 bp) and bases with a quality score below 20 were trimmed using the software Trimmomatic (version v.0.32) (Bolger et al. 2014) (Appendix B. 1). The high-quality reads were then aligned to the dog reference genome sequence (CanFam3.1) using BWA (version v.0.7.8) (Li and Durbin 2009) (Appendix B. 2). A personalized variant calling workflow was created using the software Genome Analysis Toolkit (GATK) (version v.3.3.0) (DePristo et al. 2011) and following the Best Practices recommendations (Van der Auwera et al. 2013) for Pre-processing and Variant Discovery steps (Appendix B. 3). Final filtered candidate variants, both SNPs and Indels, were obtained by analyzing the two trios separately (Trio 1 was formed by parents and Offspring II:1 and Trio 2 included the parents and Offspring II:2). Annotation was done using the software ANNOVAR (Wang et al. 2010) with the CanFam3.1 annotation. (Appendix B. 4).

2.6 Variant annotation and further filtration

After the annotation of the variants with ANNOVAR, custom-made Perl scripts were created to subdivide the obtained output and to create new files for each annotated region (*e.g.* exonic, splicing, 5' UTR, 3' UTR, *etc.*). The scripts were also used to count and subdivide each of the files according to their pattern of inheritance (Appendix B. 5.1-5.2). Three possible conditional filtering schemas were evaluated:

Schema 1. Parents presented a heterozygous genotype for a given allele and the offspring was homozygous recessive but different from the reference genome for the given loci (Fig. 3)

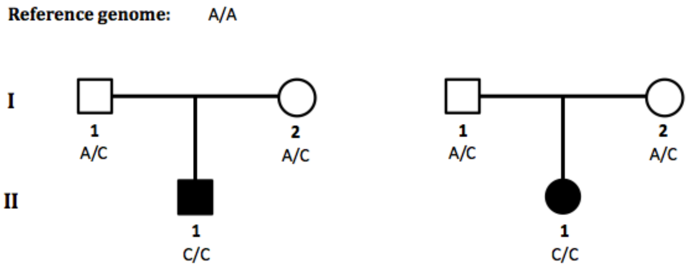


Figure 3. Example of pedigree for a given locus that could explain the mode of inheritance of the disease studied. Left trio 1 and right trio 2. Hypothetical example in which the reference genome for a given locus presented the alleles A/A, the parents were heterozygous for a variant allele A/C and the offspring is homozygous for the variant allele.

Schema 2. In this subgroup, any other type of inheritance pattern was included. These patterns of inheritance could not explain the mode of inheritance of the disease in our study. See some examples in Figure 4.

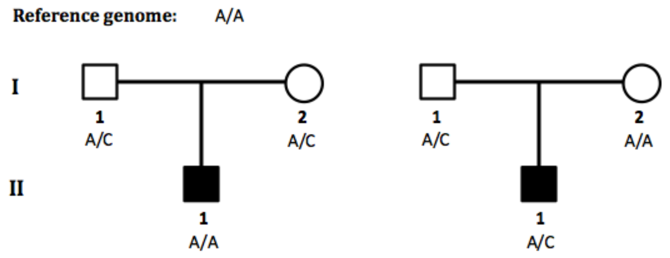


Figure 4. Examples of pedigrees for a given locus that could not explain the mode of inheritance of the disease studied. Left, parents are heterozygous for a variant and offspring is homozygous for the same alleles as the reference. Right, one of the parents was heterozygous for a variant allele and the offspring inherited the variant allele too.

Schema 3. The last subgroup contained those variants annotated in which one or more individuals from the trio did not present reads for that given locus. See examples in Figure 5.

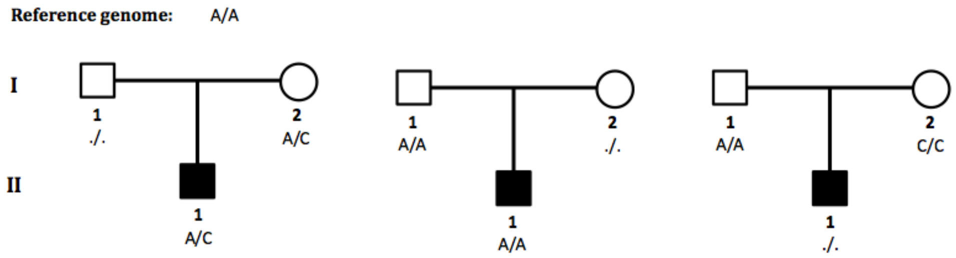


Figure 5. Examples of pedigrees in which one family member presented no data for a given locus. The three pedigrees show different examples where one of the family members did not present reads for the given locus.

Once the variants were divided in these three subgroups for SNPs and Indels in both of the trios, another custom-made Perl script was applied to check if the two analyzed trios shared the same variant as well as shared genes matching any of the previously described genes in dogs causing inherited retinal diseases (Miyadera et al. 2012) (Appendix B. 5.3).

In order to prioritize the analysis I focused on those variants which satisfied an AR inheritance (schema 1), and specifically, those variants annotated within an exonic region, that were further subdivided (See Appenix A. Table 3). To evaluate known function and disease association of genes, the GeneCards database (www.genecards.org) (Safran M. *et al.,* 2010) was used. Additionally, the genes with previously reported disease retinal genes in humans were checked in the RetNet database (http://sph.uth.edu/Retnet/).

2.7 Variant validation

In order to validate the detected variant associated with retinopathy in the investigated Labrador retriever family, two different genotyping methods were developed: Sanger Sequencing and fragment length polymorphism detection with capillary electrophoresis .

For both Sanger sequencing and length polymorphism detection with capillary electrophoresis, primers flanking the insertion c.4176insC (p.V1390fs) in the *ABCA4* gene were designed using Primer 3 (Untergasser et al. 2012). PCR reactions were performed according to the BigDye® Direct Cycle Sequencing Kit Protocol (Applied Biosystems®, Inc., [ABI], Foster City, USA) manufacturer's manual (Appendix A. Protocol 4). PCR reactions were performed in a ProFlex™ PCR System (Applied Biosystems®), later PCR-amplified fragments were purified with BigDye® Direct Cycle Sequencing Kit (Applied Biosystems®) following manufacturer's instructions. Templates were prepared to be sequenced both in forward and reverse orientation. PCR sequence products were analyzed on the ABI 3500XL Genetic Analyzer (Applied Biosystems®).

For genotyping based on fragment length polymorphism detection with capillary electrophoresis, fluorescent primers were (Cfa_ABCA4_FAM_F: 5'-Fam-CACCCACATTGCCATGTTTA-3' and Cfa_ABCA4_R: 5'-AACACATGGGGGTGAATGAT-3') were used for amplification on a ProFlex™ PCR System (Applied Biosystems,). PCR reactions were performed in a ProFlex™ PCR System. The PCR-products with an internal size standard (GeneScan™ 600 LIZ® dye Size Standard v2.0) were loaded and analyzed on a ABI 3500XL Genetic Analyzer and genotypes were called using the GeneMapper® Software (Applied Biosystems). The expected amplification length was 201 bp for the wild type allele and 202 bp for the mutant allele.

2.8 RNA extraction, cDNA synthesis and RT-PCR

RNA extraction from blood of control dog was performed using Tempus^{TM} Blood RNA Tube and Tempus^{TM} Spin RNA Isolation Kit (Applied Biosystems®), and following manufacturer's instructions. cDNA synthesis and RT-PCR was performed using the OneStep RT-PCR Kit (Qiagen) followings instructor's manual (See primers at Appendix A. Table 2).

3. Results

3.1 Clinical description

In a litter of eight Swedish Labrador retrievers, one female and one male were diagnosed with a PRA-like retinopathy at the age of nine and five years, respectively. Both the parents were tested negative for the disease-causing allele in the *PRCD* locus (Zangerl et al. 2006). An ophthalmic re-examination of the affected male offspring has been performed yearly and validated the diagnosis of a very slowly progressive, generalized, bilateral retinopathy.

3.2 Whole Genome Sequencing

3.2.1 Quantification and quality control of the libraries

Quantification of the DNA libraries showed that all the concentration of indexed DNA libraries were above 2nM (minimum required for high quality results). The final DNA libraries concentrations for both runs are presented in Appendix C. 1. Moreover, quality control of the libraries as well as the quality control from the DNA samples after the fragmentation step is shown in Appendix C. 2. Results showed that the peaks obtained were larger than expected.

3.2.2 NextSeq® 500 sequencing system performance

The sequencing system performance parameters from both runs are summarized in Table 2. Run 1 presented the flowcell ID: H5MT2BGXX and run 2: H5KWMBGXX, and the data is stored in the Animal Breeding and Genetics disk storage space.

Table 2. Summary of the NextSeq® 500 sequencing system performance parameters.

	Read length	**Total time**[1]	**Cluster density**[2]	**Estimated yield output**	**Clusters passing filter**[3]	**Quality scores**[4]
Run 1	2 x 101 bp	~ 20 h	165 K/mm^2	76.9 Gb	86.7 %	79.4 % > Q30
Run 2	2 x 150 bp	~ 29 h	185 K/mm^2	126.2 Gb	85.8%	81.7% > Q30

1. Total time of the run is subject to the read length set to sequence including cluster generation, sequencing and base calling.

2. Cluster density is dependent on the concentration of DNA inserted into the flowcell (ideally 200-210 k/mm^2). Too high cluster density can reduce the amount of data obtained due to cluster overlap.

3. Read whose cluster is sufficiently separated from other clusters in the flowcell. This filter is calculated for each cluster over the first 25 bases of the sequences.

4. Quality scores are predicted as the probability of an error in a base calling. The percentage of bases with quality over 30 is averaged across the run.

3.2.3 Reads: quality control, pre-processing and mapping

The quality control of the reads was performed with the software FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Sequence quality per base before and after trimming can be seen at Appendix C. 3. In general, more reads and higher quality was obtained in run 2. Table 1 from Appendix C shows the amount of reads before and after trimming as well as the number of mappable reads to the *Canis lupus familiaris* reference genome sequence, CanFam3.1. The fraction of aligned paired-end reads that was ~ 98 % in run 1 and 99 % in run 2. The average coverage obtained was 6.88 and 11.29 x per run 1 and 2, respectively.

When merging sequence data from both runs, the average genome sequence depth increased to ~18x per dog (Table 3).

3.2.4 Variant calling and filtering

An overview of the variant reduction process is illustrated in Figure 6 for run 1 and Figure 3 for the merged runs. Briefly, the first step of the variant calling process was performed using HaplotypeCaller software (DePristo et al. 2011), which resulted in ~ 880 M (run 1) and ~ 1 B (merged run) unfiltered SNPs and Indels calls for each sample, respectively. Afterwards, when all the samples were merged and new re-annotation was performed, the number of unfiltered SNPs and Indels calls was reduced to ~ 6,6 (run 1) and ~ 7,8 M (merged run) variants.

Table 3. Summary of the final data used for the bioinformatics analysis.

	Sample	Trimmed reads	Trimmed reads (x2)	Aligned reads (PE)	Genome Coverage
Merged reads	**Sire I:1**	180,425,877	360,851,754	357,599,644	18.33
	Dam I:2	194,182,364	388,364,728	385,367,484	19.68
	Offspring II:1	181,286,735	362,573,470	359,094,300	18.22
	Offspring II:2	164,408,226	328,816,452	324,449,983	16.39
	Total	720,303,202	1,440,606,404	1,426,511,411	18.16[1]

[1] Average coverage from all the samples.

The four dogs were then analyzed as two family trios where each trio included both parents and one of the offspring. Thus, the same raw variants were analyzed in the two different trios and used for further analysis. Then, the raw variants from each trio were separated into SNPs and Indels (same number of variants called for each trio was given) and after applying the recommended filters using the software VariantFiltration the number of variants called were slightly reduced. Last step consisted in annotating the variants with ANNOVAR.

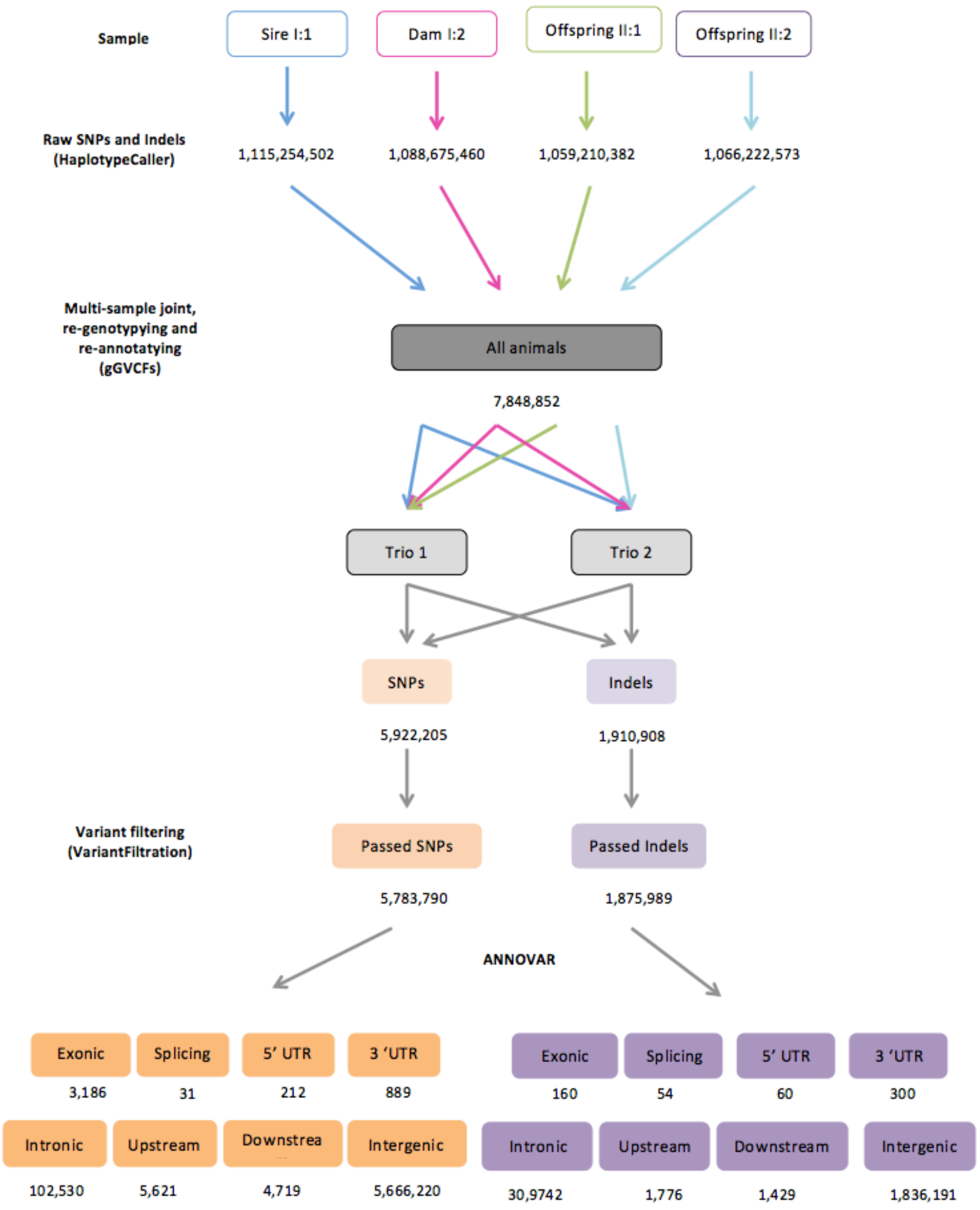


Figure 6. Schematic representation of the variant reduction process for merged runs. From raw variants called for each of the samples until the annotation of the variants that passed the filters. See text for explanation.

3.2.5 Analysis of the annotated variants

3.2.5.1 Variant filtration from run 1

A total of 5.7 M of SNPs and 1.8 M of Indel variants were annotated for both the analyzed trios. These variants were subdivided according to the genome annotation (*e.g*. exonic, intron) and also subdivided according to their conditional filtering schema. Results are shown in the tables from Appendix C. 4.1.

A conditional filtering schema was applied assuming an autosomal recessive (AR) mode of inheritance where both healthy parents were assumed to be heterozygous and the offspring in each trio was assumed to be homozygous. Results showed that none of the variants called within the exonic region that satisfied an AR inheritance pattern was shared between the two littermates. Each of the variations was analyzed but none had been previously described as a retinal inherited disease gene (Table 4).

Table 4. Shared variants found in the output of run 1 that presented an AR inheritance.

Variant region	SNPs	Indels
Exonic	0	0
Splicing	0	0
5' UTR	0	0
3' UTR	2 (*SCN5A, HIF1A*)	0
Upstream	12 *(CTNNB1, CYP4A38, RNASEL, BMPR1B, C31H21orf62, SLCO2A1, UBASH3A*[1]*, MIR487A, LHB, MIR578)*	3 (*SEP15, MMP7, DPT*)
Downstream	12 *(B3GAT1, MIR8862, MGST3, C31H21orf62*[1]*, KLRD1*[1]*, TFF2, NTF4*[2]*)*	0

1. Two different variants in this gene were found. **2.** Four different variants in this gene were found.

3.2.5.2 Variant filtration from the merged run1 and 2

Next, the merged data set from the two sequence runs was analyzed (Appendix C. 4.2). This resulted in the identification of an insertion in exon 28 of the *ATP-Binding Cassete, subfamily A* gene *(ABC1).* The insertion c.4176insC (p.F1393LfsX3) results in a frame shift that cause a premature stop codon at position 1395. The insertion was visualized in IGV, see Figure 7.

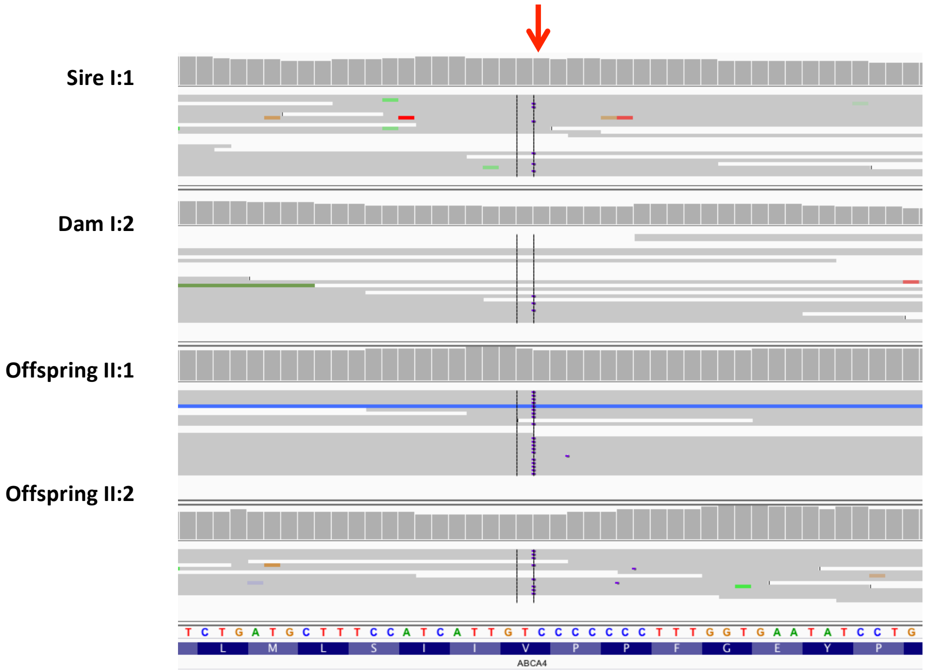

Sire I:1
Dam I:2
Offspring II:1
Offspring II:2
T C T G A T G C T T T C C A T C A T T G T C C C C C C C T T T G G T G A A T A T C C T G
L M L S I I V P P F G E Y P
ABCA4

Figure 7. Visualization of the insertion of a nucleotide in the different family members with IGV. Sire I:1 and Dam I:2 were carriers of the mutated allele (purple line). The number of reads that confirmed the presence of the mutated allele was 6 out of 17 for the sire and 3 out of 12 reads for the dam. Offspring II:1 and II:2 presented a read depth of the target variant of 21 x and 10 x respectively.

Despite the fact that the candidate mutation most likely was identified, 2 other variants annotated were annotated. The first was a variant annotated in an exonic region causing a nonsynonymous SNV: *FDX1,* but the amino acid change has a neutral effect on the prortein according to PolyPhen 2 (Adzhubei et al. 2010). The second variant was a splicing variant at *B3GAT1,* also with neutral effects.

3.3 Validation of the mutation

Sanger sequencing could validate the insertion of a nucleotide 'C' in the target position. As it can be seen in Figure 8.

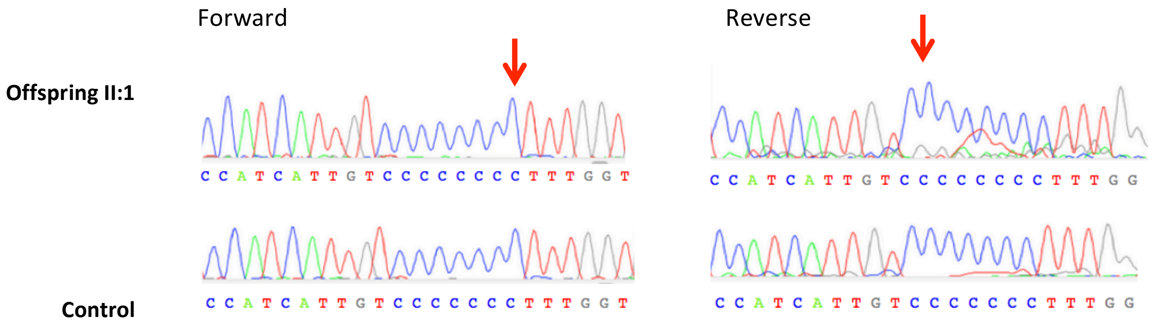


Figure 8. Visualization of the insertion with Sanger sequencing. The output from both forward- and reverse sequencing primers validated that the insertion is only found in the affected dog (Offspring II:1) and not in the control.

Additionally, the fragment length polymorphism detection method could also validate the insertion of a nucleotide in the affected dog, obtaining a fragment length polymorphism of 202 bp whilst the control dog presented an amplified fragment length polymorphism of 201 bp (images not shown).

3.4 RNA expression study

Amplification products from the RT-PCR didn’t show the expected fragment length (image not shown), thus is highly probable that the amplified products obtained are unwanted regions.

4. Discussion

4.1 Whole Genome Sequencing of a small pedigree of Labrador retriever reveals the mutation implicated to a rare form of retinopathy

To identify the genetic cause of a previously undescribed autosomal recessive form generalized, bilateral retinopathy, whole genome sequencing of a small pedigree of four Labrador retriever dogs was performed using the Illumina NextSeq500 platform-based. It was assumed that both the healthy parents were heterozygous carriers and that the two offspring's were homozygous for the disease causing allele.

In the first WGS experiment, a genome coverage of 6.5 x per individual was achieved and more than 5 x 10^6 (M) of SNPs and 1.5 M of Indels were found. However, this level of coverage appeared not to be sufficient to detect the causative mutation for this autosomal recessive disease in which the parents were heterozygous carriers for the mutation and the affected offspring was homozygous recessive for the mutation. The WGS experiment was repeated and when output data from the first and second sequencing rounds were merged and used for the analysis, 5.9 M of SNPs and 1.9 M of Indels with an average 18.1 x depth per animal was sufficient to be able to identify the mutation. Thus, having a relatively deep coverage is crucial for successfully identify causative mutations using this approach. Since none of the two sequencing runs was done with optimal cluster density (200-210 K/ mm^2), a higher yield than the obtained 77 Gb (first run) and 126 Gb (second run) could potentially have been achieved. Compared to the first run, the second run using the Illumina High Output kit v2 with 150bp PE reads from a library with a fragment size of 550bp produced significantly more sequence data and maintained the sequence quality.

For the bioinformatics analysis, computer resources provided by the Swedish National Infrastructure for Computing at the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) were used. The same bioinformatics workflow was used for the analysis of the output data from run 1 and for the merged runs. In the first run performed, the causal variant associated with the disease did not surpass any of the thresholds of confidence, bias or read position condition and might have been filtered out and consequently missed, whereas when the data from the merged runs was analyzed the causal variant was not passed over and was included in further steps of the analysis. Using both the adequate command lines as well as personalized perl scripts to analyze ANNOVAR output aided to speed up the bioinformatics analysis. The software ANNOVAR (Wang et al. 2010) gives by default two files: a file with all the variants included and a file with only the exonic variants (for SNPs and Indels and for both trios). Since these files had > 5 M SNPs and > 1.4 M Indels variants annotated, manual analysis of each of the variants would have been extremely tedious and quotidian software such as Microsoft Excel cannot deal with such big files. Thus, the scripts were created to satisfy a quickly and successful analysis of the two different format files given. In the end, after conditional filtering a single insertion variant was observed in both affected individuals that was predicted to result in frameshift and premature stop codon of the *ABCA4.*

4.2 The *ABCA4* gene

The *ABCA4* has already been evaluated and characterized in previous studies in dogs as a potential candidate to be associated with retinal diseases (Kijas et al. 2004; Lippmann et al. 2006), nevertheless, no genetic variants were found to be associated with a retinal disease. Thus, our project is the first study that reports a genetic variant associated to a canine retinopathy.

Belonging to the ATP binding cassette (ABC) transporter family, the *ABCA4* gene in dogs (Gene ID: 444852) comprises 50 exons covering 127,856 bp of genomic sequence transcribed from the forward strand of chromosome 6 (55,058,361..55,186,263). There are three alternative transcripts reported (Cunningham et al. 2015): *ABCA4-202* (ENSCAFT00000032029), *ABCA4-203* (ENSCAFT00000042929) and *ABCA4-201* (ENSCAFT00000005367), composed by 49, 47 and 51 coding exons, spanning_7,208, 6,985 and 6,897 bp and final protein products consisting of 2,268, 2,134 and 2,283 amino acid (aa) residues, respectively (Figure 9).

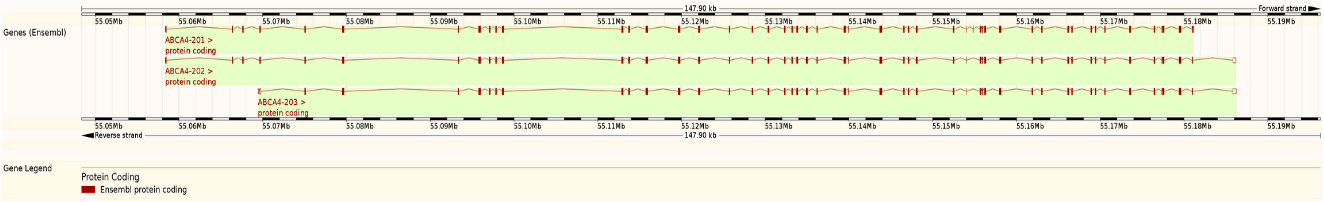


Figure 9. Transcripts from the *ABCA4* gene in dogs. Screenshot from Ensembl.

4.2.1 Introduction to the ABC Transporter Family

The ATP binding cassette (ABC) family of transporters is the largest known family of transmembrane (TM) proteins. In both eukaryotic and prokaryotic cells, the ABC transporters use energy originated from ATP hydrolysis to translocate a wide range of substrates across membranes, including ions, sugars and peptides (Dean and Allikmets 1995). Proteins are recognized as ABC transporters depending on their ATP binding domains, also known as Nucleotide Binding Domains (NBDs). A functional ABC transporter protein normally contains two NBDs and two TM domains. Each TM domain can be composed of six to 11 membrane-spanning α-helices that provide specificity for the substrates and form the translocation path, whilst the NBDs supply energy for the transport of the substrates.

The ABC proteins unidirectionally translocate different substrates. In bacteria the ABC transporters are mostly implicated in the import of basic substrates into the cell, whereas in eukaryotes, the compounds are transported from the cytoplasm to the exterior of the cell or into cellular vesicles (Vasiliou et al. 2009).

To date, 49 *ABC* genes have been annotated in the human genome, and they are divided into seven different subfamilies labeled from A – G, based on divergent evolution, gene structure and amino acid sequence similarity (Vasiliou et al. 2009). Since these family members are involved in crucial steps of many cellular processes, various diseases displaying Mendelian inheritance have been reported due to mutations occurring in some of the genes (Dean and Allikmets 1995). Subfamily A is composed of 12 proteins, mostly reported to be involved in lipid transport in several organs and cell types. The members of this family present two topologically similar halves with final protein products ranging in size from 1,543 (*ABCA10*) to 5,058 aa residues (*ABCA13*) (Cunningham et al. 2015). A distinguishing trait of members from the A family is the presence of a large extracellular domain in the N-terminal half. Mutations in this family have also been reported in humans, like the Tangier disease in *ABCA1* (Zarubica et al. 2007), harlequin type ichthyosis in *ABCA12* (Akiyama et al. 2005) and importantly for this study, the *ABCA4*, where a large number of different mutations responsible for several visual disorders in human have been reported, including the Stargardt disease (Allikmets et al. 1997).

4.2.2 Molecular view of *ABCA4*

4.2.2.1 Structure

The genomic sequence of the *ABCA4* gene is available from several species. This data can be used for sequence alignments, for example, identity scores range from 95 % (dog vs. cat) to 65 % of identity (dog vs. alpaca) (Cunningham et al. 2015). The high similarity can be explained since the primary structure is formed by conserved motifs (Figure 10). The primary structure is organized into two symmetrical but non-identical regions, named N-half (the region near the N-terminal of the polypeptide) and C-half (the region near the C-terminus), both halves contain a single transmembrane helix followed by an exocytoplasmic domain (ECD), next there is a transmembrane domain (TMD) with five membrane-spanning α -helices and followed by the nuclear binding domain (NBD). Conserved motifs are mostly found in the NBDs, they present the Walker A motif, the ABC signature motif and the Walker B. Not only the NBDs present preserved motifs, the TMD1 presents the EAA motif with unique differences among species (Mourez et al. 1997) suggesting that this ABC transporter act as an 'importer' (Rees et al. 2009) in divergence to the rest of eukaryotic ABC transporters, which have been shown to act as 'exporters'. In actual fact, not until the publication of the study of Quazi et al. (2012) was confirmed that ABCA4 functions as 'importer'. Furthermore, a highly conserved 'VFVNFA' motif close to the C-terminus has been reported to be essential for a correct folding of ABCA4 into a functional protein (Zhong et al. 2009). Last but not the least, there are three conserved single-residue motifs, the H-loop, A-loop and Q-loop (not shown in Figure B). The first two are involved in actual binding of the substrate whereas the function of the Q-loop is to transfer the energy produced by the NBDs to the TMDs (Davidson and Chen 2004).

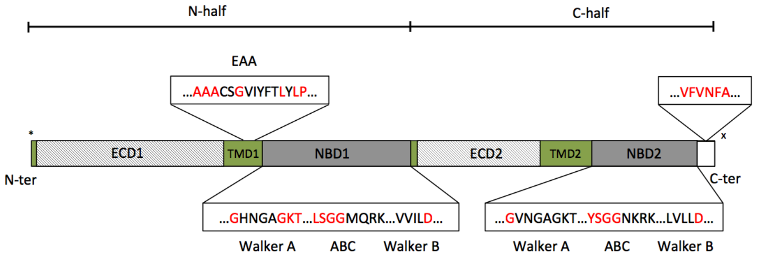


Figure 10. Figure of the primary structure of the *ABCA4* gene depicting conserved motifs. Transmembrane (TM) as well as transmembrane domains (TMDs) 1 and 2 are shown in green. Exocytoplasmic domains (ECDs) 1 and 2 are shown in light grey. Nucleotide Binding Domains (NBDs) 1 and 2 are shown in dark grey. '*' represents start codon and 'x' end codon. Figure modified from Tsybovsky et al. (2010).

4.2.2.2 Localization

ABCA4 was shown to be localized on the outer segment of cones and rods photoreceptor cells (Papermaster et al. 1978; Molday et al. 2000). In fact, the outer segment of the photoreceptor cells presents hundreds of flattened enclosed entities known as disks (Figure 11. A). ABCA4 is specifically found in the edge of these disks (Figure 11. B). Despite that the reason for this specific location remains unclear, it is hypothesized that this allocation of the protein is given due to the large exocytoplasmic domain 1 (ECD1), comprising up to 610 amino acid residues encoded by the dog transcript (Harpaz et al. 1994)

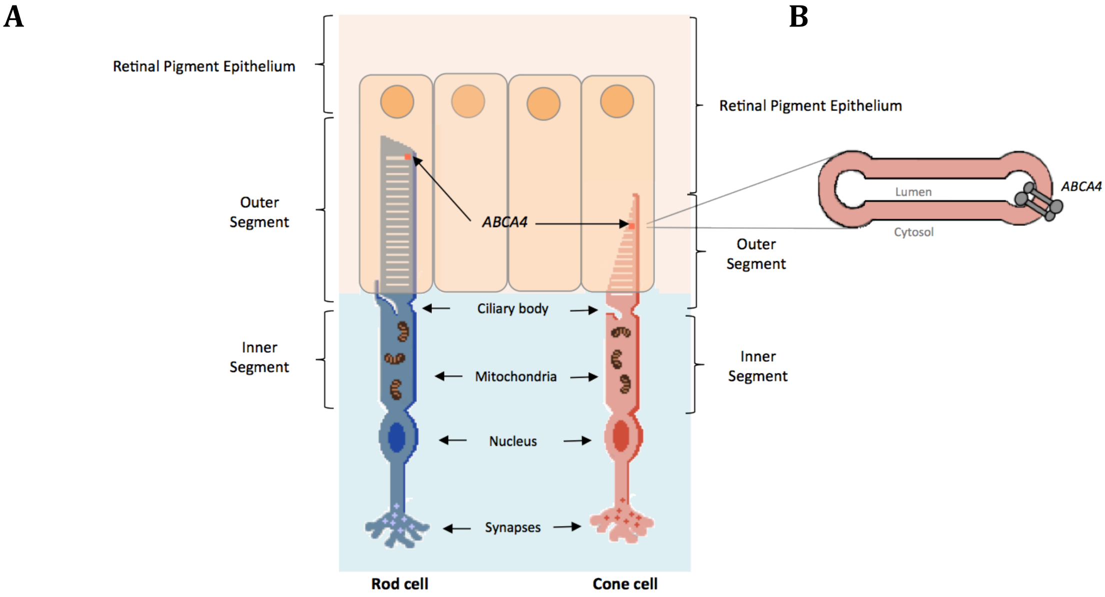


Figure 11. Localization of ABCA4. A. Schematic figure of rod and cone photoreceptor cells which are in contact with the retinal pigment epithelium monolayer of cells. The molecules of ABCA4 are localized in the disks situated in the outer segment of the cells. **B**. Zoom in of a cross-section from a disk. The ABCA4 membrane protein is localized in the rim of the disk.

4.2.2.3 Expression of ABCA4

For several years, expression of ABCA4 was thought to be exclusively in the retina (Azarian and Travis 1997). Nevertheless, posterior studies showed that expression of ABCA4 is also found in the brain, precisely located in the lateral ventricles of the rat brain (Bhongsatiern et al. 2005). Since then, mRNA expression of *ABCA4* in the brain has been studied in various mammals including humans, pigs and cows (Warren et al. 2009).

Hoeppner et al. (2014) provided a new and improved assembly of transcripts for the dog genome for coding and non-coding genes from different tissues types: blood, brain, heart, kidney, liver, lung, ovary, skeletal muscle, skin, and testis (BioProject: PRJNA78827; Accession: SRX111061 - SRX111071; SRX146606 - SRX146608). The data available from different tissue types was blasted against the *ABCA4* cDNA sequence (ENSCAFT00000032029).

Results might suggest that expression of *ABCA4* can be found in more tissues than the previously reported retina and brain. Even though gene expression data from canine retina tissue samples is not available up to date, thus no clear comparison of output obtained with blastn (Camacho et al. 2009) can be done, it seems that the expression of ABCA4 in brain is similar to its expression in kidney, ovary and testis, and in lower extent, in blood tissue (See blast alignments at Appendix C).

4.2.3 Mode of transport of ABCA4

ABCA4 is the only importer from the eukaryotic ABC transporter family (Quazi et al. 2012). Its proposed mode of transport is based then on previous knowledge obtained from comparisons with the prokaryotic type I ABC importers (Hollenstein et al. 2007). Known as the 'alternating access' or 'toppling' mechanism, it is suggested that when ATP binds to its ATP-binding site in the NBDs, bringing NBDs to close proximity until they form a dimer. This change induces a conformational rearrangement in the TMDs causing a closure of the high-affinity substrate-binding site (lumen side) which results in the translocation of the substrate to the low-affinity side of the membrane (cytoplasmic side). Then ATP is hydrolyzed, and ADP and inorganic phosphate (P_i) are released along with the dissociation of the NBDs, hence the transport cycle is complete (Figure 12) (Kos and Ford 2009; Slotboom 2014).

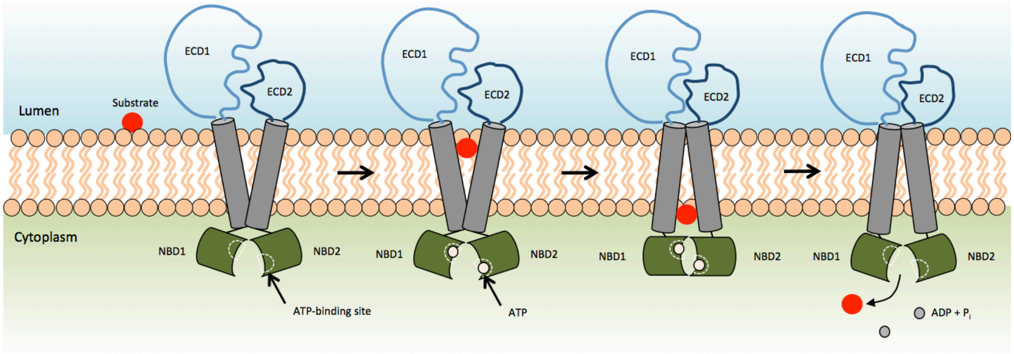


Figure 12. Schematic representation of the proposed 'alternating access' mechanism model. ABCA4 acts as an importer transporter. The figure shows the configurations of ABCA4 during the transport of the substrate (in red) across the membrane.

4.2.4 ABCA4 in the visual cycle

A choromophore undergoes thorough several transformations in the visual cycle. 11-*cis*-retinal is isomerized to all-*trans*-retinal when light absorption takes place. Following, all-*trans*-retinal is disassociated from rhodopsin and is converted to all-*trans*-retinol by retinol dehydrogenase (RDH) and transported from the cytoplasmic side to the retinal pigment epithelium (RPE) before entering to the visual cycle so as to regenerate 11-*cis*-retinal again (Figure 13) (Rees et al. 2009; Miyadera et al. 2012).

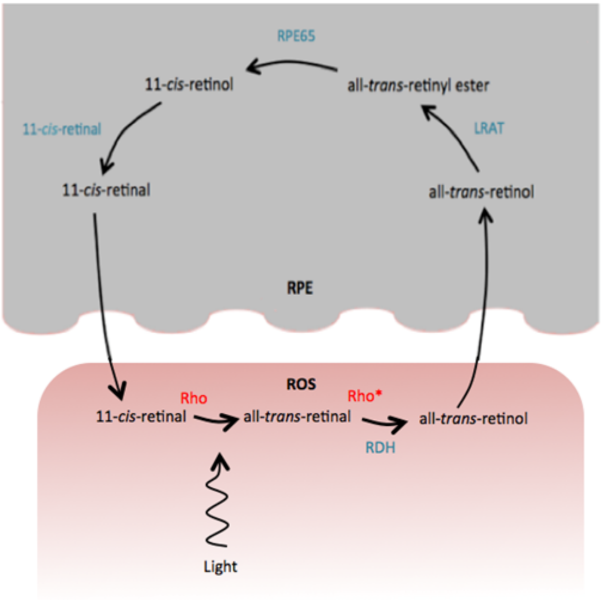


Figure 13. Schematic diagram of the chromophore in the retinoid cycle. Light absorption converts 11-*cis*-retinal bound to opsin (Rho) into all-*trans*-retinal, which will be reduced to all-*trans*-retinol by the photoreceptor retinol dehydrogenase (RDH). Then all-*trans*-retinol is exported from the retinal outer segment (ROS) to the retinal pigment epithelium (RPE), where it is esterified to all-*trans*-retinyl ester by lecithin retinol acyltransferase (LRAT). Afterwards, converted to 11-*cis*-retinol by RP65 and oxidized to 11-*cis*-retinal by 11-cis-retinol dehydrogenase and transported back to the outer segment to reassociate again with opsin.

ABCA4 acts as a critical active transporter in the visual cycle by flipping the *N*-retinylidene-PE, its preferable substrate through the photoreceptor disk membranes, from the extracellular to the cytoplasmic side (Quazi et al. 2012). *N*-retinylidene-PE is a reversible adduct formed between all-*trans*-retinal and phosphatidylethanolamine (PE) that is formed spontaneously and cannot diffuse across the membrane by itself (Figure 14). Thus, the role of ABCA4 consists on flipping the *N*-retinylidene-PE from the lumen to the cytoplasmic site where it is dissociated into PE and all-*trans*-retinal, that will re-enter in the visual cycle (Kiser et al. 2014).

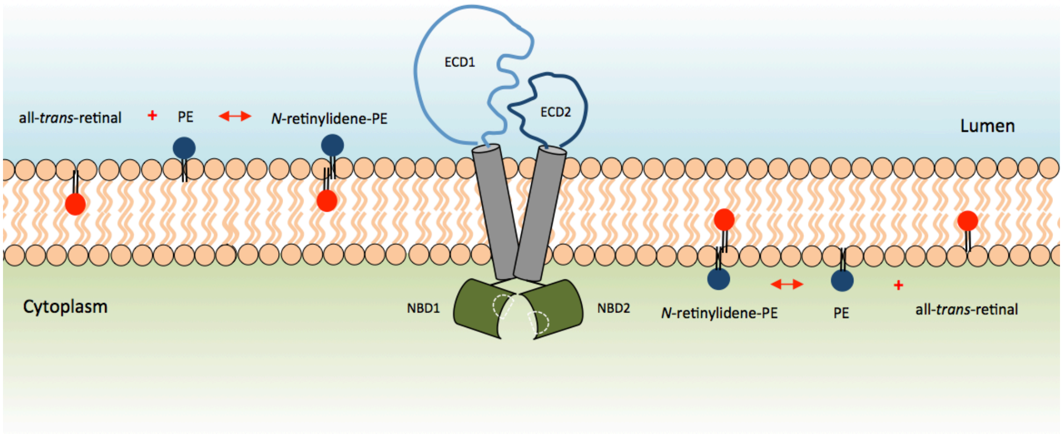

ECD1
ECD2
all-*trans*-retinal + PE ↔ *N*-retinylidene-PE
Lumen
Cytoplasm
NBD1
NBD2
N-retinylidene-PE ↔ PE + all-*trans*-retinal

Figure 14. Schematic figure showing the role of ABCA4 transporting *N*-retinylidene-PE across the OS disk membrane. ABCA4 is shown to function as a transporter of the *N*-retinylidene-PE adduct from the lumen to the cytoplasmic side, where *N*-retinylidene-PE will be dissociated into PE and all-*trans*-retinal to re-enter in the visual cycle.

When ABCA4 protein is defective, there is a progressive accumulation of *N*-retinylidene-PE on the lumen side of the membrane disk. This substrate can react with all-*trans*-retinal and produce di-retinoid-pyridinium-phosphatidylethanolamine (A2PE), which can be further hydrolyzed to phosphatidic acid (PA) and the di-retinal-pyridinium-ethanolamine (A2E), which cannot be further metabolized (Mata et al. 2000) (Figure 15). Then, the distal outer segment of the membrane is phagocytized by the RPE cells, and A2E and related bisretinoids are now accumulated to the RPE monolayer of cells (Sparrow and Boulton 2005). Since RPE cells are postmitotic, A2E cannot be diluted through cellular division and progressive aggregation of the lipofuscin will occur within the cell (Kevany and Palczewski 2010). The toxic effects of A2E on the RPE cells are produced through different mechanisms. The oxidation of the A2E triggers the activation of the complement cascade (Zhou et al. 2009). Additionally, A2E can block the cholesterol efflux from endosomes/lysosomes and subsequently cholesterol is accumulated in the RPE cells (Lakkaraju et al. 2007). A2E has also been reported to activate the retinoic acid receptor and stimulate pro-angiogenic factors (Iriyama et al. 2008). Additionally, it has been shown to destabilize the cell membranes (Sparrow et al. 2006), increase photo-damage of the cells in the blue-light wavelength (Sparrow et al. 2000) and inhibit respiration in mitochondria (Suter et al. 2000). A2E is also capable to inhibit the enzyme RPE65, a crucial enzyme in the visual cycle (see again Figure 13), decreasing the supply of 11-*cis*-retinal and causing a disrupted visual function (Moiseyev et al. 2010). Ultimately as a result of this process, RPE cells become atrophied and dies along with the adjacent photoreceptor cells, causing the vision loss.

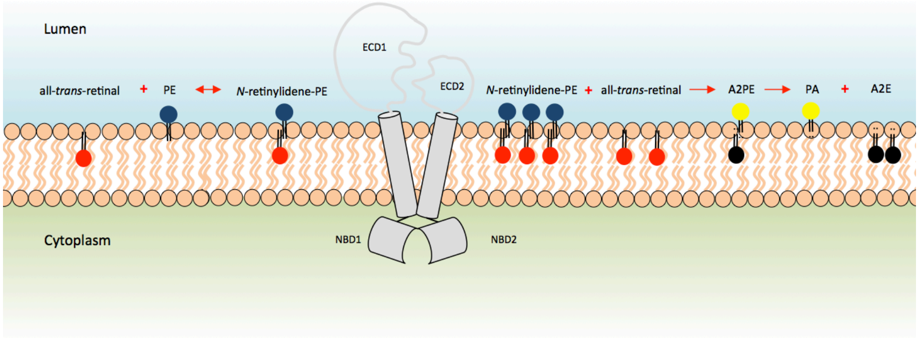

Lumen
ECD1
ECD2
all-*trans*-retinal + PE ↔ *N*-retinylidene-PE
N-retinylidene-PE + all-*trans*-retinal → A2PE → PA + A2E
Cytoplasm
NBD1
NBD2

Figure 15. Schematic representation showing the consequences if ABCA4 is defective. If the protein ABCA4 is defective *N*-retinylidene-PE can condense with all-*trans*-retinal in the lumen side of the disk forming A2PE.

4.2.5 *Abca4* knockout mice

Genetically engineered *Abca4*$^{-/-}$ knockout mice provided further insights into the role of ABCA4 even though it failed to fully reproduce all features of the human disease (Molday et al. 2009). The homozygous knockout mice showed no degradation of the photoreceptors under average light conditions, whereas extreme conditions caused significant degradation of the photoreceptor cells and showed as in many other studies, an increased accumulation of the lipofuscin A2E (Weng et al. 1999; Maeda et al. 2008; Issa et al. 2013). It was proven that *Abca4*$^{-/-}$ mice raised in total darkness did not accumulate the lipofuscin A2E or its precursors, ergo the severity is light dependent (Mata et al. 2000).

4.2.6 Mutations in *ABCA4* and vision diseases

Mutations in the *ABCA4* gene have been associated with diverse human retinal disorders (OMIM #601691) as Stargardt disease (STGD1) (Sun et al. 2000) with a prevalence of 1:10,000 (Walia and Fishman 2009), cone-rod dystrophy (CRD) (Ducroq et al. 2002), age related macular degeneration (AMD) (Allikmets et al. 1997) and retinitis pigmentosa (RP) (Allikmets et al. 1997). Over 800 different mutations have been associated with *ABCA4* (Zernant et al. 2014). Mutations include missense, nonsense, frameshift or splicing defects and up to 50 % are located in exons but no pathogenic variant is frequently associated to affected individuals (Zernant et al. 2014). In actual fact, two-disease associated alleles have been found in ~ 70 % of the cases, possibly making additive contributions and subsequently resulting in a wide spectrum of expression of the disease, from an early onset with severe retinopathy within the first decade of life to a mild retinopathy expressed at the fifth or later decade of life (Lewis et al. 1999; Zernant et al. 2011). Actually, the mechanisms in which the mutations lead to the disease remain unclear in a large number of the cases. Knowledge whether the protein has completely lost its function or if it presents toxic effects due to protein misfolding is lacking.

4.3 Pathogenicity of the insertion in exon 28 of the *ABCA4* gene

The insertion of a cytosine in the position 4176 of the coding sequence causes a frameshift mutation resulting in a stop codon two amino acid residues after the insertion. Taking into consideration the primary structure of ABCA4*,* the insertion of 'C' in exon 28 affects amino acids located in different locations of the ABCA4 structure, the end of the transmembrane domain and the beginning of the ECD2 (Figure 16.A). Specifically, the first and immediate amino acid change is Leucine instead of Phenylalanine, located in the last position of the transmembrane domain. Then, on the ECD2, the first amino acid coded is Tryptophan instead of Glycine and following, the stop codon (Figure 16.B-C). When the protein sequences from different organisms were aligned with ClustalW (Larkin et al. 2007), a high degree of sequence similarity was observed among species (Figure 16.C), as expected since the role of the gene is based on its structure.

The transcript containing the stop codon introduced by the p.F1393LfsX3 mutation, will most likely be degraded by the nonsense-mediated decay (NMD) pathway because the position of the premature stop codon fulfils all the criteria required for NMD (see below 4.3.1). If however, such transcripts were translated it would result in a truncated protein of 1,395 amino acid residues compared to the full-length ABCA4 protein (2,273 amino acid residues). The topology structure of ABCA4 and the location of the frameshift mutation is shown in Figure 17. Because a severely impaired ABCA4 protein only containing the N-terminal half would be obtained. In such case it is very unlikely that the protein could still function as an active and functional transporter.

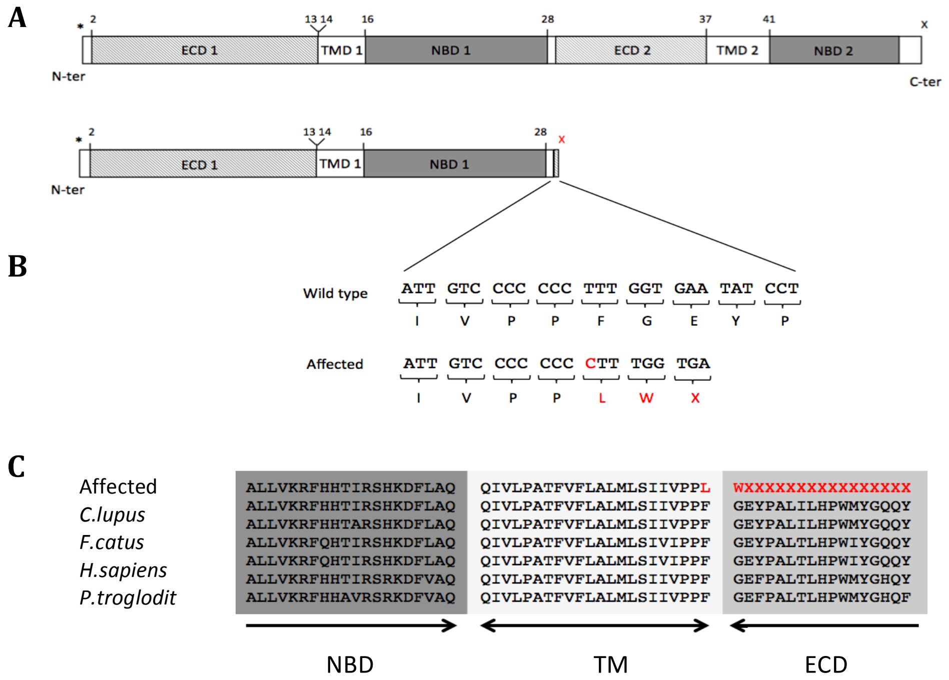

A
*
2
13 14
16
28
37
41
x
ECD 1
TMD 1
NBD 1
ECD 2
TMD 2
NBD 2
N-ter
C-ter
*
2
13 14
16
28
x
ECD 1
TMD 1
NBD 1
N-ter
B
Wild type
ATT GTC CCC CCC TTT GGT GAA TAT CCT
I V P P F G E Y P
Affected
ATT GTC CCC CCC CTT TGG TGA
I V P P L W X
C
Affected
C.lupus
F.catus
H.sapiens
P.troglodit
ALLVKRFHHTIRSHKDFLAQ QIVLPATFVFLALMLSIIVPPL WXXXXXXXXXXXXXXXX
ALLVKRFHHTIRSHKDFLAQ QIVLPATFVFLALMLSIIVPPF GEYPALILHPWMYGQQY
ALLVKRFHHTARSHKDFLAQ QIVLPATFVFLALMLSIIVPPF GEYPALILHPWMYGQQY
ALLVKRFQHTIRSHKDFLAQ QIVLPATFVFLALMLSIVIPPF GEYPALTLHPWIYGQQY
ALLVKRFQHTIRSHKDFLAQ QIVLPATFVFLALMLSIVIPPF GEYPALTLHPWIYGQQY
ALLVKRFHHTIRSRKDFVAQ QIVLPATFVFLALMLSIIVPPF GEFPALTLHPWMYGHQY
ALLVKRFHHAVRSRKDFVAQ QIVLPATFVFLALMLSIIVPPF GEFPALTLHPWMYGHQF
NBD
TM
ECD

Figure 16. Topological organization ABCA4 and conservation analysis. A. Schematic representation of the primary structure of ABCA4 including the exons. Upper, the full ABCA4 protein and below the truncated protein. '*' refers to start codon and 'x' to stop codon. **B.** Detailed schematic representation of the region where the mutation is found, the wild-type and the affected nucleotide and amino acids are compared. **C.** High sequence similarity among different species for a region close to the mutation can be seen.

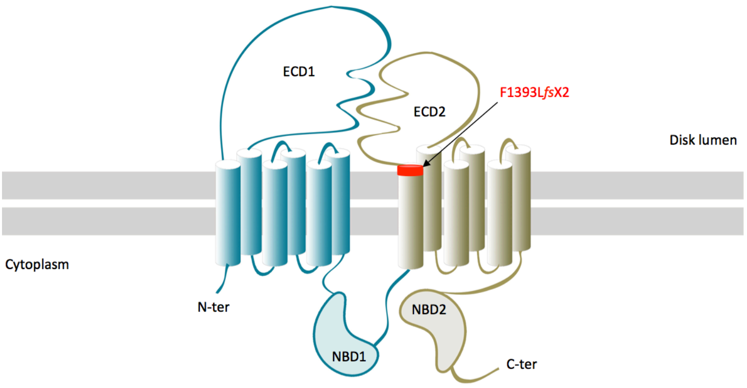


Figure 17. Topological structure of ABCA4 with the canine mutation indicated. The mutation is found in the last amino acid of the transmembrane domain and the frameshift is produced after two amino acids, corresponding to the ECD2.

4.3.1 Nonsense-mediated decay mechanism

A gene is transcribed to pre-messenger RNA by the RNA polymerase II, the pre-mRNA is processed by capping, splicing and polyadenylation to form a mature mRNA. This mRNA will be transported out of the nucleus and translated at the appropriate ribosomes. When a genetic variant results in a shortened protein, due to nonsense SNV or frameshift indel, as examples, it is known as a protein truncated variant (PTV). These variants have been described to be the cause of severe diseases (Holbrook et al. 2004; Stenson et al. 2014). PTVs can trigger the nonsense-mediated decay (NMD) mechanism. As a rule, if the PTV is located 50-55 nucleotides upstream of a splicing exon-exon junction the mRNA will go NMD (Nagy and Maquat 1998), whereas PTVs that escape NMD might create truncated proteins with gain of function or negative effects (Holbrook et al. 2004). In our study, the PTV formed due to c.4176insC in *ABCA4* is located in position 60 from the total of 125 bp of exon 28, meaning that the insertion is found 65 bases upstream the next exon-exon junction and consequently, it fulfils the criteria that trigger the NMD mechanism. Nonetheless, since exceptions of this rule have been reported (El-Bchiri et al. 2005), it is recommended to confirm whether the mRNA is a target of NMD by quantification of the level of nonsense-containing mRNA. Additionally, it will be interesting to study whether heterozygous dogs, i.e. carriers of the mutation, present any kind of dosage compensation or, as reported in the work of (Jensen 2014; Rivas et al. 2015), no evidence of up regulation was found and the expression levels of genes with PTV were found normal. Yet, both the parents of the affected dogs that are healthy carriers of the mutation should always be confirmed whether or not they are heterozygous for the mutation. Indeed this was the case for the parents analysed in our study.

4.4 Future research based on our work

I consider that compelling evidence has been obtained that the c.4176insC in *ABCA4* is the causative mutation of the retinal disease of the dogs studied in this project. However, information is lacking regarding the formation of a truncated ABCA4 protein or whether no protein at all is produced because NMD-dependent removal of the *ABCA4* mRNA containing the mutation. An RNA analysis for direct structural and expression analysis (including quantitative RT-PCR and/or RNA-Seq) would provide the missing information, but the impossibility of obtaining retinal tissue sample from *in vivo* affected dogs, where expression of

ABCA4 is the highest, makes the validation process more challenging. Further experiments would include using a specific knockout mouse or cultivation of cells with the c.4176insC mutation. Thus, for the knockout mouse, RNA extraction from retinal tissue could be used for direct structural and expression analysis as in Zhang et al. (2015). Nevertheless, if histological tissues also were to be examined, the protocol should be optimized in order to establish the time and intensity of light given as well as the age in which the animals would be studied. On the other hand, cultivation of cells could be another option, despite the fact that retinal cells would be the greatest candidates for performing histopathology or immunocytochemistry analysis using specific antibodies, they have been reported to be very difficult to grow. Cultivation of kidney or brain cells have been successful too (Bhongsatiern et al. 2005).

Very interestingly, when the Labrador retriever family pedigree was studied in detail, it was seen that there was a common ancestor five previous generations back, which could be the ancestral source of the mutation. This same sire has been extensively used for Labrador retriever breeding worldwide so chances are that there are more affected dogs with this retinopathy or carriers of the retinopathy, so development of a genetic test for the mutation that we have identified would be useful for Labrador retriever breeders worldwide. Then it would be possible to reveal if dogs carry the mutant allele or not could be very beneficial for breeding purposes, specially if they are descendant from the Bristish/Swedish Labrador retriever branch.

4.5 Future perspectives of the retinal disease

One of the unsolved issues that need to be assessed is the categorization of the disease. In humans, the hundreds of mutations found are widespread through the *ABCA4* gene, and the role of heterozygous variants still remains controversial. Whereas some have been associated to mild or later onset of the disease, others have been reported to produce more severe consequences even than homozygous variants producing a truncated protein. Thereby, further research is crucial to have a clear knowledge of the disease. How does the position of the mutation affect the disease development? And what about the position on those cases with heterozygous variants? Does the variant produce a functional protein? Does the variant cause a truncated mRNA? Is this mRNA eliminated by NMD or remain in the cell? If it remains in the cell, does it have major phenotypic consequences?

Up to date, there is no cure for the ABCA4-associated diseases. It is recommended to the patients to avoid excessive exposure of light as it was seen that *Abca4*$^{-/-}$ knockout mice didn't accumulate A2E when raised in darkness (Weng, Mata et al. 1999). Recently, the research target has become focus on *ABCA4* gene therapy. The major drawback resides on the success of gene delivery because of the large size of cDNA (6.8 Kb in humans). Different approaches have been carried out with varying results. The usage of the common lentivirus (LV) vectors showed that only 5-20 % of the photoreceptors were transduced and limited to the site of injection (Kong et al. 2008). Another approach used is to split an adeno-associated virus (AAV), known as AAV2, in which the transgene would be fragmented and packed in two halves, then both viruses have to infect the exactly same host cell and recombine (Han et al. 2014). The approach has already been successful *in vitro* and *in vivo* in the mice model (Allocca et al. 2008). Additionally, the usage of nanoparticles containing the large gene also got successful results in mice and could be a potential treatment too (Han et al. 2012).

5. Conclusions

Whole Genome Sequencing proved to be a successful method to study diseases harbouring unknown mutations and promising studies are yet to come. Here, we report for the first time a mutation in *ABCA4* gene associated to a retinal disease in Labrador retriever, homolog to Stargadt's disease in humans. Up to date, hundreds of mutations spread along the *ABCA4* have been described in humans. Yet, a clear categorization of the mutations still needs to be done. In dogs, the insertion of a cytidine in exon 28 (c.4176insC) causes an immediate frameshift and a premature stop codon after 2 amino acids, most likely targeting the mRNAs containing this premature stop codon for the Nonsense-mediated decay (NMD) pathway and then no protein product would be obtained. If however, such mRNAs are translated they would code for a truncated protein of 1,394 amino acids from total of 2,273 amino acids, thus, translating half of the flippase protein structure. In both cases, the lack of protein or lack of properly functional ABCA4 protein would cause an accumulation of toxic lipofuscin in the retinal pigment epithelium (RPE) cells causing atrophy and cell death and subsequently, vision loss in the affected dogs. In order to validate the production of protein product an RNA study needs to be performed, preferably from samples of retinal tissue. Such samples were however not yet available during the time of this project.

6. References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* **7**: 248-249.

Akiyama M, Sugiyama-Nakagiri Y, Sakai K, McMillan JR, Goto M, Arita K, Tsuji-Abe Y, Tabata N, Matsuoka K, Sasaki R et al. 2005. Mutations in lipid transporter ABCA12 in harlequin ichthyosis and functional recovery by corrective gene transfer. *J Clin Invest* **115**: 1777-1784.

Allikmets R, Singh N, Sun H, Shroyer NF, Hutchinson A, Chidambaram A, Gerrard B, Baird L, Stauffer D, Peiffer A et al. 1997. A photoreceptor cell-specific ATP-binding transporter gene (ABCR) is mutated in recessive Stargardt macular dystrophy. *Nat Genet* **15**: 236-246.

Allocca M, Doria M, Petrillo M, Colella P, Garcia-Hoyos M, Gibbs D, Kim SR, Maguire A, Rex TS, Di Vicino U et al. 2008. Serotype-dependent packaging of large genes in adeno-associated viral vectors results in effective gene delivery in mice. *J Clin Invest* **118**: 1955-1964.

Andersson L. 2001. Genetic dissection of phenotypic diversity in farm animals. *Nat Rev Genet* **2**: 130-138.

Axelsson E, Ratnakumar A, Arendt ML, Maqbool K, Webster MT, Perloski M, Liberg O, Arnemo JM, Hedhammar A, Lindblad-Toh K. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* **495**: 360-364.

Azarian SM, Travis GH. 1997. The photoreceptor rim protein is an ABC transporter encoded by the gene for recessive Stargardt's disease (ABCR). *FEBS Lett* **409**: 247-252.

Bhongsatiern J, Ohtsuki S, Tachikawa M, Hori S, Terasaki T. 2005. Retinal-specific ATP-binding cassette transporter (ABCR/ABCA4) is expressed at the choroid plexus in rat brain. *J Neurochem* **92**: 1277-1280.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.

Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. 2013. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet* **14**: 681-691.

Breen M, Hitte C, Lorentzen TD, Thomas R, Cadieu E, Sabacan L, Scott A, Evanno G, Parker HG, Kirkness EF et al. 2004. An integrated 4249 marker FISH/RH map of the canine genome. *BMC Genomics* **5**: 65.

Breen M, Jouquand S, Renier C, Mellersh CS, Hitte C, Holmes NG, Cheron A, Suter N, Vignaux F, Bristow AE et al. 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**: 1784-1795.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S et al. 2015. Ensembl 2015. *Nucleic Acids Res* **43**: D662-669.

Darwin C. 1871. *On the origin of species*. D. Appleton and Co., New York :.

Davidson AL, Chen J. 2004. ATP-binding cassette transporters in bacteria. *Annu Rev Biochem* **73**: 241-268.

Dean M, Allikmets R. 1995. Evolution of ATP-binding cassette transporter genes. *Curr Opin Genet Dev* **5**: 779-785.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491-498.

Ding ZL, Oskarsson M, Ardalan A, Angleby H, Dahlgren LG, Tepeli C, Kirkness E, Savolainen P, Zhang YP. 2012. Origins of domestic dog in southern East Asia is supported by analysis of Y-chromosome DNA. *Heredity (Edinb)* **108**: 507-514.

Ducroq D, Rozet JM, Gerber S, Perrault I, Barbet D, Hanein S, Hakiki S, Dufier JL, Munnich A, Hamel C et al. 2002. The ABCA4 gene in autosomal recessive cone-rod dystrophies. *Am J Hum Genet* **71**: 1480-1482.

El-Bchiri J, Buhard O, Penard-Lacronique V, Thomas G, Hamelin R, Duval A. 2005. Differential nonsense mediated decay of mutated mRNAs in mismatch repair deficient colorectal cancers. *Hum Mol Genet* **14**: 2435-2442.

Guyon R, Lorentzen TD, Hitte C, Kim L, Cadieu E, Parker HG, Quignon P, Lowe JK, Renier C, Gelfenbeyn B et al. 2003. A 1-Mb resolution radiation hybrid map of the canine genome. *Proc Natl Acad Sci U S A* **100**: 5296-5301.

Han Z, Conley SM, Makkia RS, Cooper MJ, Naash MI. 2012. DNA nanoparticle-mediated ABCA4 delivery rescues Stargardt dystrophy in mice. *J Clin Invest* **122**: 3221-3226.

Han Z, Conley SM, Naash MI. 2014. Gene therapy for Stargardt disease associated with ABCA4 gene. *Adv Exp Med Biol* **801**: 719-724.

Harpaz Y, Gerstein M, Chothia C. 1994. Volume changes on protein folding. *Structure* **2**: 641-649.

Hitte C, Madeoy J, Kirkness EF, Priat C, Lorentzen TD, Senger F, Thomas D, Derrien T, Ramirez C, Scott C et al. 2005. Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping. *Nat Rev Genet* **6**: 643-648.

Hoeppner MP, Lundquist A, Pirun M, Meadows JR, Zamani N, Johnson J, Sundstrom G, Cook A, FitzGerald MG, Swofford R et al. 2014. An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. *PLoS One* **9**: e91172.

Holbrook JA, Neu-Yilik G, Hentze MW, Kulozik AE. 2004. Nonsense-mediated decay approaches the clinic. *Nat Genet* **36**: 801-808.

Hollenstein K, Frei DC, Locher KP. 2007. Structure of an ABC transporter in complex with its binding protein. *Nature* **446**: 213-216.

Iriyama A, Fujiki R, Inoue Y, Takahashi H, Tamaki Y, Takezawa S, Takeyama K, Jang WD, Kato S, Yanagi Y. 2008. A2E, a pigment of the lipofuscin of retinal pigment epithelial cells, is an endogenous ligand for retinoic acid receptor. *J Biol Chem* **283**: 11947-11953.

Issa PC, Barnard AR, Singh MS, Carter E, Jiang ZC, Radu RA, Schraermeyer U, MacLaren RE. 2013. Fundus Autofluorescence in the Abca4(-/-) Mouse Model of Stargardt Disease-Correlation With Accumulation of A2E, Retinal Function, and Histology. *Invest Ophth Vis Sci* **54**: 5602-5612.

Jensen P. 2014. Behavior genetics and the domestication of animals. *Annu Rev Anim Biosci* **2**: 85-104.

Karlsson EK, Baranowska I, Wade CM, Salmon Hillbertz NH, Zody MC, Anderson N, Biagi TM, Patterson N, Pielberg GR, Kulbokas EJ, 3rd et al. 2007. Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat Genet* **39**: 1321-1328.

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

Katsanis SH, Katsanis N. 2013. Molecular genetic testing and the future of clinical genomics. *Nat Rev Genet* **14**: 415-426.

Kevany BM, Palczewski K. 2010. Phagocytosis of retinal rod and cone photoreceptors. *Physiology (Bethesda)* **25**: 8-15.

Kijas JW, Zangerl B, Miller B, Nelson J, Kirkness EF, Aguirre GD, Acland GM. 2004. Cloning of the canine ABCA4 gene and evaluation in canine cone-rod dystrophies and progressive retinal atrophies. *Mol Vis* **10**: 223-232.

Kiser PD, Golczak M, Palczewski K. 2014. Chemistry of the retinoid (visual) cycle. *Chem Rev* **114**: 194-232.

Kong J, Kim SR, Binley K, Pata I, Doi K, Mannik J, Zernant-Rajang J, Kan O, Iqball S, Naylor S et al. 2008. Correction of the disease phenotype in the mouse model of Stargardt disease by lentiviral gene therapy. *Gene Ther* **15**: 1311-1320.

Kos V, Ford RC. 2009. The ATP-binding cassette family: a structural perspective. *Cell Mol Life Sci* **66**: 3111-3126.

Lakkaraju A, Finnemann SC, Rodriguez-Boulan E. 2007. The lipofuscin fluorophore A2E perturbs cholesterol metabolism in retinal pigment epithelial cells. *Proc Natl Acad Sci U S A* **104**: 11026-11031.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947-2948.

Larson G, Karlsson EK, Perri A, Webster MT, Ho SY, Peters J, Stahl PW, Piper PJ, Lingaas F, Fredholm M et al. 2012. Rethinking dog domestication by integrating genetics, archeology, and biogeography. *Proc Natl Acad Sci U S A* **109**: 8878-8883.

Lewis RA, Shroyer NF, Singh N, Allikmets R, Hutchinson A, Li Y, Lupski JR, Leppert M, Dean M. 1999. Genotype/Phenotype analysis of a photoreceptor-specific ATP-binding cassette transporter gene, ABCR, in Stargardt disease. *Am J Hum Genet* **64**: 422-434.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Li R, Mignot E, Faraco J, Kadotani H, Cantanese J, Zhao B, Lin X, Hinton L, Ostrander EA, Patterson DF et al. 1999. Construction and characterization of an eightfold redundant dog genomic bacterial artificial chromosome library. *Genomics* **58**: 9-17.

Lindblad-Toh K Wade CM Mikkelsen TS Karlsson EK Jaffe DB Kamal M Clamp M Chang JL Kulbokas EJ, 3rd Zody MC et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**: 803-819.

Lingaas F, Sorensen A, Juneja RK, Johansson S, Fredholm M, Wintero AK, Sampson J, Mellersh C, Curzon A, Holmes NG et al. 1997. Towards construction of a canine linkage map: establishment of 16 linkage groups. *Mamm Genome* **8**: 218-221.

Lippmann T, Pasternack SM, Kraczyk B, Dudek SE, Dekomien G. 2006. Indirect exclusion of four candidate genes for generalized progressive retinal atrophy in several breeds of dogs. *J Negat Results Biomed* **5**: 19.

Maeda A, Maeda T, Golczak M, Palczewski K. 2008. Retinopathy in mice induced by disrupted all-trans-retinal clearance. *J Biol Chem* **283**: 26684-26693.

Mata NL, Weng J, Travis GH. 2000. Biosynthesis of a major lipofuscin fluorophore in mice and humans with ABCR-mediated retinal and macular degeneration. *Proc Natl Acad Sci U S A* **97**: 7154-7159.

Maxam AM, Gilbert W. 1977. A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74**: 560-564.

Mellersh CS, Langston AA, Acland GM, Fleming MA, Ray K, Wiegand NA, Francisco LV, Gibbs M, Aguirre GD, Ostrander EA. 1997. A linkage map of the canine genome. *Genomics* **46**: 326-336.

Miyadera K, Acland GM, Aguirre GD. 2012. Genetic and phenotypic variations of inherited retinal diseases in dogs: the power of within- and across-breed studies. *Mamm Genome* **23**: 40-61.

Moiseyev G, Nikolaeva O, Chen Y, Farjo K, Takahashi Y, Ma JX. 2010. Inhibition of the visual cycle by A2E through direct interaction with RPE65 and implications in Stargardt disease. *P Natl Acad Sci USA* **107**: 17551-17556.

Molday LL, Rabin AR, Molday RS. 2000. ABCR expression in foveal cone photoreceptors and its role in Stargardt macular dystrophy. *Nat Genet* **25**: 257-258.

Molday RS, Zhong M, Quazi F. 2009. The role of the photoreceptor ABC transporter ABCA4 in lipid transport and Stargardt macular degeneration. *Biochim Biophys Acta* **1791**: 573-583.

Mourez M, Hofnung M, Dassa E. 1997. Subunit interactions in ABC transporters: a conserved sequence in hydrophobic membrane proteins of periplasmic permeases defines an important site of interaction with the ATPase subunits. *EMBO J* **16**: 3066-3077.

Nagy E, Maquat LE. 1998. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem Sci* **23**: 198-199.

Neff MW, Broman KW, Mellersh CS, Ray K, Acland GM, Aguirre GD, Ziegle JS, Ostrander EA, Rine J. 1999. A second-generation genetic linkage map of the domestic dog, Canis familiaris. *Genetics* **151**: 803-820.

Neff MW, Rine J. 2006. A fetching model organism. *Cell* **124**: 229-231.

Ostrander EA, Wayne RK. 2005. The canine genome. *Genome Res* **15**: 1706-1716.

Pang JF, Kluetsch C, Zou XJ, Zhang AB, Luo LY, Angleby H, Ardalan A, Ekstrom C, Skollermo A, Lundeberg J et al. 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**: 2849-2864.

Papermaster DS, Schneider BG, Zorn MA, Kraehenbuhl JP. 1978. Immunocytochemical localization of a large intrinsic membrane protein to the incisures and margins of frog rod outer segment disks. *J Cell Biol* **78**: 415-425.

Patterson DF. 2000. Companion animal medicine in the age of medical genetics. *J Vet Intern Med* **14**: 1-9.

Quazi F, Lenevich S, Molday RS. 2012. ABCA4 is an N-retinylidene-phosphatidylethanolamine and phosphatidylethanolamine importer. *Nat Commun* **3**: 925.

Rees DC, Johnson E, Lewinson O. 2009. ABC transporters: the power to change. *Nat Rev Mol Cell Biol* **10**: 218-227.

Rivas MA, Pirinen M, Conrad DF, Lek M, Tsang EK, Karczewski KJ, Maller JB, Kukurba KR, DeLuca DS, Fromer M et al. 2015. Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* **348**: 666-669.

Salmon Hillbertz NH, Isaksson M, Karlsson EK, Hellmen E, Pielberg GR, Savolainen P, Wade CM, von Euler H, Gustafson U, Hedhammar A et al. 2007. Duplication of FGF3, FGF4, FGF19 and ORAOV1 causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. *Nat Genet* **39**: 1318-1320.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M. 1977. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**: 687-695.

Slotboom DJ. 2014. Structural and mechanistic insights into prokaryotic energy-coupling factor transporters. *Nat Rev Microbiol* **12**: 79-87.

Sparrow JR, Boulton M. 2005. RPE lipofuscin and its role in retinal pathobiology. *Exp Eye Res* **80**: 595-606.

Sparrow JR, Cai B, Jang YP, Zhou J, Nakanishi K. 2006. A2E, a fluorophore of RPE lipofuscin, can destabilize membrane. *Adv Exp Med Biol* **572**: 63-68.

Sparrow JR, Nakanishi K, Parish CA. 2000. The lipofuscin fluorophore A2E mediates blue light-induced damage to retinal pigmented epithelial cells. *Invest Ophthalmol Vis Sci* **41**: 1981-1989.

Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* **133**: 1-9.

Sun H, Smallwood PM, Nathans J. 2000. Biochemical defects in ABCR protein variants associated with human retinopathies. *Nat Genet* **26**: 242-246.

Suter M, Reme C, Grimm C, Wenzel A, Jaattela M, Esser P, Kociok N, Leist M, Richter C. 2000. Age-related macular degeneration. The lipofusion component N-retinyl-N-retinylidene ethanolamine detaches proapoptotic proteins from mitochondria and induces apoptosis in mammalian retinal pigment epithelial cells. *J Biol Chem* **275**: 39625-39630.

Thalmann O, Shapiro B, Cui P, Schuenemann VJ, Sawyer SK, Greenfield DL, Germonpre MB, Sablin MV, Lopez-Giraldez F, Domingo-Roura X et al. 2013. Complete mitochondrial genomes of ancient canids suggest a European origin of domestic dogs. *Science* **342**: 871-874.

Tsybovsky Y, Molday RS, Palczewski K. 2010. The ATP-binding cassette transporter ABCA4: structural and functional properties and role in retinal disease. *Adv Exp Med Biol* **703**: 105-125.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3-new capabilities and interfaces. *Nucleic Acids Res* **40**.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J et al. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **11**: 11 10 11-11 10 33.

van El CG, Cornel MC, Borry P, Hastings RJ, Fellmann F, Hodgson SV, Howard HC, Cambon-Thomsen A, Knoppers BM, Meijers-Heijboer H et al. 2013. Whole-genome sequencing in health care. Recommendations of the European Society of Human Genetics. *Eur J Hum Genet* **21 Suppl 1**: S1-5.

Vasiliou V, Vasiliou K, Nebert DW. 2009. Human ATP-binding cassette (ABC) transporter family. *Hum Genomics* **3**: 281-290.

Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Rosengren Pielberg G, Sigurdsson S, Fall T, Seppala EH, Hansen MS, Lawley CT et al. 2011. Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. *PLoS Genet* **7**: e1002316.

Venter JC Adams MD Myers EW Li PW Mural RJ Sutton GG Smith HO Yandell M Evans CA Holt RA et al. 2001. The sequence of the human genome. *Science* **291**: 1304-1351.

Vignaux F, Hitte C, Priat C, Chuat JC, Andre C, Galibert F. 1999. Construction and optimization of a dog whole-genome radiation hybrid panel. *Mamm Genome* **10**: 888-894.

Vila C, Savolainen P, Maldonado JE, Amorim IR, Rice JE, Honeycutt RL, Crandall KA, Lundeberg J, Wayne RK. 1997. Multiple and ancient origins of the domestic dog. *Science* **276**: 1687-1689.

Walia S, Fishman GA. 2009. Natural history of phenotypic changes in Stargardt macular dystrophy. *Ophthalmic Genet* **30**: 63-68.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**: e164.

Warren MS, Zerangue N, Woodford K, Roberts LM, Tate EH, Feng B, Li C, Feuerstein TJ, Gibbs J, Smith B et al. 2009. Comparative gene expression profiles of ABC transporters in brain microvessel endothelial cells and brain in five species including human. *Pharmacol Res* **59**: 404-413.

Watson JD, Crick FH. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**: 737-738.

Weng J, Mata NL, Azarian SM, Tzekov RT, Birch DG, Travis GH. 1999. Insights into the function of Rim protein in photoreceptors and etiology of Stargardt's disease from the phenotype in abcr knockout mice. *Cell* **98**: 13-23.

Werner P, Mellersh CS, Raducha MG, DeRose S, Acland GM, Prociuk U, Wiegand N, Aguirre GD, Henthorn PS, Patterson DF et al. 1999. Anchoring of canine linkage groups with chromosome-specific markers. *Mamm Genome* **10**: 814-823.

Wilcox B, Walkowicz C. 1995. *Atlas of dog breeds of the world*. TFH Publications ;

Distributed in the U.S. to the bookstore and library trade by National Book Network, Neptune City, NJ

Zangerl B, Goldstein O, Philp AR, Lindauer SJ, Pearce-Kelling SE, Mullins RF, Graphodatsky AS, Ripoll D, Felix JS, Stone EM et al. 2006. Identical mutation in a novel retinal gene causes progressive rod-cone degeneration in dogs and retinitis pigmentosa in humans. *Genomics* **88**: 551-563.

Zarubica A, Trompier D, Chimini G. 2007. ABCA1, from pathology to membrane function. *Pflugers Arch* **453**: 569-579.

Zernant J, Schubert C, Im KM, Burke T, Brown CM, Fishman GA, Tsang SH, Gouras P, Dean M, Allikmets R. 2011. Analysis of the ABCA4 gene by next-generation sequencing. *Invest Ophthalmol Vis Sci* **52**: 8479-8487.

Zernant J, Xie YA, Ayuso C, Riveiro-Alvarez R, Lopez-Martinez MA, Simonelli F, Testa F, Gorin MB, Strom SP, Bertelsen M et al. 2014. Analysis of the ABCA4 genomic locus in Stargardt disease. *Hum Mol Genet* **23**: 6797-6806.

Zhang N, Tsybovsky Y, Kolesnikov AV, Rozanowska M, Swider M, Schwartz SB, Stone EM, Palczewska G, Maeda A, Kefalov VJ et al. 2015. Protein misfolding and the pathogenesis of ABCA4-associated retinal degenerations. *Hum Mol Genet* **24**: 3220-3237.

Zhong M, Molday LL, Molday RS. 2009. Role of the C terminus of the photoreceptor ABCA4 transporter in protein folding, function, and retinal degenerative diseases. *J Biol Chem* **284**: 3640-3649.

Zhou J, Kim SR, Westlund BS, Sparrow JR. 2009. Complement activation by bisretinoid constituents of RPE lipofuscin. *Invest Ophthalmol Vis Sci* **50**: 1392-1399.

Appendix A. Laboratory protocols

Protocol 1. Protocol: General Purification Protocol (Qiagen®). QIAsymphony DNA Handbook 09/2012 (pag. 18-20).

1. Close all drawers and the hood.

2. Switch on the QIAsymphony SP, and wait until the "Sample Preparation" screen appears and the initialization procedure has finished.

3. Log on to the instrument.

4. Ensure the "Waste" drawer is loaded properly, and perform an inventory scan of the "Waste" drawer, including the tip chute and liquid waste.

5. Load the required elution rack into the "Eluate" drawer.

6. Load the required reagent cartridge(s) and consumables into the "Reagents and Consumables" drawer.

7. Perform an inventory scan of the "Reagents and Consumables" drawer.

8. Place the samples into the appropriate sample carrier, and load them into the "Sample" drawer.

9. For Virus Blood applications: The tube(s) containing the internal control–Buffer

1. ATE mixture should be placed in slot A of the "Sample" drawer.

10. Using the touchscreen, enter the required information for each batch of samples to be processed.

11. Press the "Run" button to start the purification procedure.

2. All processing steps are fully automated. At the end of the protocol run, the status of the batch changes from "RUNNING" to "COMPLETED".

12. Retrieve the elution rack containing the purified nucleic acids from the "Eluate" drawer. The DNA is ready to use or can be stored at 2–8°C, –20°C, or –80°C.

Protocol 2. DNA extraction from sperm samples using QIAsymphony (Qiagen®). Alexander Falk 07/2013.

1. Cut off the plugs in one end of the straw, place it in a marked eppendorf tube with the uncut end up. Then cut the other end of the straw letting the sperm go down into the tube. This step might take some time.
2. Transfer 100 µl of sperm to a new eppendorf tube (free the rest of the sperm).
3. Add 400 µl of PBS and vortex for 10 s.
4. Centrifuge for 30 s at 1300 rpm (14900 x g).
5. Remove the supernatant carefully.
6. Resuspend the pellet by adding 200 µl of Qiagen buffer ATL and 25 µl of DTT 1 M. DTT is made by solving 4,62 mg of DTT powder in 30 µl sterile DNA free ddH$_2$O.
7. Vortex until the pellet is completely dissolved. You can make the pellet detach from the tube by pipetting.
8. When the pellet is suspended add 25 µl of Proteinase K (20 mg/ml) and vortex gently.
9. Wrap the tube lid with Para film, incubate at 56°C for 2-3 h or overnight until the pellet is completely dissolved.
10. After the incubation, transfer the eppendof content into a Sarstedt 2 ml screw skirt tube. You will need the same tube for DNA elution. Then use the QIAsymphony machine and use the Mini Kit.

Protocol 3. KAPA Library Quantification Kit for Illumina® platforms (Kapa Biosystems, Inc., Wilmington, MA) 07/2014 (Kits: KK4835; KK4906)

The library quantification was performed according to manufacturer's instructions. Since the platform that we use is the StepOnePlus™ Real-Time PCR System (Life Technologies™), the reaction setup was: 12 µl 2X KAPPA SYBR® FAST qPCR Master Mix + 10X Primers Premix; 4 µl PCR-grade water. Furthermore, Standard 0 was used as a library quantification control. Libraries quantification was performed at concentrations 1:5,000, 1:10,000 and 1:20,0000 and three replicas of each (using the same master mix) were performed.

Protocol 4. BigDye® Direct Cycle Sequencing Kit (Applied Biosystems®, Foster City, USA) 02/2011 (Rev. C)

Prepare and run the PCR reactions:

1. For each forward or reverse reaction, add the components to an appropriate reaction plate:

    Genomic DNA (4 ng/µL) 1.0 µL

    M13-tailed PCR primer mix (0.8 µM each primer) 1.5 µL

    BigDye® Direct PCR Master Mix 5.0 µL

    Deionized water 2.5 µ

2. Pipet up and down to mix well, seal the plate with adhesive film or caps, then spin the plate briefly.

3. Run the reactions in a thermal cycler (See Material and Methods)

4. Store the amplified DNA at 4°C overnight or at –15°C or –25°C for long-term storage.

Perform cycle sequencing:

1. Prepare a forward or reverse sequencing reaction mix in a tube on ice:

    BigDye® Direct Sequencing Master Mix 2.0 µL

    One sequencing primer: 1.0 µL

    - BigDye® Direct M13 Fwd Primer **or**
    - BigDye® Direct M13 Rev Primer

2. Seal the reaction plate with adhesive film or caps, then spin the plate briefly.

3. Run the reactions in a thermal cycler: 15 min at 37 °C; 2 min at 80 °C; 1 min at 96 °C; 25 cycles of 10 s at 96 °C, 5 s at 50 °C and 75 s at 60 °C.

The primers were synthesized by TAG Copenhagen A/S, Frederiksberg, Denmark with M13 sequencing tails (Cfa_ABCA4_Frw 5’- TGT AAA ACG ACG GCC AGT CAC CCA CAT TGC CAT GTT TA-3’ and Cfa_ABCA4_Rev 5’-CAG GAA ACA GCT ATG ACC AAC ACA TGG GGG TGA ATG AT-3’).

Protocol 5. Fragment separation. Swedish University of Agricultural Sciences. 05/2015

1. Prepare a Master mix with the following (x 1 reaction):

   10x HotStar PCR Buffer 1 µL

   dNTP 25 mM 0.096 µL

   Forward primer 0.05 µL

   Reverse primer 0.05 µL

   HotStar Taq 0.12 µL

   H_2O 8.19 µL

2. Run the reactions in a thermal cycler: 10 min at 94 °C; 29 cycles of 1 min at 94 °C, 1 min at 60 °C and 2 min at 72 °C; and a final extension of 10 min at 72 °C.
3. Add 1 µL formaldehide.

Protocol 6. RNA extraction and cDNA synthesis

RNA extraction was perfromed using Tempus™ Blood RNA Tube and Tempus™ Spin RNA Isolation Kit (Applied Biosystems®). cDNA synthesis and RT-PCR was performed using the OneStep RT-PCR Kit (Qiagen). The PCR reaction with a final volume of 50 µl contained: 10 µl of QIAGEN OneStep RT-PCR Buffer 5 X, 2 µl of dNTPs (10mM), 3 µl of each of the primers (10 µM) (See Table RT-PCR primers), 2 µl of the QIAGEN OneStep RT-PCR Enzyme, 1 µl of RNA and 29 µl of water. The PCR thermocycler conditions were: 30 min at 54 °C; 15 min at 95 °C; 35 cycles of 1 m at 94 °C, 1 m at 58 °C and 1 m at 72 °C; and a final extension of 10 min at 72 °C.

Table 1. Index adapter sequences used for the samples of the study

Family member	Adapter	Sequence
Sire I:1	AD002	CGATGT
Dam I:2	AD004	TGACCA
Offspring II:1	AD005	ACAGTG
Offspring II:2	AD006	GCCAAT

Different indexes of 6 bp were used for each sample in both runs. The indexed adapter sequences belong to the TruSeq LT Kit Set A Indexed Adapter Sequences for DNA Kits (Illumina).

Primers RT-PCR

Primers used for the RT-PCR

Gene	Exons	Primer Sequence (5' → 3')	Product size (bp)
ABCA4	2	ATTCGCTTTGTGGTGGAACT	
	3	ATTCTCCCGGGGTAGGATTT	236
	27	CAAGCGATTCCACCACACTA	
	28	TACTGCTGCCCATACATCCA	379
	30	CAATTCAACCCCTTGGAAGA	
	31	TTCAGTGCTGCGCTGTATTC	282
GAPDH		TCCTGCACCACCAACTGCTT	
		GTCTTCTGGGTGGCAGTGAT	334

Table 3. ANNOVAR annotation

Annotation	Explanation
frameshift insertion	an insertion of one or more nucleotides that cause frameshift changes in protein coding sequence
frameshift deletion	a deletion of one or more nucleotides that cause frameshift changes in protein coding sequence
frameshift block substitution	a block substitution of one or more nucleotides that cause frameshift changes in protein coding sequence
stopgain	a nonsynonymous SNV, frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that lead to the immediate creation of stop codon at the variant site.
stoploss	a nonsynonymous SNV, frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that lead to the immediate elimination of stop codon at the variant site
nonframeshift insertion	an insertion of 3 or multiples of 3 nucleotides that do not cause frameshift changes in protein coding sequence
nonframeshift deletion	a deletion of 3 or mutliples of 3 nucleotides that do not cause frameshift changes in protein coding sequence
nonframeshift block substitution	a block substitution of one or more nucleotides that do not cause frameshift changes in protein coding sequence
nonsynonymous SNV	a single nucleotide change that cause an amino acid change
synonymous SNV	a single nucleotide change that does not cause an amino acid change
unknown	unknown function (due to various errors in the gene structure definition in the database file)

Information extracted from http://annovar.openbioinformatics.org/en/latest/user-guide/gene/.

Appendix B. Bioinformatics scripts

1. Generation of FastQ files, merging, quality control and pre-processing of the reads

```
#!/bin/bash
#SBATCH -A b2015069
#SBATCH -p core
#SBATCH -n 6
#SBATCH -t 1-20:00:00

#Create working directory.
#mkdir /proj/b2015069/labbe_retinopathy
#mkdir /proj/b2015069/labbe_retinopathy/raw_bcl_data_labbe

#----------
#Bcl conversion and demultiplexing with bcl2fastq.
#SampleSheet.csv has to be in the run folder.

module load bioinfo-tools
module load bioinfo-tools bcl2fastq/2.15.0

bcl2fastq --runfolder-dir /proj/b2015069/labbe_retinopathy/raw_bcl_data_labbe/150330_NS500636_0005_AH5KWMBGXX -
p 6

wait

#mkdir /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe
#mkdir /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/reads_per_lane

cp
/proj/b2015069/labbe_retinopathy/raw_bcl_data_labbe/150330_NS500636_0005_AH5KWMBGXX/Data/Intensities/BaseCall
s/*.gz /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/reads_per_lane

#Concatenate forward reads of each animal instead of having them by lanes.
#Also the reverse reads.
cat /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/reads_per_lane/sire_S1_*_R1_*.fastq.gz >
/proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/sire_S1_fwd_fastq.gz &
cat /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/reads_per_lane/sire_S1_*_R2_*.fastq.gz >
/proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/sire_S1_rev_fastq.gz &
cat /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/reads_per_lane/dam_S2_*_R1_*.fastq.gz >
/proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/dam_S2_fwd_fastq.gz &
cat /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/reads_per_lane/dam_S2_*_R2_*.fastq.gz >
/proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/dam_S2_rev_fastq.gz &
cat /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/reads_per_lane/off1_S3_*_R1_*.fastq.gz >
/proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/off1_S3_fwd_fastq.gz &
cat /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/reads_per_lane/off1_S3_*_R2_*.fastq.gz >
/proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/off1_S3_rev_fastq.gz &
cat /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/reads_per_lane/off2_S4_*_R1_*.fastq.gz >
/proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/off2_S4_fwd_fastq.gz &
cat /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/reads_per_lane/off2_S4_*_R2_*.fastq.gz >
/proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/off2_S4_rev_fastq.gz &

#Check quality with FastQC.
mkdir /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/first_fastqc
mkdir /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/first_fastqc/sire
mkdir /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/first_fastqc/dam
mkdir /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/first_fastqc/off1
mkdir /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/first_fastqc/off2

module load bioinfo-tools
module load bioinfo-tools FastQC/0.11.2

fastqc -t 6 -o /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/first_fastqc/sire \
      /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/sire_*.gz

fastqc -t 6 -o /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/first_fastqc/dam \
      /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/dam_*.gz

fastqc -t 6 -o /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/first_fastqc/off1 \
      /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/off1_*.gz

fastqc -t 6 -o /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/first_fastqc/off2 \
      /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/off2_*.gz

wait

#------------
#!/bin/bash
#SBATCH -A b2015069
#SBATCH -p core
#SBATCH -n 5
#SBATCH -t 1-10:00:00

#Preprocessing
#mkdir /proj/b2015069/labbe_retinopathy/trimming

#Trimmomatic
module load bioinfo-tools
module load bioinfo-tools trimmomatic/0.32
module load bioinfo-tools java/sun_jdk1.7.0_25

java -jar -Xmx10g /sw/apps/bioinfo/trimmomatic/0.32/milou/trimmomatic-0.32.jar PE -threads 5 -phred33 \
              /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/sire_S1_fwd_fastq.gz \
              /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/sire_S1_rev_fastq.gz \
              /proj/b2015069/labbe_retinopathy/trimming/paired_sire_fwd.fq.gz \
              /proj/b2015069/labbe_retinopathy/trimming/unpaired_sire_fwd.fq.gz \
              /proj/b2015069/labbe_retinopathy/trimming/paired_sire_rev.fq.gz \
              /proj/b2015069/labbe_retinopathy/trimming/unpaired_sire_rev.fq.gz \
              ILLUMINACLIP:/proj/b2015069/labbe_retinopathy/trimming/adapters.fasta:2:30:15 LEADING:20
SLIDINGWINDOW:40:20 MINLEN:30

java -jar -Xmx10g /sw/apps/bioinfo/trimmomatic/0.32/milou/trimmomatic-0.32.jar PE -threads 5 -phred33 \
              /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/dam_S2_fwd_fastq.gz \
              /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/dam_S2_rev_fastq.gz \
              /proj/b2015069/labbe_retinopathy/trimming/paired_dam_fwd.fq.gz \
              /proj/b2015069/labbe_retinopathy/trimming/unpaired_dam_fwd.fq.gz \
              /proj/b2015069/labbe_retinopathy/trimming/paired_dam_rev.fq.gz \
              /proj/b2015069/labbe_retinopathy/trimming/unpaired_dam_rev.fq.gz \
              ILLUMINACLIP:/proj/b2015069/labbe_retinopathy/trimming/adapters.fasta:2:30:15 LEADING:20
SLIDINGWINDOW:40:20 MINLEN:30

java -jar -Xmx10g /sw/apps/bioinfo/trimmomatic/0.32/milou/trimmomatic-0.32.jar PE -threads 5 -phred33 \
              /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/off1_S3_fwd_fastq.gz \
              /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/off1_S3_rev_fastq.gz \
              /proj/b2015069/labbe_retinopathy/trimming/paired_off1_fwd.fq.gz \
              /proj/b2015069/labbe_retinopathy/trimming/unpaired_off1_fwd.fq.gz \
              /proj/b2015069/labbe_retinopathy/trimming/paired_off1_rev.fq.gz \
              /proj/b2015069/labbe_retinopathy/trimming/unpaired_off1_rev.fq.gz \
              ILLUMINACLIP:/proj/b2015069/labbe_retinopathy/trimming/adapters.fasta:2:30:15 LEADING:20
SLIDINGWINDOW:40:20 MINLEN:30

java -jar -Xmx10g /sw/apps/bioinfo/trimmomatic/0.32/milou/trimmomatic-0.32.jar PE -threads 5 -phred33 \
              /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/off2_S4_fwd_fastq.gz \
        /proj/b2015069/labbe_retinopathy/raw_fastq_data_labbe/off2_S4_rev_fastq.gz \
        /proj/b2015069/labbe_retinopathy/trimming/paired_off2_fwd.fq.gz \
        /proj/b2015069/labbe_retinopathy/trimming/unpaired_off2_fwd.fq.gz \
        /proj/b2015069/labbe_retinopathy/trimming/paired_off2_rev.fq.gz \
        /proj/b2015069/labbe_retinopathy/trimming/unpaired_off2_rev.fq.gz \
        ILLUMINACLIP:/proj/b2015069/labbe_retinopathy/trimming/adapters.fasta:2:30:15 LEADING:20
SLIDINGWINDOW:40:20 MINLEN:30

wait

#Check in fastqc again after trimming

mkdir /proj/b2015069/labbe_retinopathy/trimming/fastqc
mkdir /proj/b2015069/labbe_retinopathy/trimming/fastqc/sire
mkdir /proj/b2015069/labbe_retinopathy/trimming/fastqc/dam
mkdir /proj/b2015069/labbe_retinopathy/trimming/fastqc/off1
mkdir /proj/b2015069/labbe_retinopathy/trimming/fastqc/off2

module load bioinfo-tools FastQC/0.11.2

fastqc -t 5 -o /proj/b2015069/labbe_retinopathy/trimming/fastqc/sire \
        /proj/b2015069/labbe_retinopathy/trimming/paired_sire_*.gz

fastqc -t 5 -o /proj/b2015069/labbe_retinopathy/trimming/fastqc/dam \
        /proj/b2015069/labbe_retinopathy/trimming/paired_dam_*.gz

fastqc -t 5 -o /proj/b2015069/labbe_retinopathy/trimming/fastqc/off1 \
        /proj/b2015069/labbe_retinopathy/trimming/paired_off1_*.gz

fastqc -t 5 -o /proj/b2015069/labbe_retinopathy/trimming/fastqc/off2 \
        /proj/b2015069/labbe_retinopathy/trimming/paired_off2_*.gz


#----------
#Unzip the files with the reads after trimmomatic.
gunzip /proj/b2015069/labbe_retinopathy/trimming/paired_*.gz

#Merge reads from both runs: per animal and per direction.
mkdir /proj/b2015069/labbe_retinopathy/merged

cat /proj/b2015069/labbe_retinopathy/trimming/paired_sire_fwd.fq
/proj/b2015069/WGS1_labbe_retinopathy/trimming/paired_sire_fwd.fq >
/proj/b2015069/labbe_retinopathy/merged/merged_reads_sire_fwd.fq &

cat /proj/b2015069/labbe_retinopathy/trimming/paired_sire_rev.fq
/proj/b2015069/WGS1_labbe_retinopathy/trimming/paired_sire_rev.fq >
/proj/b2015069/labbe_retinopathy/merged/merged_reads_sire_rev.fq &

cat /proj/b2015069/labbe_retinopathy/trimming/paired_dam_fwd.fq
/proj/b2015069/WGS1_labbe_retinopathy/trimming/paired_dam_fwd.fq >
/proj/b2015069/labbe_retinopathy/merged/merged_reads_dam_fwd.fq &

cat /proj/b2015069/labbe_retinopathy/trimming/paired_dam_rev.fq
/proj/b2015069/WGS1_labbe_retinopathy/trimming/paired_dam_rev.fq >
/proj/b2015069/labbe_retinopathy/merged/merged_reads_dam_rev.fq &

cat /proj/b2015069/labbe_retinopathy/trimming/paired_off1_fwd.fq
/proj/b2015069/WGS1_labbe_retinopathy/trimming/paired_off1_fwd.fq >
/proj/b2015069/labbe_retinopathy/merged/merged_reads_off1_fwd.fq &

cat /proj/b2015069/labbe_retinopathy/trimming/paired_off1_rev.fq
/proj/b2015069/WGS1_labbe_retinopathy/trimming/paired_off1_rev.fq >
/proj/b2015069/labbe_retinopathy/merged/merged_reads_off1_rev.fq &
```

```
cat /proj/b2015069/labbe_retinopathy/trimming/paired_off2_fwd.fq
/proj/b2015069/WGS1_labbe_retinopathy/trimming/paired_off2_fwd.fq >
/proj/b2015069/labbe_retinopathy/merged/merged_reads_off2_fwd.fq &

cat /proj/b2015069/labbe_retinopathy/trimming/paired_off2_rev.fq
/proj/b2015069/WGS1_labbe_retinopathy/trimming/paired_off2_rev.fq >
/proj/b2015069/labbe_retinopathy/merged/merged_reads_off2_rev.fq &

#Count number of reads from the first run, second run, and the merged.
echo "Stats reads" > merged_stats.txt
echo "###########" >> merged_stats.txt
#Sire
echo "FWD" >> merged_stats.txt
echo "Number of reads WGS1 (350 bp)" >> merged_stats.txt
grep "@" /proj/b2015069/WGS1_labbe_retinopathy/trimming/paired_sire_fwd.fq | wc -l >> merged_stats.txt
echo "Number of reads WGS2 (550 bp)" >> merged_stats.txt
grep "@" /proj/b2015069/labbe_retinopathy/trimming/paired_sire_fwd.fq | wc -l >> merged_stats.txt
echo "Number of merged reads aligned to the reference genome for Sire:" >> merged_stats.txt
grep "@" /proj/b2015069/labbe_retinopathy/merged/merged_reads_sire_fwd.fq | wc -l >> merged_stats.txt
echo "REV" >> merged_stats.txt
echo "Number of reads WGS1 (350 bp)" >> merged_stats.txt
grep "@" /proj/b2015069/WGS1_labbe_retinopathy/trimming/paired_sire_rev.fq | wc -l >> merged_stats.txt
echo "Number of reads WGS2 (550 bp)" >> merged_stats.txt
grep "@" /proj/b2015069/labbe_retinopathy/trimming/paired_sire_rev.fq | wc -l >> merged_stats.txt
echo "Number of merged reads aligned to the reference genome for Sire:" >> merged_stats.txt
grep "@" /proj/b2015069/labbe_retinopathy/merged/merged_reads_sire_rev.fq | wc -l >> merged_stats.txt
echo "----------" >> merged_stats.txt
#Dam
echo "FWD" >> merged_stats.txt
echo "Number of reads WGS1 (350 bp)" >> merged_stats.txt
grep "@" /proj/b2015069/WGS1_labbe_retinopathy/trimming/paired_dam_fwd.fq | wc -l >> merged_stats.txt
echo "Number of reads WGS2 (550 bp)" >> merged_stats.txt
grep "@" /proj/b2015069/labbe_retinopathy/trimming/paired_dam_fwd.fq | wc -l >> merged_stats.txt
echo "Number of merged reads aligned to the reference genome for Dam:" >> merged_stats.txt
grep "@" /proj/b2015069/labbe_retinopathy/merged/merged_reads_dam_fwd.fq | wc -l >> merged_stats.txt
echo "REV" >> merged_stats.txt
echo "Number of reads WGS1 (350 bp)" >> merged_stats.txt
grep "@" /proj/b2015069/WGS1_labbe_retinopathy/trimming/paired_dam_rev.fq | wc -l >> merged_stats.txt
echo "Number of reads WGS2 (550 bp)" >> merged_stats.txt
grep "@" /proj/b2015069/labbe_retinopathy/trimming/paired_dam_rev.fq | wc -l >> merged_stats.txt
echo "Number of merged reads aligned to the reference genome for Dam:" >> merged_stats.txt
grep "@" /proj/b2015069/labbe_retinopathy/merged/merged_reads_dam_rev.fq | wc -l >> merged_stats.txt
echo "----------" >> merged_stats.txt
# Off1
echo "FWD" >> merged_stats.txt
echo "Number of reads WGS1 (350 bp)" >> merged_stats.txt
grep "@" /proj/b2015069/WGS1_labbe_retinopathy/trimming/paired_off1_fwd.fq | wc -l >> merged_stats.txt
echo "Number of reads WGS2 (550 bp)" >> merged_stats.txt
grep "@" /proj/b2015069/labbe_retinopathy/trimming/paired_off1_fwd.fq | wc -l >> merged_stats.txt
echo "Number of merged reads aligned to the reference genome for Off1:" >> merged_stats.txt
grep "@" /proj/b2015069/labbe_retinopathy/merged/merged_reads_off1_fwd.fq | wc -l >> merged_stats.txt
echo "REV" >> merged_stats.txt
echo "Number of reads WGS1 (350 bp)" >> merged_stats.txt
grep "@" /proj/b2015069/WGS1_labbe_retinopathy/trimming/paired_off1_rev.fq | wc -l >> merged_stats.txt
echo "Number of reads WGS2 (550 bp)" >> merged_stats.txt
grep "@" /proj/b2015069/labbe_retinopathy/trimming/paired_off1_rev.fq | wc -l >> merged_stats.txt
echo "Number of merged reads aligned to the reference genome for Off1:" >> merged_stats.txt
grep "@" /proj/b2015069/labbe_retinopathy/merged/merged_reads_off1_rev.fq | wc -l >> merged_stats.txt
echo "----------" >> merged_stats.txt
# Off2
echo "FWD" >> merged_stats.txt
echo "Number of reads WGS1 (350 bp)" >> merged_stats.txt
```

```
grep "@" /proj/b2015069/WGS1_labbe_retinopathy/trimming/paired_off2_fwd.fq | wc -l >> merged_stats.txt
echo "Number of reads WGS2 (550 bp)" >> merged_stats.txt
grep "@" /proj/b2015069/labbe_retinopathy/trimming/paired_off2_fwd.fq | wc -l >> merged_stats.txt
echo "Number of merged reads aligned to the reference genome for Off2:" >> merged_stats.txt
grep "@" /proj/b2015069/labbe_retinopathy/merged/merged_reads_off2_fwd.fq | wc -l >> merged_stats.txt
echo "REV" >> merged_stats.txt
echo "Number of reads WGS1 (350 bp)" >> merged_stats.txt
grep "@" /proj/b2015069/WGS1_labbe_retinopathy/trimming/paired_off2_rev.fq | wc -l >> merged_stats.txt
echo "Number of reads WGS2 (550 bp)" >> merged_stats.txt
grep "@" /proj/b2015069/labbe_retinopathy/trimming/paired_off2_rev.fq | wc -l >> merged_stats.txt
echo "Number of merged reads aligned to the reference genome for Off2:" >> merged_stats.txt
grep "@" /proj/b2015069/labbe_retinopathy/merged/merged_reads_off2_rev.fq | wc -l >> merged_stats.txt
echo "----------" >> merged_stats.txt
```

2. Prepare the dog reference genome (CanFam 3.1) for use with BWA and GATK

```
#Getting ready the reference genome
module load bioinfo-tools samtools
module load bioinfo-tools java/sun_jdk1.7.0_25
module load bwa

#Generation BWA index
bwa index /proj/b2015069/canfam3.fasta

#Generate Fasta file index
samtools faidx /proj/b2015069/canfam3.fasta

#Generate sequence dictionary
java -Xmx4g /sw/apps/bioinfo/picard/1.92/milou/CreateSequenceDictionary.jar \
        REFERENCE=/proj/b2015069/canfam3.fasta \
        OUTPUT=/proj/b2015069/canfam3.dict

#Run this only once if necessary (in case of incompatibilities between contigs reference vs. vcf file eg. "chr1" vs. chr1"
perl -pe 's/^([^#])/chr\1/' dbSNP_canis_familiaris.vcf > dbSNP_canis_familiaris_with_chr.vcf
```

3. Alignment of the reads to the reference genome with BWA

```
#----------
#!/bin/bash
#SBATCH -A b2015069
#SBATCH -p core
#SBATCH -n 6
#SBATCH -t 4-20:00:00

module load bioinfo-tools
module load bioinfo-tools samtools
module load bioinfo-tools java/sun_jdk1.7.0_25
module load bwa

#Align the pair end reads to the reference genome

#mkdir /proj/b2015069/labbe_retinopathy/map_and_mark

bwa mem -t 6 -M -R "@RG\tID:@NS500636\tSM:sire\tPL:ILLUMINA\tLB:lib3" \
        /proj/b2015069/canfam3.fasta \
        /proj/b2015069/labbe_retinopathy/merged/merged_reads_sire_fwd.fq \
        /proj/b2015069/labbe_retinopathy/merged/merged_reads_sire_rev.fq \
        > /proj/b2015069/labbe_retinopathy/map_and_mark/sire_aligned_reads.sam &

bwa mem -t 6 -M -R "@RG\tID:@NS500636\tSM:dam\tPL:ILLUMINA\tLB:lib3" \
        /proj/b2015069/canfam3.fasta \
        /proj/b2015069/labbe_retinopathy/merged/merged_reads_dam_fwd.fq \
        /proj/b2015069/labbe_retinopathy/merged/merged_reads_dam_rev.fq \
        > /proj/b2015069/labbe_retinopathy/map_and_mark/dam_aligned_reads.sam &

bwa mem -t 6 -M -R "@RG\tID:@NS500636\tSM:off1\tPL:ILLUMINA\tLB:lib3" \
        /proj/b2015069/canfam3.fasta \
        /proj/b2015069/labbe_retinopathy/merged/merged_reads_off1_fwd.fq \
        /proj/b2015069/labbe_retinopathy/merged/merged_reads_off1_rev.fq \
        > /proj/b2015069/labbe_retinopathy/map_and_mark/off1_aligned_reads.sam &

bwa mem -t 6 -M -R "@RG\tID:@NS500636\tSM:off2\tPL:ILLUMINA\tLB:lib3" \
        /proj/b2015069/canfam3.fasta \
        /proj/b2015069/labbe_retinopathy/merged/merged_reads_off2_fwd.fq \
        /proj/b2015069/labbe_retinopathy/merged/merged_reads_off2_rev.fq \
        > /proj/b2015069/labbe_retinopathy/map_and_mark/off2_aligned_reads.sam &

wait
echo "#####################"
echo "BWA mem done"

#Check statistics
  #Check number of reads aligning to the reference genome
    echo "Number of paired reads aligned to the reference genome for Sire:" > first_stats.txt
    samtools view -S -F0x4 /proj/b2015069/labbe_retinopathy/map_and_mark/sire_aligned_reads.sam | wc -l >>
first_stats.txt
    echo "Number of single reads aligned to the reference genome for Sire " >> first_stats.txt
    samtools view -S -f0x4 /proj/b2015069/labbe_retinopathy/map_and_mark/sire_aligned_reads.sam | wc -l >>
first_stats.txt

    echo "Number of paired reads aligned to the reference genome for Dam:" >> first_stats.txt
    samtools view -S -F0x4 /proj/b2015069/labbe_retinopathy/map_and_mark/dam_aligned_reads.sam | wc -l >>
first_stats.txt
    echo "Number of single reads aligned to the reference genome for Dam" >> first_stats.txt
    samtools view -S -f0x4 /proj/b2015069/labbe_retinopathy/map_and_mark/dam_aligned_reads.sam | wc -l >>
first_stats.txt

    echo "Number of paired reads aligned to the reference genome for Off1:" >> first_stats.txt
    samtools view -S -F0x4 /proj/b2015069/labbe_retinopathy/map_and_mark/off1_aligned_reads.sam | wc -l >>
first_stats.txt
    echo "Number of single reads aligned to the reference genome for Off1" >> first_stats.txt
    samtools view -S -f0x4 /proj/b2015069/labbe_retinopathy/map_and_mark/off1_aligned_reads.sam | wc -l >>
first_stats.txt

    echo "Number of paired reads aligned to the reference genome for off2:" >> first_stats.txt
    samtools view -S -F0x4 /proj/b2015069/labbe_retinopathy/map_and_mark/off2_aligned_reads.sam | wc -l >>
first_stats.txt
    echo "Number of single reads aligned to the reference genome for off2" >> first_stats.txt
    samtools view -S -f0x4 /proj/b2015069/labbe_retinopathy/map_and_mark/off2_aligned_reads.sam | wc -l >>
first_stats.txt
  echo "Statistics done"
        #Now we can check the statistics in the first_stats.txt file
```

4. GATK worflow

```
#Sort the SAM and convert it to BAM
java -jar -Xmx11g /sw/apps/bioinfo/picard/1.92/milou/SortSam.jar \
        INPUT=/proj/b2015069/labbe_retinopathy/map_and_mark/sire_aligned_reads.sam \
        OUTPUT=/proj/b2015069/labbe_retinopathy/map_and_mark/sorted_reads_sire.bam \
        SORT_ORDER=coordinate &

java -jar -Xmx11g /sw/apps/bioinfo/picard/1.92/milou/SortSam.jar \
        INPUT=/proj/b2015069/labbe_retinopathy/map_and_mark/dam_aligned_reads.sam \
        OUTPUT=/proj/b2015069/labbe_retinopathy/map_and_mark/sorted_reads_dam.bam \
        SORT_ORDER=coordinate &

java -jar -Xmx11g /sw/apps/bioinfo/picard/1.92/milou/SortSam.jar \
        INPUT=/proj/b2015069/labbe_retinopathy/map_and_mark/off1_aligned_reads.sam \
        OUTPUT=/proj/b2015069/labbe_retinopathy/map_and_mark/sorted_reads_off1.bam \
        SORT_ORDER=coordinate &

java -jar -Xmx11g /sw/apps/bioinfo/picard/1.92/milou/SortSam.jar \
        INPUT=/proj/b2015069/labbe_retinopathy/map_and_mark/off2_aligned_reads.sam \
        OUTPUT=/proj/b2015069/labbe_retinopathy/map_and_mark/sorted_reads_off2.bam \
        SORT_ORDER=coordinate &

wait
echo "#####################"
echo "SortSam done"


#Mark duplicates
java -jar -Xmx11g /sw/apps/bioinfo/picard/1.92/milou/MarkDuplicates.jar \
        INPUT=/proj/b2015069/labbe_retinopathy/map_and_mark/sorted_reads_sire.bam \
        OUTPUT=/proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_sire.bam \
        METRICS_FILE=metrics_sire.txt &

java -jar -Xmx11g /sw/apps/bioinfo/picard/1.92/milou/MarkDuplicates.jar \
        INPUT=/proj/b2015069/labbe_retinopathy/map_and_mark/sorted_reads_dam.bam \
        OUTPUT=/proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_dam.bam \
        METRICS_FILE=metrics_dam.txt &

java -jar -Xmx11g /sw/apps/bioinfo/picard/1.92/milou/MarkDuplicates.jar \
        INPUT=/proj/b2015069/labbe_retinopathy/map_and_mark/sorted_reads_off1.bam \
        OUTPUT=/proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_off1.bam \
        METRICS_FILE=metrics_off1.txt &

java -jar -Xmx11g /sw/apps/bioinfo/picard/1.92/milou/MarkDuplicates.jar \
        INPUT=/proj/b2015069/labbe_retinopathy/map_and_mark/sorted_reads_off2.bam \
        OUTPUT=/proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_off2.bam \
        METRICS_FILE=metrics_off2.txt &

wait
echo "#####################"
echo "MarkDuplicates done"

#Index BAM files
java -jar -Xmx8g /sw/apps/bioinfo/picard/1.92/milou/BuildBamIndex.jar \
        INPUT=/proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_sire.bam &

java -jar -Xmx8g /sw/apps/bioinfo/picard/1.92/milou/BuildBamIndex.jar \
        INPUT=/proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_dam.bam &

java -jar -Xmx8g /sw/apps/bioinfo/picard/1.92/milou/BuildBamIndex.jar \
        INPUT=/proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_off1.bam &
```

```
java -jar -Xmx8g /sw/apps/bioinfo/picard/1.92/milou/BuildBamIndex.jar \
        INPUT=/proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_off2.bam &

wait
echo "####################"
echo "BuilBamIndex done; now time for IGV"

#----------
#!/bin/bash
#SBATCH -A b2015069
#SBATCH -p core
#SBATCH -n 4
#SBATCH -t 2-10:00:00

#Check statistics of the reads used: quality, genome coverage, GC%, etc
module load bioinfo-tools
module load bioinfo-tools java/sun_jdk1.7.0_25
module load R/3.1.0
module load BEDTools/2.23.0

mkdir /proj/b2015069/labbe_retinopathy/map_and_mark/reads_stats_qual

        #1 Check number of reads pair-end aligned, single end aligned, duplicates, etc.

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -T FlagStat \
        -nct 4 \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_sire.bam \
        -o /proj/b2015069/labbe_retinopathy/map_and_mark/reads_stats_qual/stats_num_reads_sire.txt &

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -T FlagStat \
        -nct 4 \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_dam.bam \
        -o /proj/b2015069/labbe_retinopathy/map_and_mark/reads_stats_qual/stats_num_reads_dam.txt &

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -T FlagStat \
        -nct 4 \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_off1.bam \
        -o /proj/b2015069/labbe_retinopathy/map_and_mark/reads_stats_qual/stats_num_reads_off1.txt &

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -T FlagStat \
        -nct 4 \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_off2.bam \
        -o /proj/b2015069/labbe_retinopathy/map_and_mark/reads_stats_qual/stats_num_reads_off2.txt &

        #2 Check metrics quality of the reads used for mapping

java -Xmx7g -jar /sw/apps/bioinfo/picard/1.92/milou/CollectMultipleMetrics.jar \
        R=/proj/b2015069/canfam3.fasta \
        I=/proj/b2015069/labbe_retinopathy/map_and_mark/sorted_reads_sire.bam \
        O=/proj/b2015069/labbe_retinopathy/map_and_mark/reads_stats_qual/sire_metrics &

java -Xmx7g -jar /sw/apps/bioinfo/picard/1.92/milou/CollectMultipleMetrics.jar \
        R=/proj/b2015069/canfam3.fasta \
        I=/proj/b2015069/labbe_retinopathy/map_and_mark/sorted_reads_dam.bam \
        O=/proj/b2015069/labbe_retinopathy/map_and_mark/reads_stats_qual/dam_metrics &
```

```
java -Xmx7g -jar /sw/apps/bioinfo/picard/1.92/milou/CollectMultipleMetrics.jar \
        R=/proj/b2015069/canfam3.fasta \
        I=/proj/b2015069/labbe_retinopathy/map_and_mark/sorted_reads_off1.bam \
        O=/proj/b2015069/labbe_retinopathy/map_and_mark/reads_stats_qual/off1_metrics &

java -Xmx7g -jar /sw/apps/bioinfo/picard/1.92/milou/CollectMultipleMetrics.jar \
        R=/proj/b2015069/canfam3.fasta \
        I=/proj/b2015069/labbe_retinopathy/map_and_mark/sorted_reads_off2.bam \
        O=/proj/b2015069/labbe_retinopathy/map_and_mark/reads_stats_qual/off2_metrics &

        #3 Check coverage

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -T DepthOfCoverage \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_sire.bam \
        -o /proj/b2015069/labbe_retinopathy/map_and_mark/reads_stats_qual/coverage_sire.txt

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -T DepthOfCoverage \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_dam.bam \
        -o /proj/b2015069/labbe_retinopathy/map_and_mark/reads_stats_qual/coverage_dam.txt

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -T DepthOfCoverage \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_off1.bam \
        -o /proj/b2015069/labbe_retinopathy/map_and_mark/reads_stats_qual/coverage_off1.txt

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -T DepthOfCoverage \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_off2.bam \
        -o /proj/b2015069/labbe_retinopathy/map_and_mark/reads_stats_qual/coverage_off2.txt

#---------
#!/bin/bash
#SBATCH -A b2015069
#SBATCH -p core
#SBATCH -n 6
#SBATCH -t 2-20:00:00

#Local Realignment around indels
#Determining small suspicious intervals which need realignment with RealignerTargetCreator

module load bioinfo-tools
module load bioinfo-tools java/sun_jdk1.7.0_25

#mkdir /proj/b2015069/labbe_retinopathy/indel_realignment

java -Xmx10g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -nt 6 \
        -T RealignerTargetCreator \
        -R /proj/b2015069/canfam3.fasta  \
        -I /proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_sire.bam \
        -o /proj/b2015069/labbe_retinopathy/indel_realignment/target_intervals_list_sire.list &

java -Xmx10g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -nt 6 \
        -T RealignerTargetCreator \
        -R /proj/b2015069/canfam3.fasta  \
        -I /proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_dam.bam \
        -o /proj/b2015069/labbe_retinopathy/indel_realignment/target_intervals_list_dam.list &

java -Xmx10g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -nt 6 \
        -T RealignerTargetCreator \
        -R /proj/b2015069/canfam3.fasta  \
        -I /proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_off1.bam \
        -o /proj/b2015069/labbe_retinopathy/indel_realignment/target_intervals_list_off1.list &

java -Xmx10g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -nt 6 \
        -T RealignerTargetCreator \
        -R /proj/b2015069/canfam3.fasta  \
        -I /proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_off2.bam \
        -o /proj/b2015069/labbe_retinopathy/indel_realignment/target_intervals_list_off2.list &

echo "RealignerTargetCreator done"
wait


        #Running the realigner over those suspicious intervals with IndelRealigner
java -Xmx5g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -T IndelRealigner \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_sire.bam \
        -targetIntervals /proj/b2015069/labbe_retinopathy/indel_realignment/target_intervals_list_sire.list \
        -o /proj/b2015069/labbe_retinopathy/indel_realignment/realigned_reads_sire.bam

java -Xmx5g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -T IndelRealigner \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_dam.bam \
        -targetIntervals /proj/b2015069/labbe_retinopathy/indel_realignment/target_intervals_list_dam.list \
        -o /proj/b2015069/labbe_retinopathy/indel_realignment/realigned_reads_dam.bam

java -Xmx5g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -T IndelRealigner \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_off1.bam \
        -targetIntervals /proj/b2015069/labbe_retinopathy/indel_realignment/target_intervals_list_off1.list \
        -o /proj/b2015069/labbe_retinopathy/indel_realignment/realigned_reads_off1.bam

java -Xmx5g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -T IndelRealigner \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/map_and_mark/dedup_reads_off2.bam \
        -targetIntervals /proj/b2015069/labbe_retinopathy/indel_realignment/target_intervals_list_off2.list \
        -o /proj/b2015069/labbe_retinopathy/indel_realignment/realigned_reads_off2.bam

echo "IndelRealigner done"
wait


#Base Quality Score Recalibration
mkdir /proj/b2015069/labbe_retinopathy/base_recalibrator

        #Analyze patterns of covariation in the sequence dataset using BaseRecalibrator
java -Xmx10g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -nct 6 \
        -T BaseRecalibrator \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/indel_realignment/realigned_reads_sire.bam \
        -knownSites /proj/b2015069/dbSNP_canis_familiaris.vcf \
        -o /proj/b2015069/labbe_retinopathy/base_recalibrator/recalibrated_data_sire.table &

java -Xmx10g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -nct 6 \
        -T BaseRecalibrator \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/indel_realignment/realigned_reads_dam.bam \
        -knownSites /proj/b2015069/dbSNP_canis_familiaris.vcf \
        -o /proj/b2015069/labbe_retinopathy/base_recalibrator/recalibrated_data_dam.table &

java -Xmx10g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -nct 6 \
        -T BaseRecalibrator \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/indel_realignment/realigned_reads_off1.bam \
        -knownSites /proj/b2015069/dbSNP_canis_familiaris.vcf \
        -o /proj/b2015069/labbe_retinopathy/base_recalibrator/recalibrated_data_off1.table &

java -Xmx10g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -nct 6 \
        -T BaseRecalibrator \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/indel_realignment/realigned_reads_off2.bam \
        -knownSites /proj/b2015069/dbSNP_canis_familiaris.vcf \
        -o /proj/b2015069/labbe_retinopathy/base_recalibrator/recalibrated_data_off2.table &

wait
echo "#####################"
echo "BaseRecalibrator 1 done"

        #Analysis of the covariation remaining after recalibration using BaseRecalibrator again.
java -Xmx10g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -nct 6 \
        -T BaseRecalibrator \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/indel_realignment/realigned_reads_sire.bam \
        -knownSites /proj/b2015069/dbSNP_canis_familiaris.vcf \
        -BQSR /proj/b2015069/labbe_retinopathy/base_recalibrator/recalibrated_data_sire.table \
        -o /proj/b2015069/labbe_retinopathy/base_recalibrator/post_recalibrated_data_sire.table &

java -Xmx10g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -nct 6 \
        -T BaseRecalibrator \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/indel_realignment/realigned_reads_dam.bam \
        -knownSites /proj/b2015069/dbSNP_canis_familiaris.vcf \
        -BQSR /proj/b2015069/labbe_retinopathy/base_recalibrator/recalibrated_data_dam.table \
        -o /proj/b2015069/labbe_retinopathy/base_recalibrator/post_recalibrated_data_dam.table &

java -Xmx10g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -nct 6 \
        -T BaseRecalibrator \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/indel_realignment/realigned_reads_off1.bam \
        -knownSites /proj/b2015069/dbSNP_canis_familiaris.vcf \
        -BQSR /proj/b2015069/labbe_retinopathy/base_recalibrator/recalibrated_data_off1.table \
        -o /proj/b2015069/labbe_retinopathy/base_recalibrator/post_recalibrated_data_off1.table &

java -Xmx10g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -nct 6 \
        -T BaseRecalibrator \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/indel_realignment/realigned_reads_off2.bam \
        -knownSites /proj/b2015069/dbSNP_canis_familiaris.vcf \
        -BQSR /proj/b2015069/labbe_retinopathy/base_recalibrator/recalibrated_data_off2.table \
        -o /proj/b2015069/labbe_retinopathy/base_recalibrator/post_recalibrated_data_off2.table &

wait
echo "#####################"
echo "BaseRecalibrator 2 done"


        #Apply recalibration to the sequence data.
java -Xmx10g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -nct 6 \
        -T PrintReads \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/indel_realignment/realigned_reads_sire.bam \
        -BQSR /proj/b2015069/labbe_retinopathy/base_recalibrator/recalibrated_data_sire.table \
        -o /proj/b2015069/labbe_retinopathy/base_recalibrator/final_recalibrated_reads_sire.bam &

java -Xmx10g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -nct 6 \
        -T PrintReads \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/indel_realignment/realigned_reads_dam.bam \
        -BQSR /proj/b2015069/labbe_retinopathy/base_recalibrator/recalibrated_data_dam.table \
        -o /proj/b2015069/labbe_retinopathy/base_recalibrator/final_recalibrated_reads_dam.bam &

java -Xmx10g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -nct 6 \
        -T PrintReads \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/indel_realignment/realigned_reads_off1.bam \
        -BQSR /proj/b2015069/labbe_retinopathy/base_recalibrator/recalibrated_data_off1.table \
        -o /proj/b2015069/labbe_retinopathy/base_recalibrator/final_recalibrated_reads_off1.bam &

java -Xmx10g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
        -nct 6 \
        -T PrintReads \
        -R /proj/b2015069/canfam3.fasta \
        -I /proj/b2015069/labbe_retinopathy/indel_realignment/realigned_reads_off2.bam \
        -BQSR /proj/b2015069/labbe_retinopathy/base_recalibrator/recalibrated_data_off2.table \
        -o /proj/b2015069/labbe_retinopathy/base_recalibrator/final_recalibrated_reads_off2.bam &

wait
echo "#####################"
echo "PrintReads done"

#-------------
#!/bin/bash
#SBATCH -A b2015069
#SBATCH -p core
#SBATCH -n 6
#SBATCH -t 5-10:00:00

#Variant discovery

module load bioinfo-tools
module load bioinfo-tools java/sun_jdk1.7.0_25

#mkdir /proj/b2015069/labbe_retinopathy/variant_discovery
#mkdir /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants
```

```
#Calling variants with Haplotype Caller in gVCF file format

java -Xmx48g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
	-T HaplotypeCaller \
	-nct 6 \
	-R /proj/b2015069/canfam3.fasta \
	-I /proj/b2015069/labbe_retinopathy/base_recalibrator/final_recalibrated_reads_sire.bam \
	--emitRefConfidence GVCF \
	--variant_index_type LINEAR \
	--variant_index_parameter 128000 \
	--dbsnp /proj/b2015069/dbSNP_canis_familiaris.vcf \
	-stand_emit_conf 25 \
	-stand_call_conf 10 \
	-o /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/raw_snps_indels_1_sire.g.vcf

#-------------
#!/bin/bash
#SBATCH -A b2015069
#SBATCH -p core
#SBATCH -n 8
#SBATCH -t 2-10:00:00

module load bioinfo-tools
module load bioinfo-tools java/sun_jdk1.7.0_25

java -Xmx48g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
	-T HaplotypeCaller \
	-nct 8 \
	-R /proj/b2015069/canfam3.fasta \
	-I /proj/b2015069/labbe_retinopathy/base_recalibrator/final_recalibrated_reads_dam.bam \
	--emitRefConfidence GVCF \
	--variant_index_type LINEAR \
	--variant_index_parameter 128000 \
	--dbsnp /proj/b2015069/dbSNP_canis_familiaris.vcf \
	-stand_emit_conf 25 \
	-stand_call_conf 10 \
	-o /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/raw_snps_indels_1_dam.g.vcf


#-------------
#!/bin/bash
#SBATCH -A b2015069
#SBATCH -p core
#SBATCH -n 6
#SBATCH -t 2-10:00:00

module load bioinfo-tools
module load bioinfo-tools java/sun_jdk1.7.0_25

java -Xmx48g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
	-T HaplotypeCaller \
	-nct 6 \
	-R /proj/b2015069/canfam3.fasta \
	-I /proj/b2015069/labbe_retinopathy/base_recalibrator/final_recalibrated_reads_off1.bam \
	--emitRefConfidence GVCF \
	--variant_index_type LINEAR \
	--variant_index_parameter 128000 \
	--dbsnp /proj/b2015069/dbSNP_canis_familiaris.vcf \
	-stand_emit_conf 25 \
	-stand_call_conf 10 \
	-o /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/raw_snps_indels_1_off1.g.vcf
```

```
#-------------
#!/bin/bash
#SBATCH -A b2015069
#SBATCH -p core
#SBATCH -n 8
#SBATCH -t 3-10:00:00

module load bioinfo-tools
module load bioinfo-tools java/sun_jdk1.7.0_25

java -Xmx48g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
            -T HaplotypeCaller \
            -nct 6 \
            -R /proj/b2015069/canfam3.fasta \
            -I /proj/b2015069/labbe_retinopathy/base_recalibrator/final_recalibrated_reads_off2.bam \
            --emitRefConfidence GVCF \
            --variant_index_type LINEAR \
            --variant_index_parameter 128000 \
            --dbsnp /proj/b2015069/dbSNP_canis_familiaris.vcf \
            -stand_emit_conf 25 \
            -stand_call_conf 10 \
            -o /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/raw_snps_indels_1_off2.g.vcf


#-------------
#!/bin/bash
#SBATCH -A b2015069
#SBATCH -p core
#SBATCH -n 6
#SBATCH -t 2-10:00:00

module load bioinfo-tools
module load bioinfo-tools java/sun_jdk1.7.0_25

#Joint the gVCFs files with GenotypeGVCFs per family trios

java -Xmx16g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
            -T GenotypeGVCFs \
            -nt 6 \
            -R /proj/b2015069/canfam3.fasta \
            --variant /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/raw_snps_indels_1_sire.g.vcf \
            --variant /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/raw_snps_indels_1_dam.g.vcf \
            --variant /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/raw_snps_indels_1_off1.g.vcf \
            --variant /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/raw_snps_indels_1_off2.g.vcf \
   -o /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/all_joined_gvcf.vcf


#--------------------
#!/bin/bash
#SBATCH -A b2015069
#SBATCH -p core
#SBATCH -n 4
#SBATCH -t 10:00:00

#Extract the family trios given a single VCF file, creating 2 VCF files for each trio

module load bioinfo-tools
module load bioinfo-tools java/sun_jdk1.7.0_25


java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
            -R /proj/b2015069/canfam3.fasta \
            -nt 4 \
            -T SelectVariants \
            --variant /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/all_joined_gvcf.vcf \
            -o /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/trio_1_with_off1.vcf \
            -sn sire \
            -sn dam \
            -sn off1 &

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
            -R /proj/b2015069/canfam3.fasta \
            -nt 4 \
            -T SelectVariants \
            --variant /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/all_joined_gvcf.vcf \
            -o /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/trio_2_with_off2.vcf \
            -sn sire \
            -sn dam \
            -sn off2 &

wait
echo "#####################"
echo "Select variant done!"


#--------------
#!/bin/bash
#SBATCH -A b2015069
#SBATCH -p core
#SBATCH -n 4
#SBATCH -t 3-10:00:00

#Variant filtering for SNPs and indels separately using VariantFiltration
module load bioinfo-tools
module load bioinfo-tools java/sun_jdk1.7.0_25

#mkdir /proj/b2015069/labbe_retinopathy/variant_discovery/variant_filtering

            #1. Extract the SNPs from the call set

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
            -T SelectVariants \
            -nt 4 \
            -R /proj/b2015069/canfam3.fasta \
            -V /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/trio_1_with_off1.vcf \
            -selectType SNP \
            -o /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/raw_snps_trio_1.vcf &

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
            -T SelectVariants \
            -nt 4 \
            -R /proj/b2015069/canfam3.fasta \
            -V /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/trio_2_with_off2.vcf \
            -selectType SNP \
            -o /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/raw_snps_trio_2.vcf &

wait
echo "SNPs extracted successfully!"

            #2. Apply hard filters for SNPs

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
            -T VariantFiltration \
            -R /proj/b2015069/canfam3.fasta \
            -V /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/raw_snps_trio_1.vcf \
            --filterExpression "QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0" \
            --filterName "my_snp_filter" \
            -o /proj/b2015069/labbe_retinopathy/variant_discovery/variant_filtering/filtered_snps_trio_1.vcf &

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
            -T VariantFiltration \
            -R /proj/b2015069/canfam3.fasta \
            -V /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/raw_snps_trio_2.vcf \
            --filterExpression "QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0" \
            --filterName "my_snp_filter" \
            -o /proj/b2015069/labbe_retinopathy/variant_discovery/variant_filtering/filtered_snps_trio_2.vcf &

wait
echo "SNPs filters with exit"


            #3. Extract the indels from the call set

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
            -T SelectVariants \
            -nt 4 \
            -R /proj/b2015069/canfam3.fasta \
            -V /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/trio_1_with_off1.vcf \
            -selectType INDEL \
            -o /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/raw_indels_trio_1.vcf &

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
            -T SelectVariants \
            -nt 4 \
            -R /proj/b2015069/canfam3.fasta \
            -V /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/trio_2_with_off2.vcf \
            -selectType INDEL \
            -o /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/raw_indels_trio_2.vcf &

wait
echo "INDELs extracted successfully!"


                #4. Apply hard filters for indels

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
            -T VariantFiltration \
            -R /proj/b2015069/canfam3.fasta \
            -V /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/raw_indels_trio_1.vcf \
            --filterExpression "QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0" \
            --filterName "my_indel_filter" \
            -o /proj/b2015069/labbe_retinopathy/variant_discovery/variant_filtering/filtered_indels_trio_1.vcf &

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
            -T VariantFiltration \
            -R /proj/b2015069/canfam3.fasta \
            -V /proj/b2015069/labbe_retinopathy/variant_discovery/calling_variants/raw_indels_trio_2.vcf \
            --filterExpression "QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0" \
            --filterName "my_indel_filter" \
            -o /proj/b2015069/labbe_retinopathy/variant_discovery/variant_filtering/filtered_indels_trio_2.vcf &


wait
echo "Hard filters done!"


            #5. Select those variants which passed the filter

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
	-T SelectVariants \
	-nt 4 \
	-R /proj/b2015069/canfam3.fasta \
	-V /proj/b2015069/labbe_retinopathy/variant_discovery/variant_filtering/filtered_snps_trio_1.vcf \
	-select 'vc.isNotFiltered()' \
	-o /proj/b2015069/labbe_retinopathy/variant_discovery/variant_filtering/pass_filtered_snps_trio_1.vcf &

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
	-T SelectVariants \
	-nt 4 \
	-R /proj/b2015069/canfam3.fasta \
	-V /proj/b2015069/labbe_retinopathy/variant_discovery/variant_filtering/filtered_snps_trio_2.vcf \
	-select 'vc.isNotFiltered()' \
	-o /proj/b2015069/labbe_retinopathy/variant_discovery/variant_filtering/pass_filtered_snps_trio_2.vcf &

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
	-T SelectVariants \
	-nt 4 \
	-R /proj/b2015069/canfam3.fasta \
	-V /proj/b2015069/labbe_retinopathy/variant_discovery/variant_filtering/filtered_indels_trio_1.vcf \
	-select 'vc.isNotFiltered()' \
	-o /proj/b2015069/labbe_retinopathy/variant_discovery/variant_filtering/pass_filtered_indels_trio_1.vcf &

java -Xmx7g -jar /sw/apps/bioinfo/GATK/3.3.0/GenomeAnalysisTK.jar \
	-T SelectVariants \
	-nt 4 \
	-R /proj/b2015069/canfam3.fasta \
	-V /proj/b2015069/labbe_retinopathy/variant_discovery/variant_filtering/filtered_indels_trio_2.vcf \
	-select 'vc.isNotFiltered()' \
	-o /proj/b2015069/labbe_retinopathy/variant_discovery/variant_filtering/pass_filtered_indels_trio_2.vcf &

wait
echo "ALL done!"
```

5. Annotation of the detected variants with ANNOVAR

```
#From now on working from the command line, no BATCH
#We need the perl scripts and canFam3 database on the working directory

mkdir /proj/b2015069/labbe_retinopathy/annovar/input_annovar
mkdir /proj/b2015069/labbe_retinopathy/annovar/annovar_output

perl convert2annovar.pl -format vcf4old
/proj/b2015069/labbe_retinopathy/variant_discovery/variant_filtering/pass_filtered_snps_trio_1.vcf >
input_annovar/outfile.snps_trio_1_annovar_input.vcf -include &
perl convert2annovar.pl -format vcf4old
/proj/b2015069/labbe_retinopathy/variant_discovery/variant_filtering/pass_filtered_snps_trio_2.vcf >
input_annovar/outfile.snps_trio_2_annovar_input.vcf -include &
perl convert2annovar.pl -format vcf4old
/proj/b2015069/labbe_retinopathy/variant_discovery/variant_filtering/pass_filtered_indels_trio_1.vcf >
input_annovar/outfile.indels_trio_1_annovar_input.vcf -include &
perl convert2annovar.pl -format vcf4old
/proj/b2015069/labbe_retinopathy/variant_discovery/variant_filtering/pass_filtered_indels_trio_2.vcf >
input_annovar/outfile.indels_trio_2_annovar_input.vcf -include &

perl annotate_variation.pl -out annovar_output/annovar_snps_trio_1_input.vcf -build canFam3
input_annovar/outfile.snps_trio_1_annovar_input.vcf canFam3db/ &
perl annotate_variation.pl -out annovar_output/annovar_snps_trio_2_input.vcf -build canFam3
input_annovar/outfile.snps_trio_2_annovar_input.vcf canFam3db/ &
perl annotate_variation.pl -out annovar_output/annovar_indels_trio_1_input.vcf -build canFam3
input_annovar/outfile.indels_trio_1_annovar_input.vcf canFam3db/ &
perl annotate_variation.pl -out annovar_output/annovar_indels_trio_2_input.vcf -build canFam3
input_annovar/outfile.indels_trio_2_annovar_input.vcf canFam3db/ &
```

6. Perl scripts to analyze ANNOVAR output

6.1 Perl script for exonic variant function files

```
#!/usr/bin/perl
use strict;

#These script will be used for the files:
        #annovar_snps_trio_1_input.vcf.exonic_variant_function
        #annovar_snps_trio_2_input.vcf.exonic_variant_function
        #annovar_indels_trio_1_input.vcf.exonic_variant_function
        #annovar_indels_trio_2_input.vcf.exonic_variant_function

my $filename = 'annovar_snps_trio_1_input.vcf.exonic_variant_function';
open my $fh, $filename or die "Could not open file '$filename': $!";

#Create the directory "output_exonic_variants" to save the outputs classified.
my $existingdir = './snps_trio_1_exonic_variant_function';
mkdir $existingdir unless -d $existingdir; # Check if dir exists. If not create it.

#Create an array with the different types of annotation outputs.
my @arg = ('frameshift insertion','frameshift deletion','frameshift block substitution','stopgain','stoploss','nonframeshift
insertion','nonframeshift deletion','nonframeshift block substitution','nonsynonymous SNV','synonymous
SNV','unknown','other');

#Set the number of total variants in the input file at 0.
my $num_variants = 0;

#Open the input file and read each line.
while (my $line = <$fh>) {
    #Split the information given by ANNOVAR by columns and give a name.
    my @infoRow = split (/\t/, $line);

my($line_pos,$annotation,$gene,$chr,$pos,$pos2,$ref,$alt,$chr2,$pos3,$ID,$ref2,$alt2,$qual,$filter,$info,$format,$sample_
offspring,$dam,$sire)=@infoRow[0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20];

    #Count number of variants in the file.
    $num_variants++;

    #For each type of argument, create a specific file in the './output_exonic_variants' directory and append the row in the
file.
    for my $argument(@arg) {
        if ($annotation eq $argument) {
            open my $fileHandle, ">>", "$existingdir/$argument" or die "Can't open '$existingdir/$argument\n";
            print $fileHandle "$line";
            close $fileHandle;
        }
    }
}
#Print total number of variants.
print "########\n";
print "The TOTAL number of exonic variants is $num_variants\n\n";

#Create new directory to save the varients with the wanted genotype.
my $newdir = './snps_trio_1_exonic_variant_function/wanted_genotype';
mkdir $newdir unless -d $newdir; # Check if dir exists. If not create it.

#Now time to check the genotypes.
#Create an array with all the variant files created.
my @FILES = glob('./snps_trio_1_exonic_variant_function/*');

#Enter to each file created, count the number of variants annotated and check the genotypes of the animals.
```

```
foreach my $file(@FILES) {

    #Set different combination of genotypes at 0.
    my $wanted_genotype = 0;
    my $variant_with_not_wanted_genotype = 0;
    my $one_or_more_genotypes_missing = 0;
    my $num_lines = 0;

    if (-e "$file") {
    #Print the type of file in a fancy way.
    my @name = split /\//, $file;
    my $name_file = $name[2];
    print "For the $name_file\n";
    print "----------------\n";
        open (my $fileHandle, "<", "$file") or die "Can't open $file: $!";
            while (my $line = <$fileHandle>) {
                $num_lines ++;
                my @infoRow = split /\t/, $line;

my($line_pos,$annotation,$gene,$chr,$pos,$pos2,$ref,$alt,$chr2,$pos3,$ID,$ref2,$alt2,$qual,$filter,$info,$format,$sample_
offspring,$dam,$sire)=@infoRow[0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20];

                #Extract the alleles of each of the samples
                my @info_split_offspring = split /:/, $sample_offspring;
                my $allele_offspring = $info_split_offspring[0];

                my @info_split_dam= split /:/, $dam;
                my $allele_dam = $info_split_dam[0];

                my @info_split_sire = split /:/, $sire;
                my $allele_sire = $info_split_sire[0];

                #Offspring homozygous recessive and parents heterozygous.
                if ($allele_offspring eq "1/1")  {
                    if ($allele_dam eq "1/0") {
                        if ($allele_sire eq "1/0") {
                            $wanted_genotype++;
                            open my $fh, ">>", "$newdir/wanted_genotype" or die "Can't open
'newdir/wanted_genotype'\n";
                            print $fh "$line";
                            close $fh;
                        } elsif ($allele_sire eq "0/1") {
                            $wanted_genotype++;
                            open my $fh, ">>", "$newdir/wanted_genotype" or die "Can't open
'newdir/wanted_genotype'\n";
                            print $fh "$line";
                            close $fh;
                        }
                    } elsif ($allele_dam eq "0/1") {
                        if ($allele_sire eq "1/0") {
                            $wanted_genotype++;
                            open my $fh, ">>", "$newdir/wanted_genotype" or die "Can't open
'newdir/wanted_genotype'\n";
                            print $fh "$line";
                            close $fh;
                        } elsif ($allele_sire eq "0/1") {
                            $wanted_genotype++;
                            open my $fh, ">>", "$newdir/wanted_genotype" or die "Can't open
'newdir/wanted_genotype'\n";
                            print $fh "$line";
                            close $fh;
                        }
                    }
```

```
                  #Start checking if any of the animals has a missing genotype (e.g. sire)
                } elsif ($allele_offspring =~ /(\d)(\/)(\d)/) {
                    if ($allele_dam =~ /(\d)(\/)(\d)/) {
                        if ($allele_sire =~ /\./) {
                          $one_or_more_genotypes_missing++;
                    } elsif ($allele_dam eq "0/1") {
                        if ($allele_sire eq "1/0") {
                            $wanted_genotype++;
                            open my $fh, ">>", "$newdir/wanted_genotype" or die "Can't open
'newdir/wanted_genotype'\n";
                            print $fh "$line";
                            close $fh;
                        } elsif ($allele_sire eq "0/1") {
                            $wanted_genotype++;
                            open my $fh, ">>", "$newdir/wanted_genotype" or die "Can't open
'newdir/wanted_genotype'\n";
                            print $fh "$line";
                            close $fh;
                        }
                    }

                #Check if any of the animals has a missing genotype: sire, dam and offspring.
                } elsif ($allele_offspring =~ /(\d)(\/)(\d)/) {
                    if ($allele_dam =~ /(\d)(\/)(\d)/) {
                        if ($allele_sire =~ /\./) {
                            $one_or_more_genotypes_missing++;
                        }
                    } elsif ($allele_dam =~ /\./) {
                            $one_or_more_genotypes_missing++;
                    }
                } elsif ($allele_offspring =~ /\./) {
                    $one_or_more_genotypes_missing++;
                }
            }
    }

#Check the number of variants which don't have the genotype that we are interested in and do not have any of the animals
with missing data.
my $variant_with_not_wanted_genotype = ($num_lines-$one_or_more_genotypes_missing-$wanted_genotype);

print "The total number of variants is: $num_lines\n";
print "The total number of variants with the wanted genotype profile is: $wanted_genotype\n";
print "The total number of variants without the wanted genotype profile is: $variant_with_not_wanted_genotype\n";
print "The total number of variants with one of more genotypes missing is: $one_or_more_genotypes_missing\n";
print "############\n";
}
}
```

6.2 Perl script for variant function files

```
#!/usr/bin/perl
use strict;

#These script will be used for the files:
        #annovar_snps_trio_1_input.vcf.variant_function
        #annovar_snps_trio_2_input.vcf.variant_function
        #annovar_indels_trio_1_input.vcf.variant_function
        #annovar_indels_trio_2_input.vcf.variant_function

my $filename = 'annovar_snps_trio_2_input.vcf.variant_function';
open my $fh, $filename or die "Could not open file '$filename': $!";

#Create the directory "output_exonic_variants" to save the outputs classified.
my $existingdir = './snps_trio_2_variant_function';
mkdir $existingdir unless -d $existingdir; # Check if dir exists. If not create it.

#Create an array with the different types of annotation outputs.
my @arg = ('exonic','splicing','ncRNA','UTR5','UTR3','intronic','upstream','downstream','intergenic');


#Set the number of total variants in the input file at 0.
my $num_variants = 0;

#Open the input file and read each line.
while (my $line = <$fh>) {
        #Split the information given by ANNOVAR by columns and give a name.
        my @infoRow = split (/\t/, $line);
        my($annotation,$gene,$chr,$pos,$pos2,$ref,$alt,$chr2,$pos3,$ID,$ref2,$alt2,$qual,$filter,$info,$format,$sample_
offspring,$dam,$sire)=@infoRow[0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19];

        #Count number of variants in the file.
        $num_variants++;

        #For each type of argument, create a specific file in the './output_exonic_variants' directory and append the row in
the file.
        for my $argument(@arg) {
                if ($annotation eq $argument) {
                        open my $fileHandle, ">>", "$existingdir/$argument" or die "Can't open
'$existingdir/$argument\n";
                        print $fileHandle "$line";
                        close $fileHandle;
                }
        }
}


#Print total number of variants.
print "########\n";
print "The TOTAL number of variants is $num_variants\n\n";

#Create new directory to save the varients with the wanted genotype.
my $newdir = './snps_trio_2_variant_function/wanted_genotype';
mkdir $newdir unless -d $newdir; # Check if dir exists. If not create it.

#Now time to check the genotypes.
#Create an array with all the variant files created.
my @FILES = glob('./snps_trio_2_variant_function/*');

#Enter to each file created, count the number of variants annotated and check the genotypes of the animals.
foreach my $file(@FILES) {
        #Set different combination of genotypes at 0.
        my $wanted_genotype = 0;
        my $variant_with_not_wanted_genotype = 0;
        my $one_or_more_genotypes_missing = 0;
        my $num_lines = 0;

        if (-e "$file") {
        #my @name = split /\//, $file;
  #my $name_file = $name[2];
  print "For the $file\n";
        print "----------------\n";
                open (my $fileHandle, "<", "$file") or die "Can't open $file: $!";
                        while (my $line = <$fileHandle>) {
                                $num_lines ++;
                                my @infoRow = split /\t/, $line;

        my($annotation,$gene,$chr,$pos,$pos2,$ref,$alt,$chr2,$pos3,$ID,$ref2,$alt2,$qual,$filter,$info,$format,$sample_
offspring,$dam,$sire)=@infoRow[0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19];

                                #Extract the alleles of each of the samples
                                my @info_split_offspring = split /:/, $sample_offspring;
                                my $allele_offspring = $info_split_offspring[0];

                                my @info_split_dam= split /:/, $dam;
                                my $allele_dam = $info_split_dam[0];

                                my @info_split_sire = split /:/, $sire;
                                my $allele_sire = $info_split_sire[0];

                                #Offspring homozygous recessive and parents heterozygous.
                if ($allele_offspring eq "1/1")  {
                  if ($allele_dam eq "1/0") {
                      if ($allele_sire eq "1/0") {
                          $wanted_genotype++;
                          open my $fh, ">>", "$newdir/wanted_genotype" or die "Can't open
'newdir/wanted_genotype'\n";
                          print $fh "$line";
                          close $fh;
                      } elsif ($allele_sire eq "0/1") {
                          $wanted_genotype++;
                          open my $fh, ">>", "$newdir/wanted_genotype" or die "Can't open
'newdir/wanted_genotype'\n";
                          print $fh "$line";
                          close $fh;
                      }
                  } elsif ($allele_dam eq "0/1") {
                      if ($allele_sire eq "1/0") {
                          $wanted_genotype++;
                          open my $fh, ">>", "$newdir/wanted_genotype" or die "Can't open
'newdir/wanted_genotype'\n";
                          print $fh "$line";
                          close $fh;
                      } elsif ($allele_sire eq "0/1") {
                          $wanted_genotype++;
                          open my $fh, ">>", "$newdir/wanted_genotype" or die "Can't open
'newdir/wanted_genotype'\n";
                          print $fh "$line";
                          close $fh;
                      }
                  }

              #Check if any of the animals has a missing genotype: sire, dam and/or offspring.
```

```
                } elsif ($allele_offspring =~ /(\d)(\/)(\d)/) {
                    if ($allele_dam =~ /(\d)(\/)(\d)/) {
                        if ($allele_sire =~ /\./) {
                            $one_or_more_genotypes_missing++;
                        }
                    } elsif ($allele_dam =~ /\./) {
                            $one_or_more_genotypes_missing++;
                    }
                } elsif ($allele_offspring =~ /\./) {
                    $one_or_more_genotypes_missing++;
                }

                           }
        }

#Check the number of variants which don't have the genotype that we are interested in and do not have any of the animals with missing data.
my $variant_with_not_wanted_genotype = ($num_lines-$one_or_more_genotypes_missing-$wanted_genotype);

print "The total number of variants is: $num_lines\n";
print "The total number of variants with the wanted genotype profile is: $wanted_genotype\n";
print "The total number of variants without the wanted genotype profile is: $variant_with_not_wanted_genotype\n";
print "The total number of variants with one of more genotypes missing is: $one_or_more_genotypes_missing\n";
print "#############\n";
}
```

6.3 Perl script for checking shared exonic variants in both trios and possible matching with previously reported retina inhered disease genes.

```
#!/usr/bin/perl
use strict;

#Create a directory to save the common variants presenting the wanted genotype.
my $newdir = './common_variants_indels_exonic';
mkdir $newdir unless -d $newdir; # Check if dir exists. If not create it.

#Give the path to the file with wanted genotype for all the variants. Both trios.
my $file_1 = './indels_trio_1_exonic_variant_function/wanted_genotype/wanted_genotype';
my $file_2 = './indels_trio_2_exonic_variant_function/wanted_genotype/wanted_genotype';
my $num_common = 0;

#Open both files and define genome position of the variants.
open (my $fh, "<", "$file_1") or die "Can't open $file_1: $!";
        while (my $line = <$fh>) {
                my @infoRow = split /\t/, $line;
                #Define position in the genome of the variant in trio 1.
                my $position_trio1 = $infoRow[4];
open (my $fh2, "<", "$file_2") or die "Could not open file '$file_2': $!";
        while (my $line2 = <$fh2>) {
                my @infoRow2 = split (/\t/, $line2);
                #Define position in the genome of the variant in trio 2.
                my $position_trio2 = $infoRow2[4];

                #If variant at the same position for both trios is found, print it in a new file.
                if ($position_trio2 eq $position_trio1) {
                        $num_common ++;
                        open my $fileHandle, ">>", "$newdir/indels_exonic_variant" or die "Can't open
'$newdir/indels_exonic_variant'\n";
                                print $fileHandle "$line";
                        close $fileHandle;
                }
        }
        }
print "$num_common\n";

#Check if there is any possible candidate gene found
my @RETINAL_GENES =
('PDE6B','RD3','PDE6A','STK38L','PRCD','SLC4A3','RPGR','RHO','CCDC66','ADAM9','RPGRIP1','NPHP4','CNGB3','RPE65','COL9A3'
,'COL9A2','NHEJ1','BEST1');

#Open the file only if exists.
if (-e "$newdir/indels_exonic_variant") {
        open (my $fileHandle, "<", "$newdir/indels_exonic_variant") or die "Can't open '$newdir/indels_exonic_variant':
$!";
                while (my $line = <$fileHandle>) {
                my @infoRow = split /\t/, $line;
                #Define the gene which is in common in both trios.
                my $common_gene = $infoRow[2];

                for my $candidate_gene(@RETINAL_GENES) {
                        if ($common_gene =~ /$candidate_gene/) {
                        print "Found $candidate_gene\n";
                        }
                }
                }
}
close $file_1;
close $file_2;
```

6.4 Perl script for checking shared variants in both trios and possible matching with previously reported retina inhered disease genes.

```
#!/usr/bin/perl
use strict;

#Create a directory to save the common variants presenting the wanted genotype.
my $newdir = './common_variants_snps';
mkdir $newdir unless -d $newdir; # Check if dir exists. If not create it.

#Give the path to the file with wanted genotype for all the variants. Both trios.
my $file_1 = './snps_trio_1_variant_function/wanted_genotype/wanted_genotype';
my $file_2 = './snps_trio_2_variant_function/wanted_genotype/wanted_genotype';
my $num_common = 0;

#Open both files and define genome position of the variants.
open (my $fh, "<", "$file_1") or die "Can't open $file_1: $!";
    while (my $line = <$fh>) {
        my @infoRow = split /\t/, $line;
        #Define position in the genome of the variant in trio 1.
        my $position_trio1 = $infoRow[3];
open (my $fh2, "<", "$file_2") or die "Could not open file '$file_2': $!";
    while (my $line2 = <$fh2>) {
        my @infoRow2 = split (/\t/, $line2);
        #Define position in the genome of the variant in trio 2.
        my $position_trio2 = $infoRow2[3];

        #If variant at the same position for both trios is found, print it in a new file.
        if ($position_trio2 eq $position_trio1) {
            $num_common ++;
            open my $fileHandle, ">>", "$newdir/snps_variant" or die "Can't open
'$newdir/snps_variant'\n";
                print $fileHandle "$line";
            close $fileHandle;
        }
    }
    }
print "$num_common\n";

#Check if there is any possible candidate gene found
my @RETINAL_GENES =
('PDE6B','RD3','PDE6A','STK38L','PRCD','SLC4A3','RPGR','RHO','CCDC66','ADAM9','RPGRIP1','NPHP4','CNGB3','RPE65','COL9A3'
,'COL9A2','NHEJ1','BEST1');

#Open the file only if exists.
if (-e "$newdir/snps_variant") {
    open (my $fileHandle, "<", "$newdir/snps_variant") or die "Can't open '$newdir/snps_variant': $!";
        while (my $line = <$fileHandle>) {
        my @infoRow = split /\t/, $line;
        #Define the gene which is in common in both trios.
        my $common_gene = $infoRow[2];

        for my $candidate_gene(@RETINAL_GENES) {
            if ($common_gene =~ /$candidate_gene/) {
            print "Found $candidate_gene\n";
            }
        }
        }
}

close $file_1;
close $file_2;
```

Appendix C. Results

Table1.

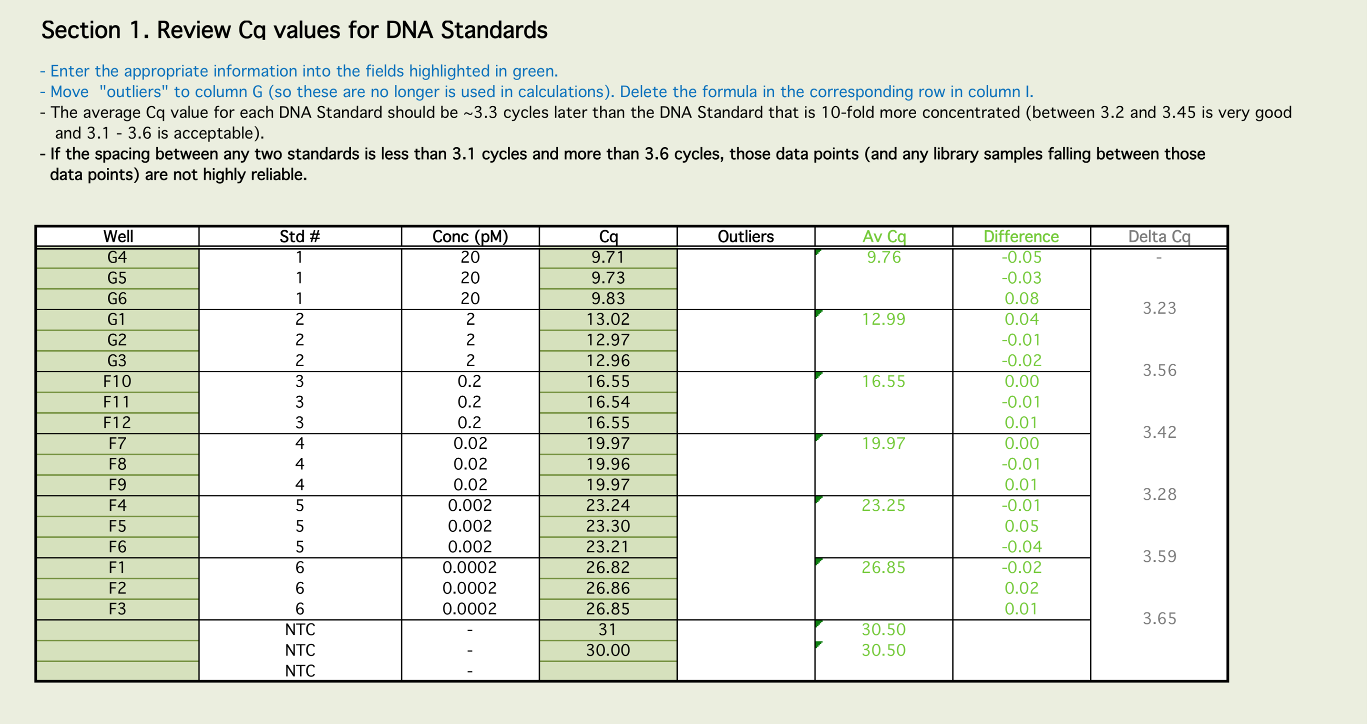
Summary of the reads for all the samples in both runs.

	Sample	Raw reads	Trimmed reads	Trimmed reads (x2)	Aligned reads (PE)	Genome Coverage
Run 1	**Sire I:1**	87,744,533	85,110,263	170,220,526	167,459,818	6.89 x
	Dam I:2	93,481,858	90,434,190	180,868,380	178,187,113	7.31 x
	Offspring II:1	90,477,985	86,521,473	173,042,946	170,123,976	6.97 x
	Offspring II:2	86,606,183	79,558,231	159,116,462	155,742,221	6.36 x
	Total	358,310,559	341,624,157	683,248,314	671,513,128	6.88 x[1]
Run 2	**Sire I:1**	97,403,683	95,315,614	190,631,228	190,139,797	11.45 x
	Dam I:2	105,820,150	103,748,174	207,496,348	207,180,231	12.39 x
	Offspring II:1	97,332,856	94,765,262	189,530,524	188,970,159	11.28 x
	Offspring II:2	88,190,346	84,849,995	169,699,990	168,707,631	10.04 x
	Total	388,747,035	378,679,045	757,358,090	754,997,818	11.29 x[1]

[1] Average coverage from all the samples.

1. Libraries quantification

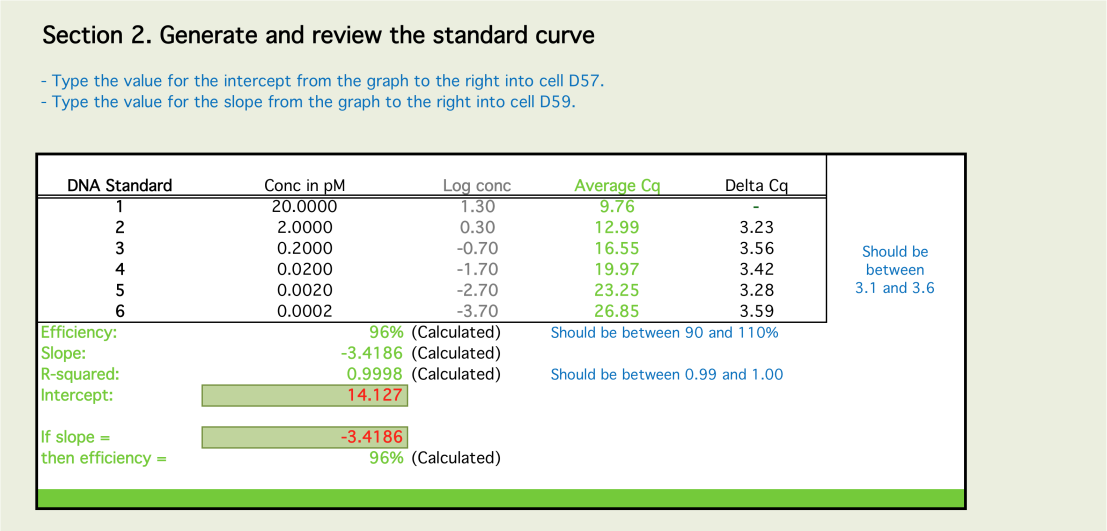
1.1 Run 1



Section 1. Review Cq values for DNA Standards

- Enter the appropriate information into the fields highlighted in green.
- Move "outliers" to column G (so these are no longer is used in calculations). Delete the formula in the corresponding row in column I.
- The average Cq value for each DNA Standard should be ~3.3 cycles later than the DNA Standard that is 10-fold more concentrated (between 3.2 and 3.45 is very good and 3.1 - 3.6 is acceptable).
- If the spacing between any two standards is less than 3.1 cycles and more than 3.6 cycles, those data points (and any library samples falling between those data points) are not highly reliable.

Well	Std #	Conc (pM)	Cq	Outliers	Av Cq	Difference	Delta Cq
G4	1	20	9.71		9.76	-0.05	-
G5	1	20	9.73			-0.03	
G6	1	20	9.83			0.08	3.23
G1	2	2	13.02		12.99	0.04	
G2	2	2	12.97			-0.01	
G3	2	2	12.96			-0.02	3.56
F10	3	0.2	16.55		16.55	0.00	
F11	3	0.2	16.54			-0.01	
F12	3	0.2	16.55			0.01	3.42
F7	4	0.02	19.97		19.97	0.00	
F8	4	0.02	19.96			-0.01	
F9	4	0.02	19.97			0.01	3.28
F4	5	0.002	23.24		23.25	-0.01	
F5	5	0.002	23.30			0.05	
F6	5	0.002	23.21			-0.04	3.59
F1	6	0.0002	26.82		26.85	-0.02	
F2	6	0.0002	26.86			0.02	
F3	6	0.0002	26.85			0.01	3.65
	NTC	-	31		30.50		
	NTC	-	30.00		30.50		
	NTC	-					



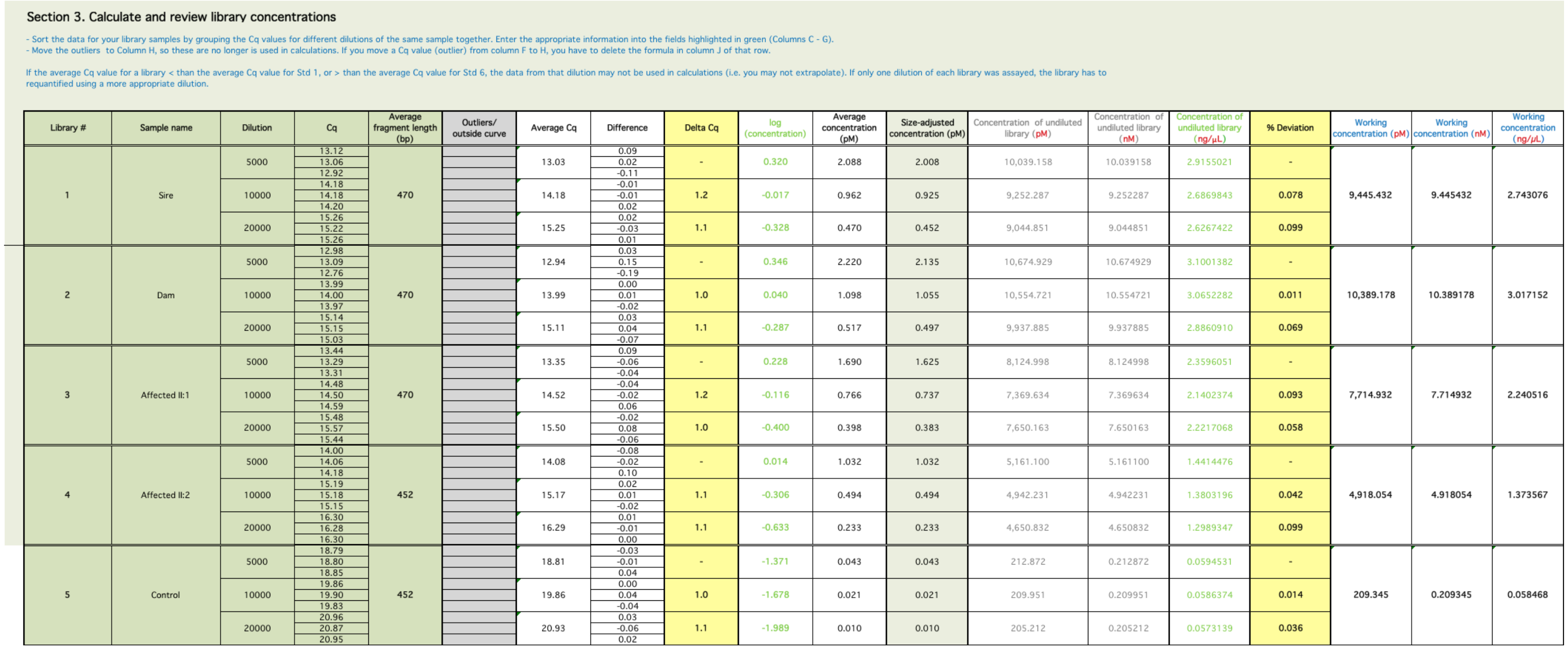
Section 2. Generate and review the standard curve

- Type the value for the intercept from the graph to the right into cell D57.
- Type the value for the slope from the graph to the right into cell D59.

DNA Standard	Conc in pM	Log conc	Average Cq	Delta Cq
1	20.0000	1.30	9.76	-
2	2.0000	0.30	12.99	3.23
3	0.2000	-0.70	16.55	3.56
4	0.0200	-1.70	19.97	3.42
5	0.0020	-2.70	23.25	3.28
6	0.0002	-3.70	26.85	3.59

Should be between 3.1 and 3.6

Efficiency:	96%	(Calculated)	Should be between 90 and 110%
Slope:	-3.4186	(Calculated)	
R-squared:	0.9998	(Calculated)	Should be between 0.99 and 1.00
Intercept:	14.127		
If slope =	-3.4186		
then efficiency =	96%	(Calculated)	



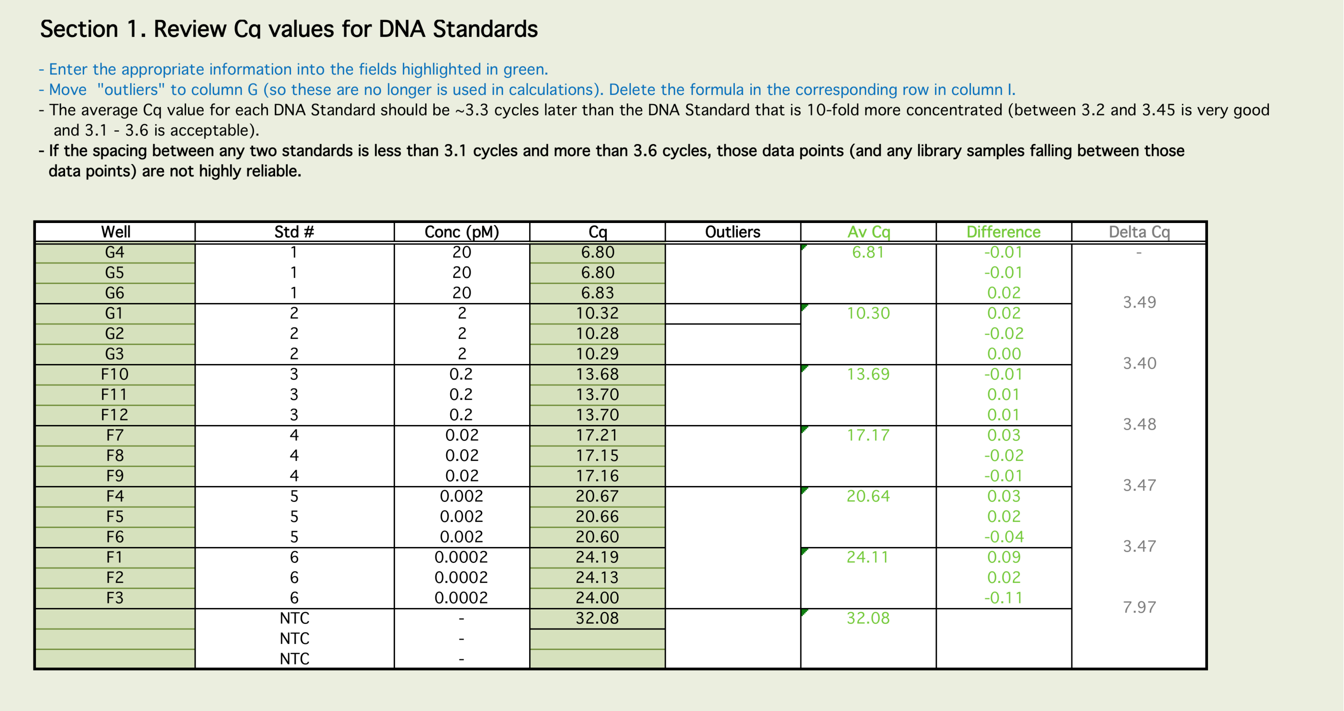
Section 3. Calculate and review library concentrations

- Sort the data for your library samples by grouping the Cq values for different dilutions of the same sample together. Enter the appropriate information into the fields highlighted in green (Columns C - G).
- Move the outliers to Column H, so these are no longer is used in calculations. If you move a Cq value (outlier) from column F to H, you have to delete the formula in column J of that row.

If the average Cq value for a library < than the average Cq value for Std 1, or > than the average Cq value for Std 6, the data from that dilution may not be used in calculations (i.e. you may not extrapolate). If only one dilution of each library was assayed, the library has to requantified using a more appropriate dilution.

Library #	Sample name	Dilution	Cq	Average fragment length (bp)	Outliers/ outside curve	Average Cq	Difference	Delta Cq	log (concentration)	Average concentration (pM)	Size-adjusted concentration (pM)	Concentration of undiluted library (pM)	Concentration of undiluted library (nM)	Concentration of undiluted library (ng/µL)	% Deviation	Working concentration (pM)	Working concentration (nM)	Working concentration (ng/µL)
1	Sire	5000	13.12	470		13.03	0.09	-	0.320	2.088	2.008	10,039.158	10.039158	2.9155021	-	9,445.432	9.445432	2.743076
			13.06				0.02											
			12.92				-0.11											
		10000	14.18			14.18	-0.01	1.2	-0.017	0.962	0.925	9,252.287	9.252287	2.6869843	0.078			
			14.18				-0.01											
			14.20				0.02											
		20000	15.26			15.25	0.02	1.1	-0.328	0.470	0.452	9,044.851	9.044851	2.6267422	0.099			
			15.22				-0.03											
			15.26				0.01											
2	Dam	5000	12.98	470		12.94	0.03	-	0.346	2.220	2.135	10,674.929	10.674929	3.1001382	-	10,389.178	10.389178	3.017152
			13.09				0.15											
			12.76				-0.19											
		10000	13.99			13.99	0.00	1.0	0.040	1.098	1.055	10,554.721	10.554721	3.0652282	0.011			
			14.00				0.01											
			13.97				-0.02											
		20000	15.14			15.11	0.03	1.1	-0.287	0.517	0.497	9,937.885	9.937885	2.8860910	0.069			
			15.15				0.04											
			15.03				-0.07											
3	Affected II:1	5000	13.44	470		13.35	0.09	-	0.228	1.690	1.625	8,124.998	8.124998	2.3596051	-	7,714.932	7.714932	2.240516
			13.29				-0.06											
			13.31				-0.04											
		10000	14.48			14.52	-0.04	1.2	-0.116	0.766	0.737	7,369.634	7.369634	2.1402374	0.093			
			14.50				-0.02											
			14.59				0.06											
		20000	15.48			15.50	-0.02	1.0	-0.400	0.398	0.383	7,650.163	7.650163	2.2217068	0.058			
			15.57				0.08											
			15.44				-0.06											
4	Affected II:2	5000	14.00	452		14.08	-0.08	-	0.014	1.032	1.032	5,161.100	5.161100	1.4414476	-	4,918.054	4.918054	1.373567
			14.06				-0.02											
			14.18				0.10											
		10000	15.19			15.17	0.02	1.1	-0.306	0.494	0.494	4,942.231	4.942231	1.3803196	0.042			
			15.18				0.01											
			15.15				-0.02											
		20000	16.30			16.29	0.01	1.1	-0.633	0.233	0.233	4,650.832	4.650832	1.2989347	0.099			
			16.28				-0.01											
			16.30				0.00											
5	Control	5000	18.79	452		18.81	-0.03	-	-1.371	0.043	0.043	212.872	0.212872	0.0594531	-	209.345	0.209345	0.058468
			18.80				-0.01											
			18.85				0.04											
		10000	19.86			19.86	0.00	1.0	-1.678	0.021	0.021	209.951	0.209951	0.0586374	0.014			
			19.90				0.04											
			19.83				-0.04											
		20000	20.96			20.93	0.03	1.1	-1.989	0.010	0.010	205.212	0.205212	0.0573139	0.036			
			20.87				-0.06											
			20.95				0.02											

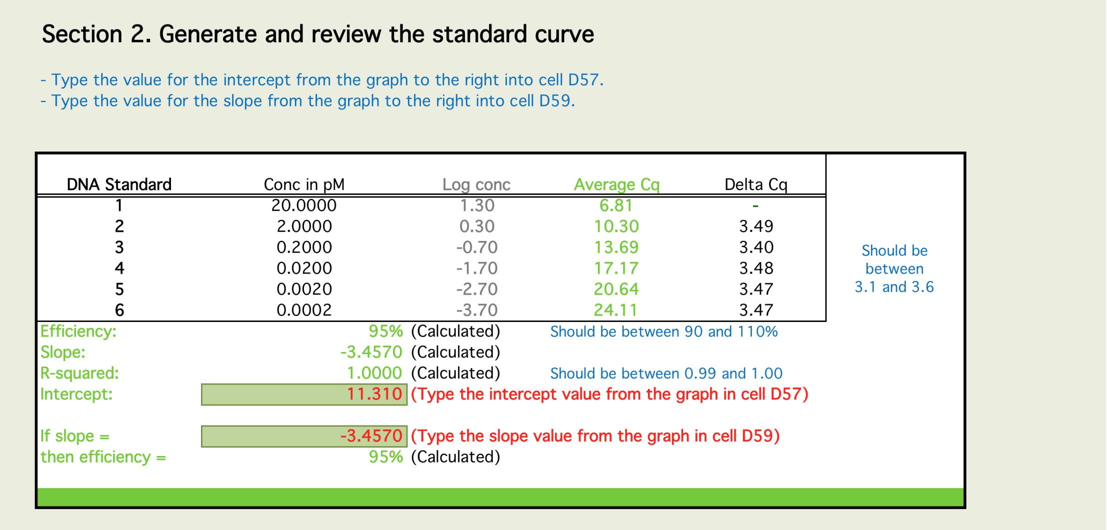
1.1 Run 2



Section 1. Review Cq values for DNA Standards

- Enter the appropriate information into the fields highlighted in green.
- Move "outliers" to column G (so these are no longer is used in calculations). Delete the formula in the corresponding row in column I.
- The average Cq value for each DNA Standard should be ~3.3 cycles later than the DNA Standard that is 10-fold more concentrated (between 3.2 and 3.45 is very good and 3.1 - 3.6 is acceptable).
- If the spacing between any two standards is less than 3.1 cycles and more than 3.6 cycles, those data points (and any library samples falling between those data points) are not highly reliable.

Well	Std #	Conc (pM)	Cq	Outliers	Av Cq	Difference	Delta Cq
G4	1	20	6.80		6.81	-0.01	-
G5	1	20	6.80			-0.01	
G6	1	20	6.83			0.02	3.49
G1	2	2	10.32		10.30	0.02	
G2	2	2	10.28			-0.02	
G3	2	2	10.29			0.00	3.40
F10	3	0.2	13.68		13.69	-0.01	
F11	3	0.2	13.70			0.01	
F12	3	0.2	13.70			0.01	3.48
F7	4	0.02	17.21		17.17	0.03	
F8	4	0.02	17.15			-0.02	
F9	4	0.02	17.16			-0.01	3.47
F4	5	0.002	20.67		20.64	0.03	
F5	5	0.002	20.66			0.02	
F6	5	0.002	20.60			-0.04	3.47
F1	6	0.0002	24.19		24.11	0.09	
F2	6	0.0002	24.13			0.02	
F3	6	0.0002	24.00			-0.11	7.97
	NTC	-	32.08		32.08		
	NTC	-					
	NTC	-					



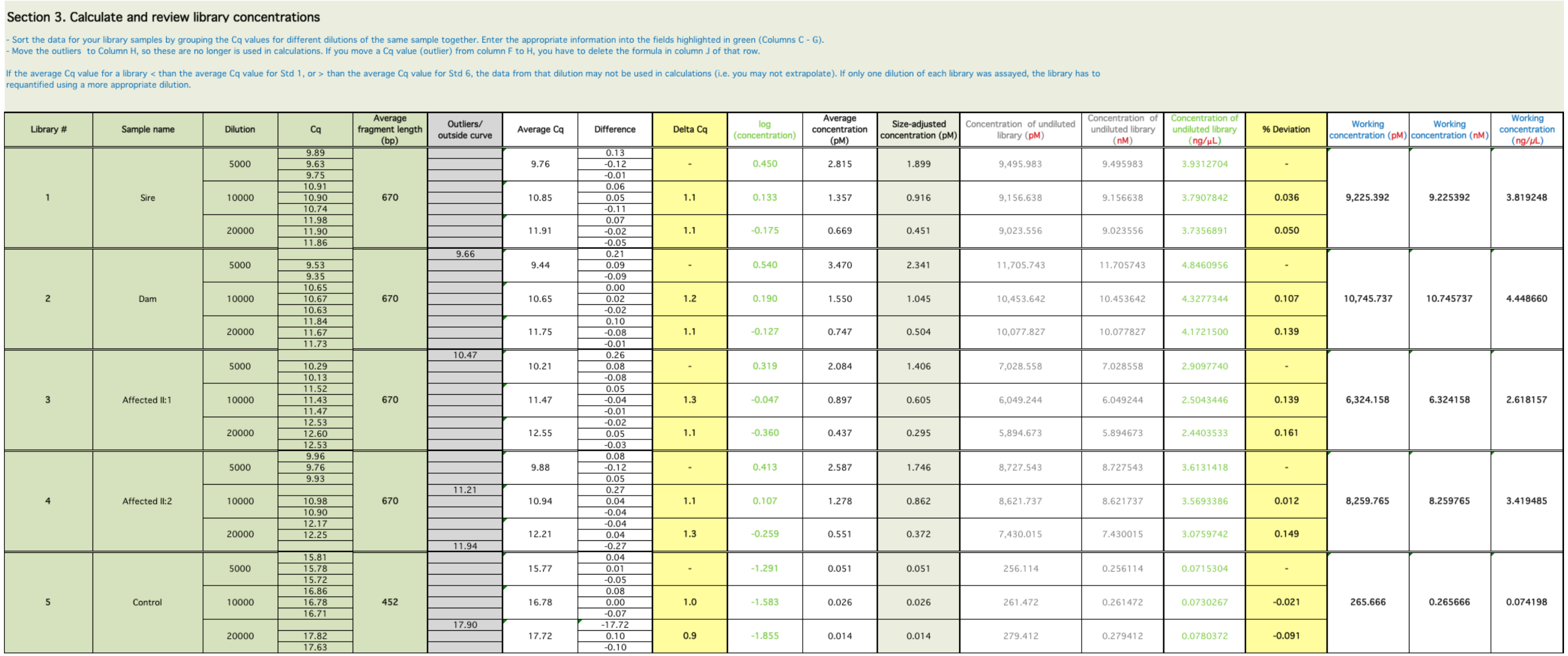
Section 2. Generate and review the standard curve

- Type the value for the intercept from the graph to the right into cell D57.
- Type the value for the slope from the graph to the right into cell D59.

DNA Standard	Conc in pM	Log conc	Average Cq	Delta Cq
1	20.0000	1.30	6.81	-
2	2.0000	0.30	10.30	3.49
3	0.2000	-0.70	13.69	3.40
4	0.0200	-1.70	17.17	3.48
5	0.0020	-2.70	20.64	3.47
6	0.0002	-3.70	24.11	3.47

Should be between 3.1 and 3.6

Efficiency:	95%	(Calculated)	Should be between 90 and 110%
Slope:	-3.4570	(Calculated)	
R-squared:	1.0000	(Calculated)	Should be between 0.99 and 1.00
Intercept:	11.310	(Type the intercept value from the graph in cell D57)	
If slope =	-3.4570	(Type the slope value from the graph in cell D59)	
then efficiency =	95%	(Calculated)	



Section 3. Calculate and review library concentrations

- Sort the data for your library samples by grouping the Cq values for different dilutions of the same sample together. Enter the appropriate information into the fields highlighted in green (Columns C - G).
- Move the outliers to Column H, so these are no longer is used in calculations. If you move a Cq value (outlier) from column F to H, you have to delete the formula in column J of that row.

If the average Cq value for a library < than the average Cq value for Std 1, or > than the average Cq value for Std 6, the data from that dilution may not be used in calculations (i.e. you may not extrapolate). If only one dilution of each library was assayed, the library has to requantified using a more appropriate dilution.

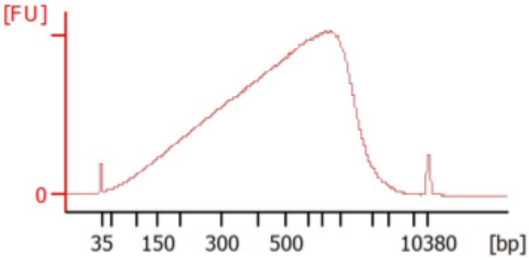
Library #	Sample name	Dilution	Cq	Average fragment length (bp)	Outliers/ outside curve	Average Cq	Difference	Delta Cq	log (concentration)	Average concentration (pM)	Size-adjusted concentration (pM)	Concentration of undiluted library (pM)	Concentration of undiluted library (nM)	Concentration of undiluted library (ng/µL)	% Deviation	Working concentration (pM)	Working concentration (nM)	Working concentration (ng/µL)
1	Sire	5000	9.89	670		9.76	0.13	-	0.450	2.815	1.899	9,495.983	9.495983	3.9312704	-	9,225.392	9.225392	3.819248
			9.63				-0.12											
			9.75				-0.01											
		10000	10.91			10.85	0.06	1.1	0.133	1.357	0.916	9,156.638	9.156638	3.7907842	0.036			
			10.90				0.05											
			10.74				-0.11											
		20000	11.98			11.91	0.07	1.1	-0.175	0.669	0.451	9,023.556	9.023556	3.7356891	0.050			
			11.90				-0.02											
			11.86				-0.05											
2	Dam	5000		670	9.66	9.44	0.21	-	0.540	3.470	2.341	11,705.743	11.705743	4.8460956	-	10,745.737	10.745737	4.448660
			9.53				0.09											
			9.35				-0.09											
		10000	10.65			10.65	0.00	1.2	0.190	1.550	1.045	10,453.642	10.453642	4.3277344	0.107			
			10.67				0.02											
			10.63				-0.02											
		20000	11.84			11.75	0.10	1.1	-0.127	0.747	0.504	10,077.827	10.077827	4.1721500	0.139			
			11.67				-0.08											
			11.73				-0.01											
3	Affected II:1	5000		670	10.47	10.21	0.26	-	0.319	2.084	1.406	7,028.558	7.028558	2.9097740	-	6,324.158	6.324158	2.618157
			10.29				0.08											
			10.13				-0.08											
		10000	11.52			11.47	0.05	1.3	-0.047	0.897	0.605	6,049.244	6.049244	2.5043446	0.139			
			11.43				-0.04											
			11.47				-0.01											
		20000	12.53			12.55	-0.02	1.1	-0.360	0.437	0.295	5,894.673	5.894673	2.4403533	0.161			
			12.60				0.05											
			12.53				-0.03											
4	Affected II:2	5000	9.96	670		9.88	0.08	-	0.413	2.587	1.746	8,727.543	8.727543	3.6131418	-	8,259.765	8.259765	3.419485
			9.76				-0.12											
			9.93				0.05											
		10000			11.21	10.94	0.27	1.1	0.107	1.278	0.862	8,621.737	8.621737	3.5693386	0.012			
			10.98				0.04											
			10.90				-0.04											
		20000	12.17			12.21	-0.04	1.3	-0.259	0.551	0.372	7,430.015	7.430015	3.0759742	0.149			
			12.25				0.04											
					11.94		-0.27											
5	Control	5000	15.81	452		15.77	0.04	-	-1.291	0.051	0.051	256.114	0.256114	0.0715304	-	265.666	0.265666	0.074198
			15.78				0.01											
			15.72				-0.05											
		10000	16.86			16.78	0.08	1.0	-1.583	0.026	0.026	261.472	0.261472	0.0730267	-0.021			
			16.78				0.00											
			16.71				-0.07											
		20000			17.90	17.72	-17.72	0.9	-1.855	0.014	0.014	279.412	0.279412	0.0780372	-0.091			
			17.82				0.10											
			17.63				-0.10											

2. Quality control of the libraries

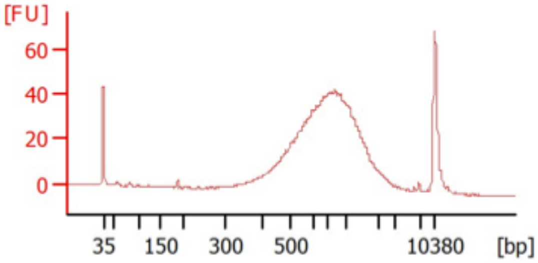
2.1 Run 1

Sire I:1

Fragmented DNA (expected 350 bp)

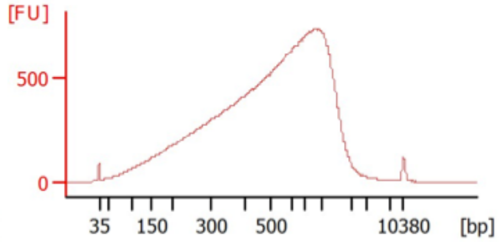


DNA libraries (expected 350 bp)

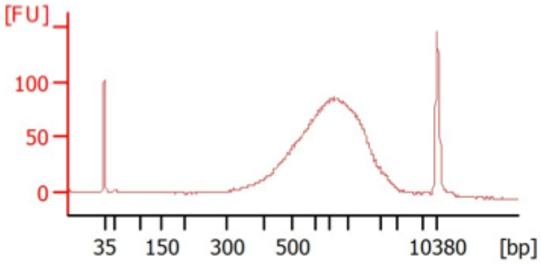


Dam I:2

Fragmented DNA (expected 350 bp)

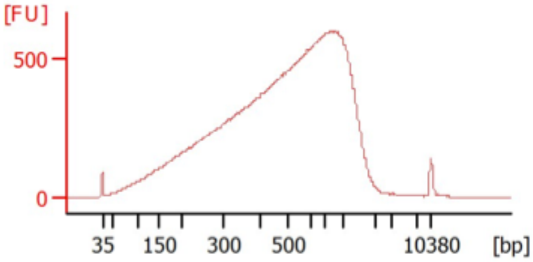


DNA libraries (expected 350 bp)

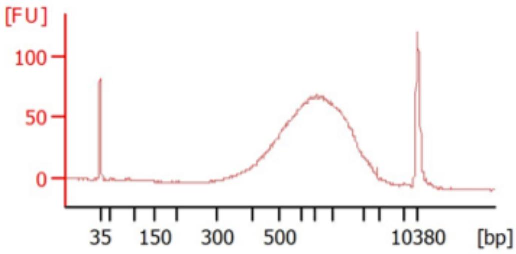


Offspring II:1

Fragmented DNA (expected 350 bp)

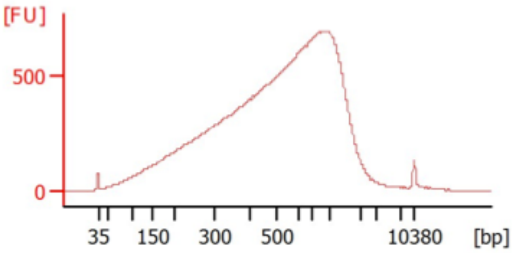


DNA libraries (expected 350 bp)

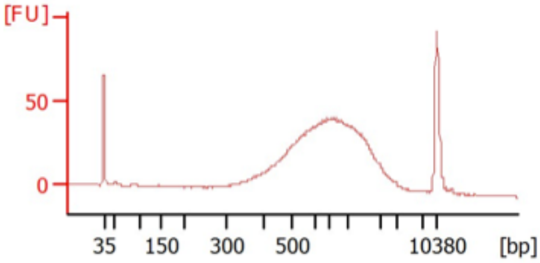


Offspring II:2

Fragmented DNA (expected 350 bp)



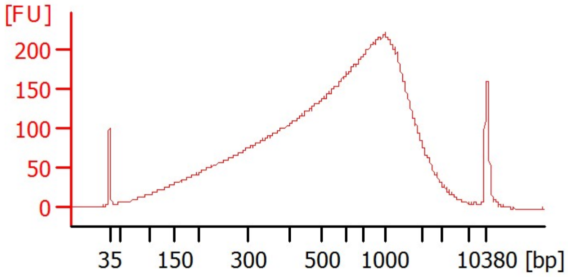
DNA libraries (expected 350 bp)



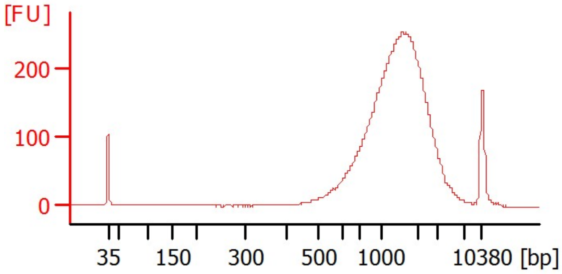
2.2 Run 2

Sire I:1

Fragmented DNA (expected 550 bp)



DNA libraries (expected 550 bp)

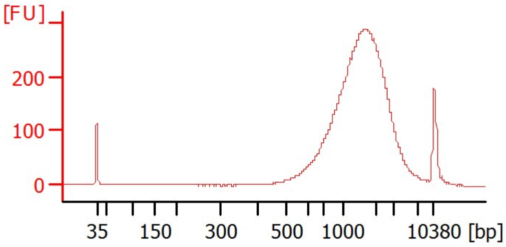


Dam I:2

Fragmented DNA (expected 550 bp)

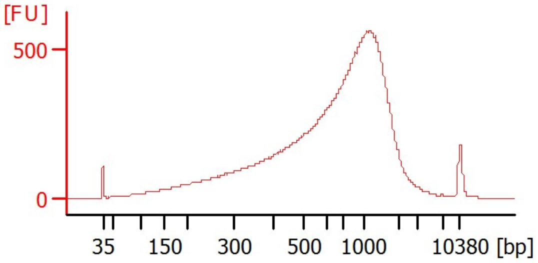
Data Not Available

DNA libraries (expected 550 bp)

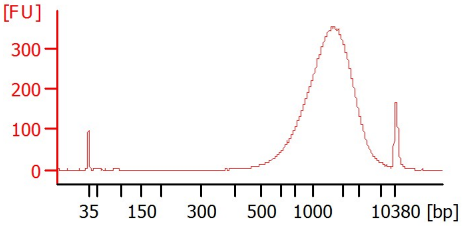


Offspring II:1

Fragmented DNA (expected 550 bp)



DNA libraries (expected 550 bp)

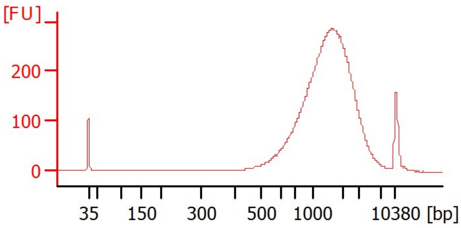


Offspring II:2

Fragmented DNA (expected 550 bp)

Data Not Available

DNA libraries (expected 550 bp)

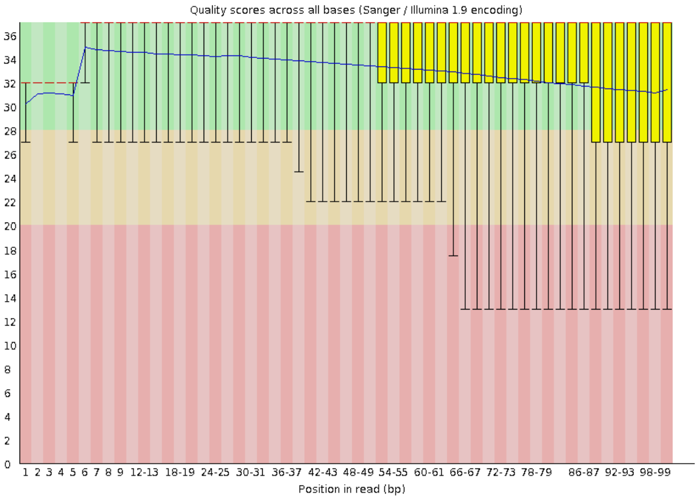


3. Quality control per base sequence quality with FastQC

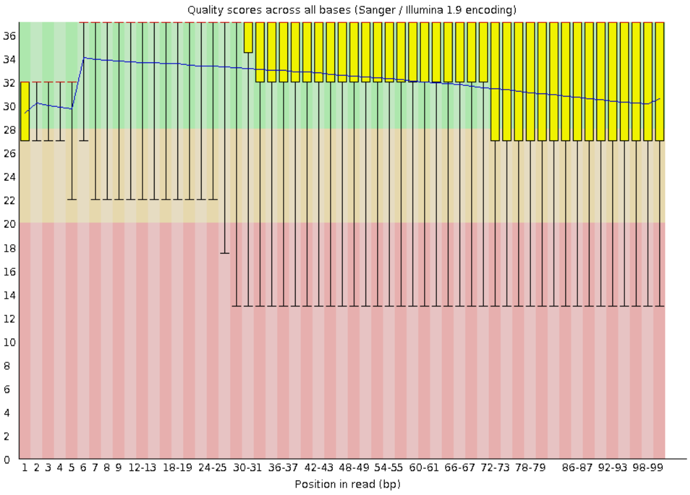
3.1 Run 1 (raw reads)

Sire I:1

Forward reads

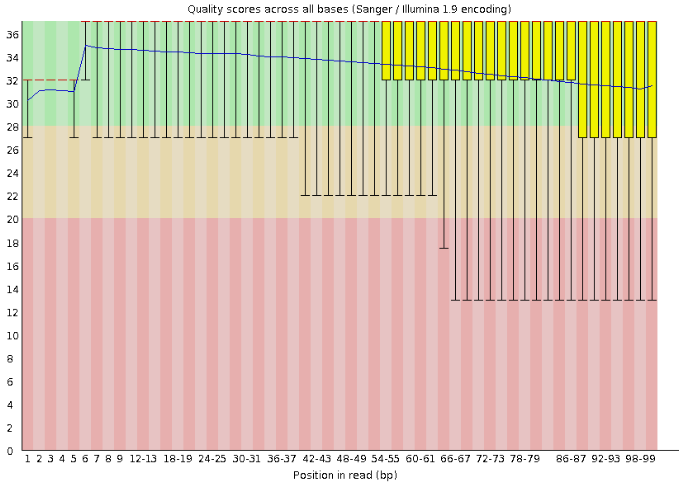


Reverse reads

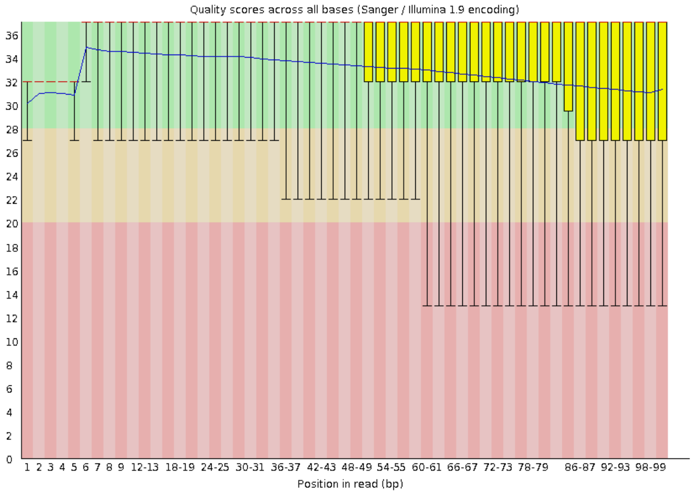


Dam I:2

Forward reads



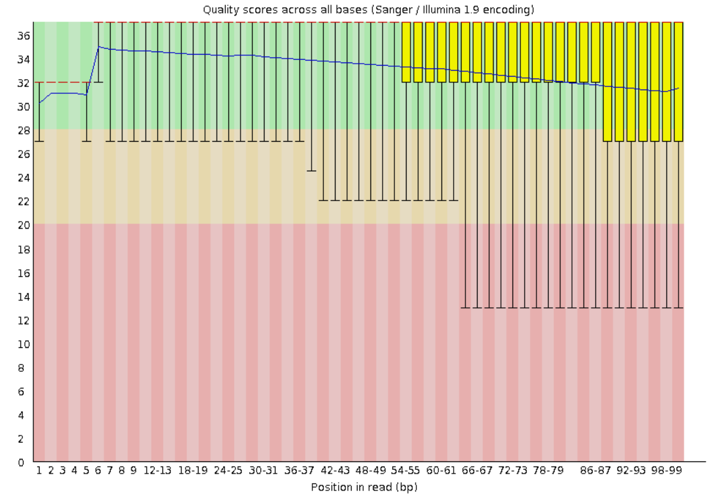
Reverse reads

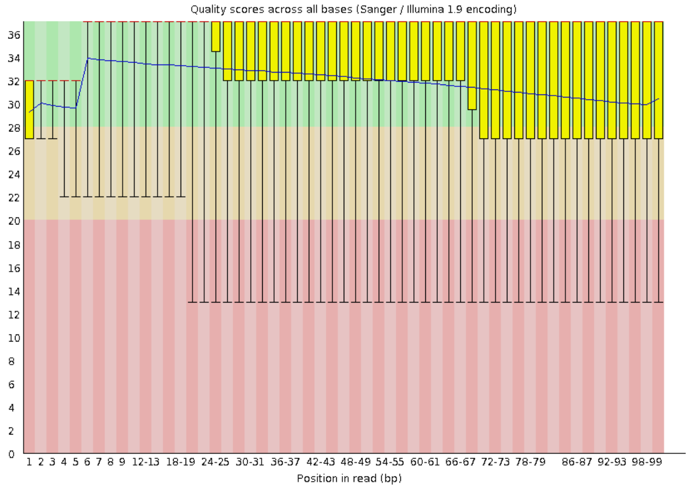


Offspring II:1

Forward reads

Reverse reads

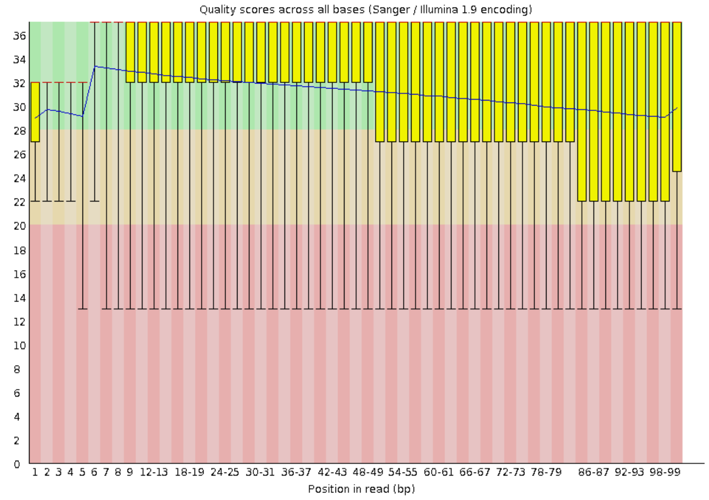


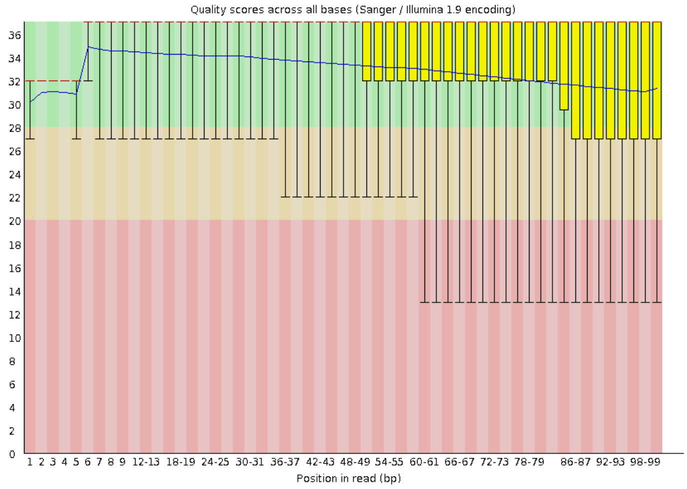


Offspring II:2

Forward reads

Reverse reads



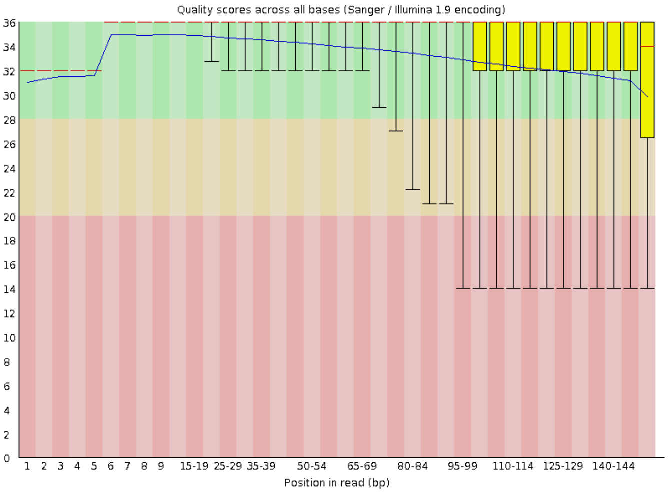


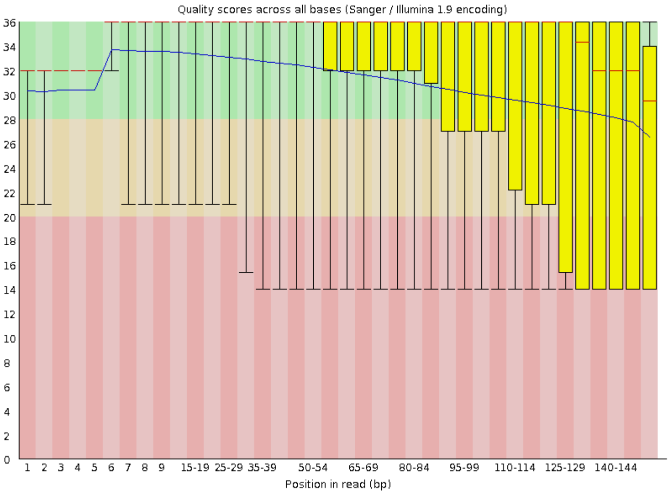
3.2 Run 2 (raw reads)

Sire I:1

Forward reads

Reverse reads

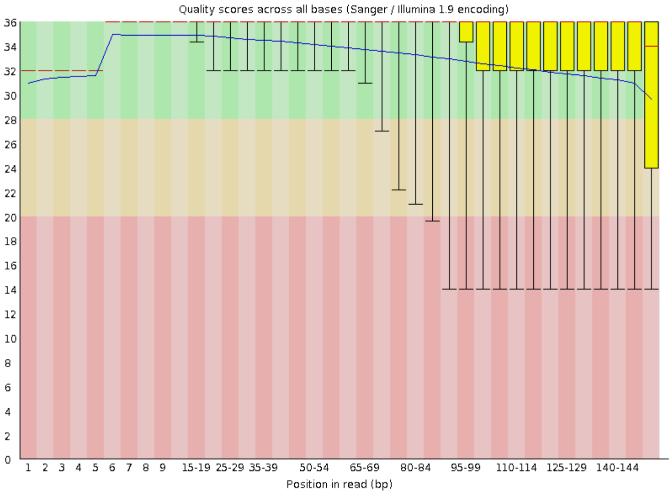


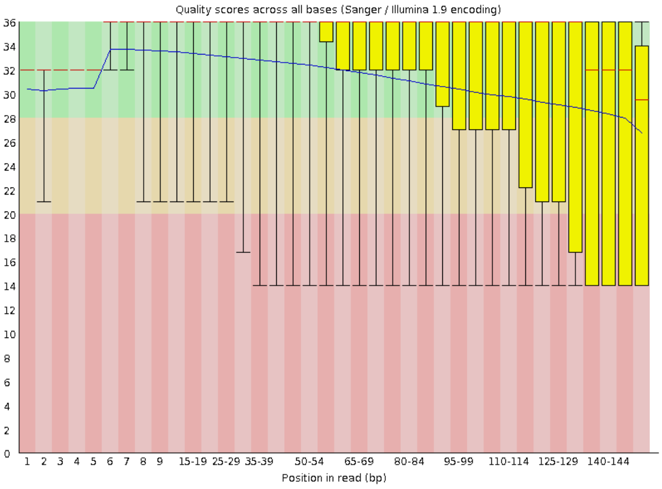


Dam I:2

Forward reads

Reverse reads

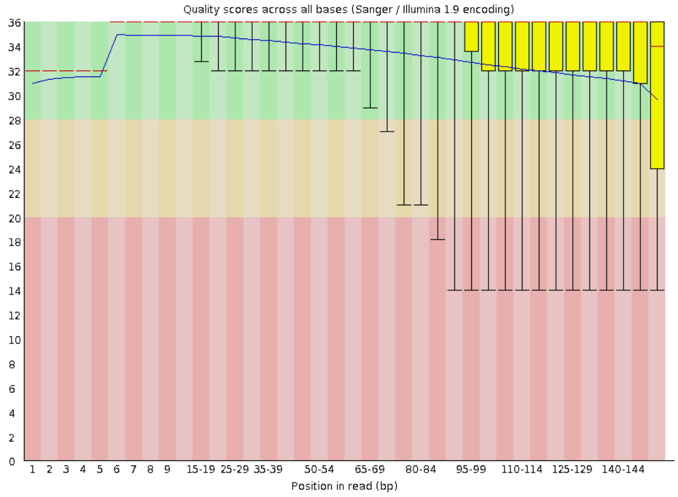


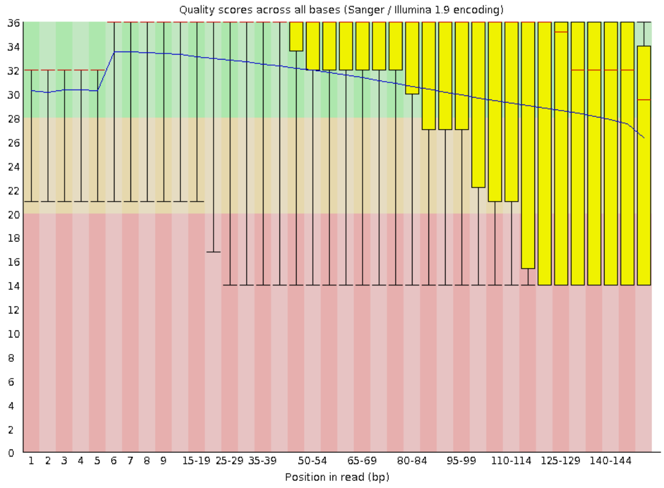


Offspring II:1

Forward reads

Reverse reads

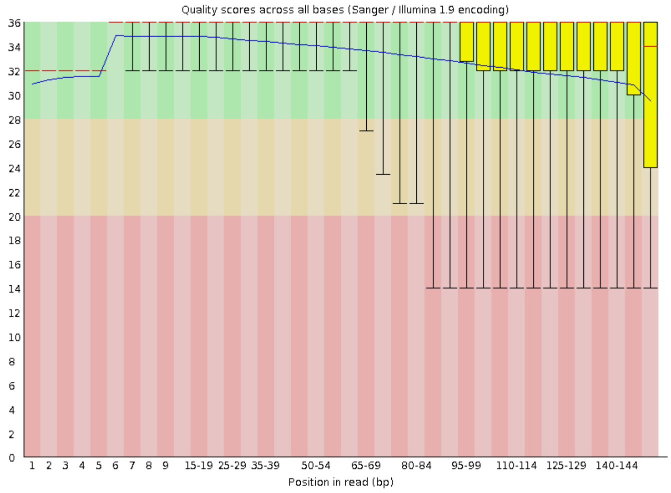


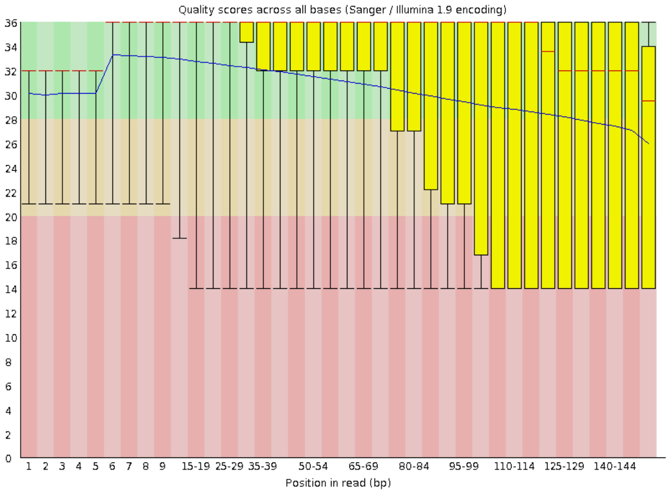


Offspring II:2

Forward reads

Reverse reads

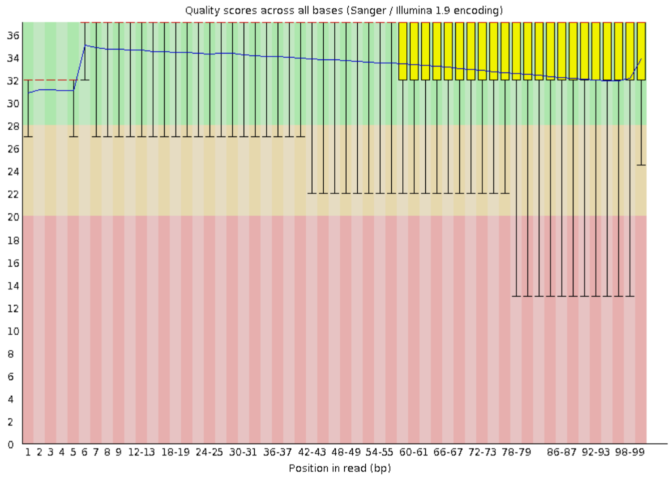




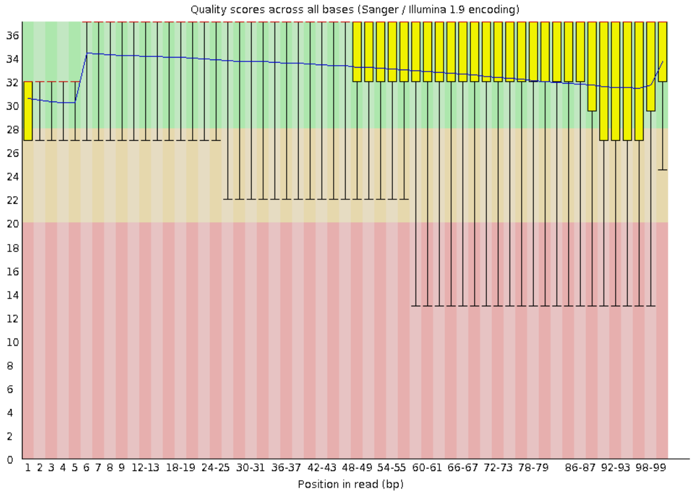
3.3 Run 1 (trimmed reads)

Sire I:1

Forward reads

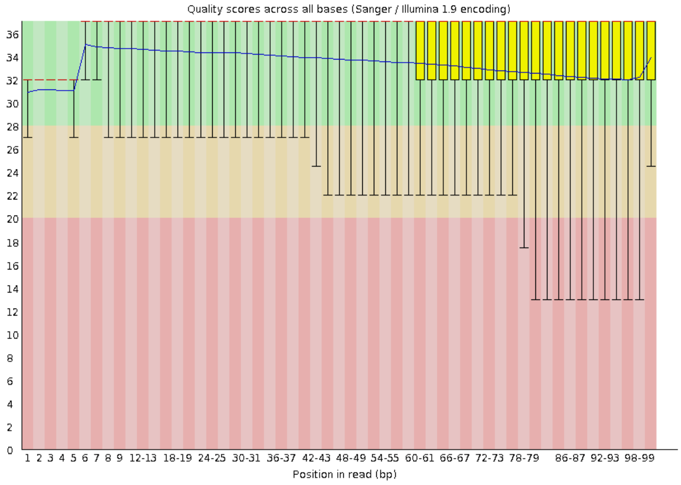


Reverse reads

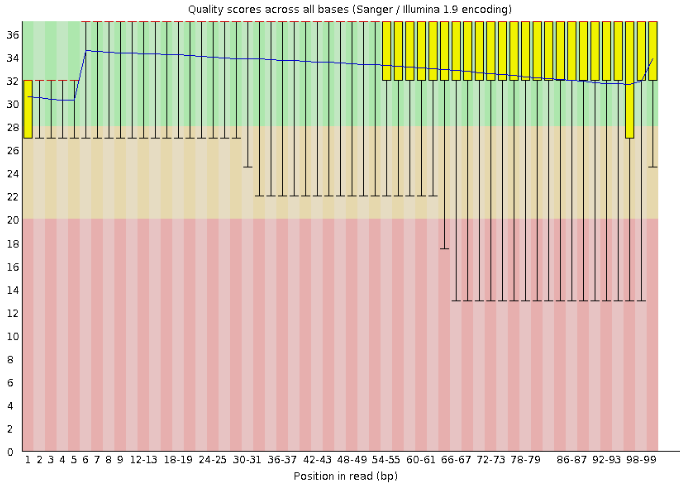


Dam I:2

Forward reads

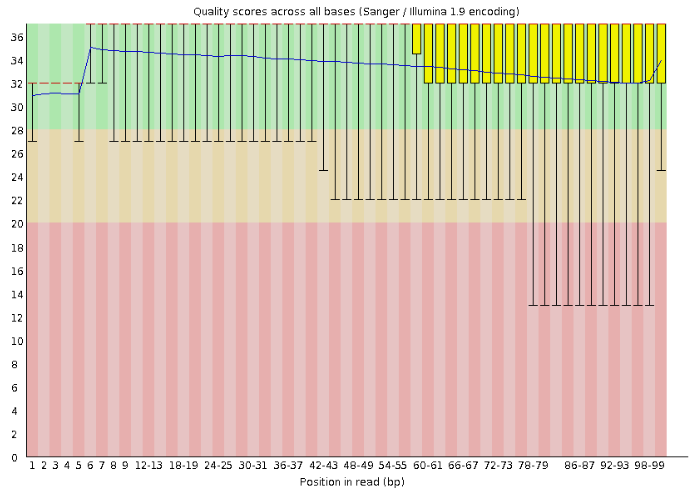


Reverse reads

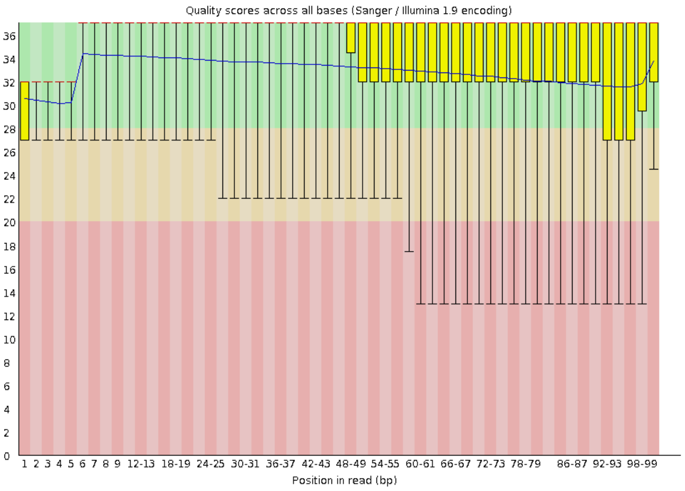


Offspring II:1

Forward reads

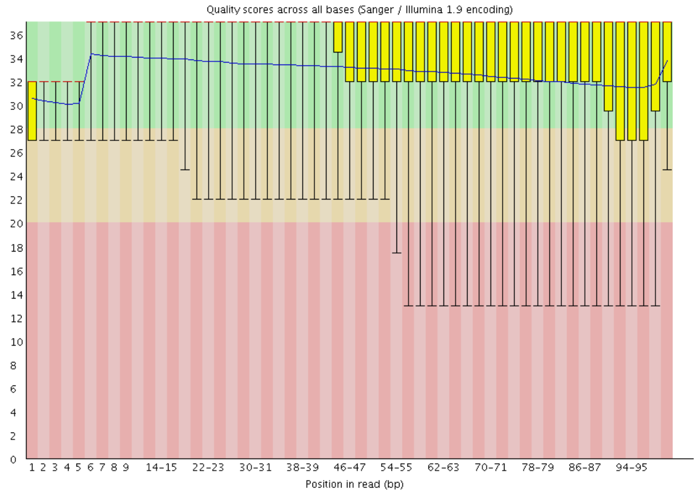


Reverse reads

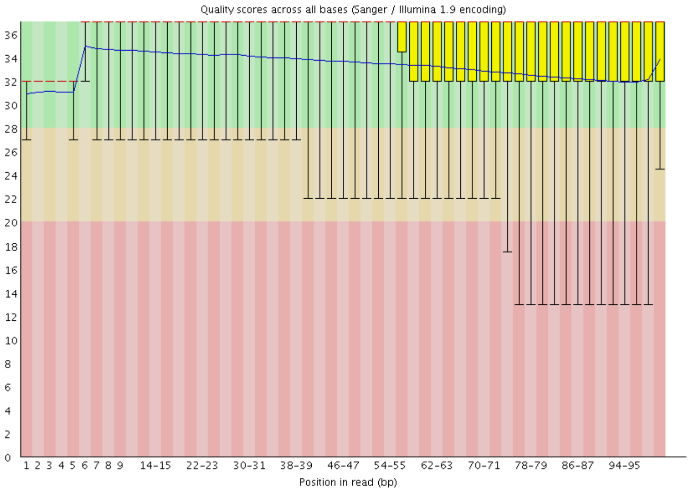


Offspring II:2

Forward reads



Reverse reads

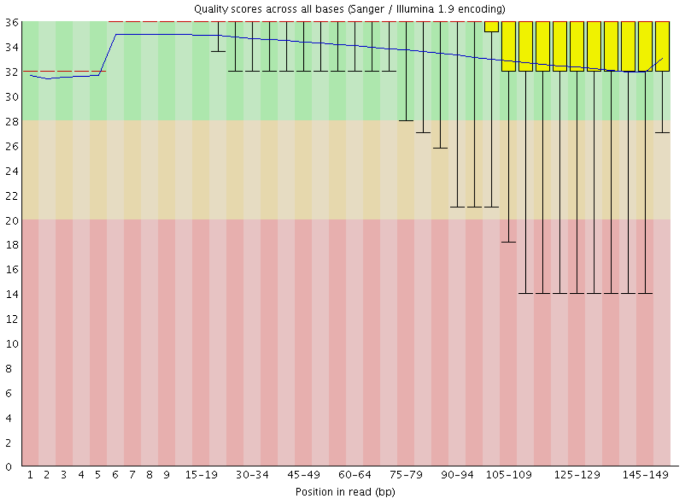


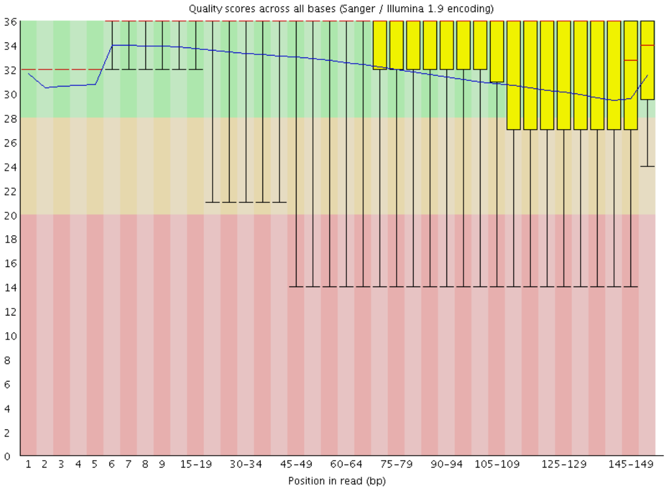
3.4 Run 2 (trimmed reads)

Sire I:1

Forward reads

Reverse reads

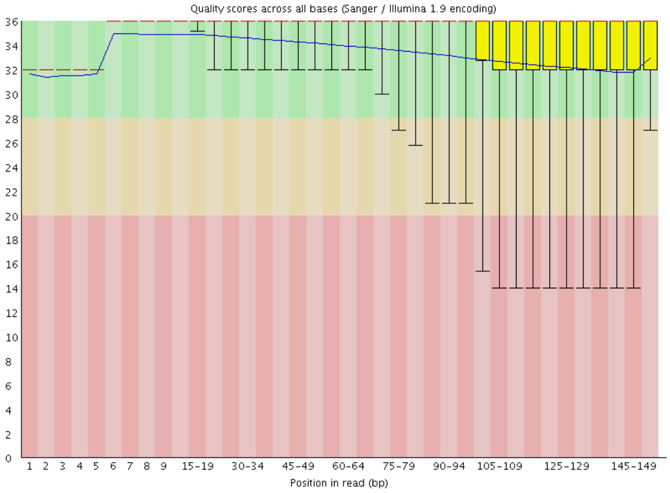


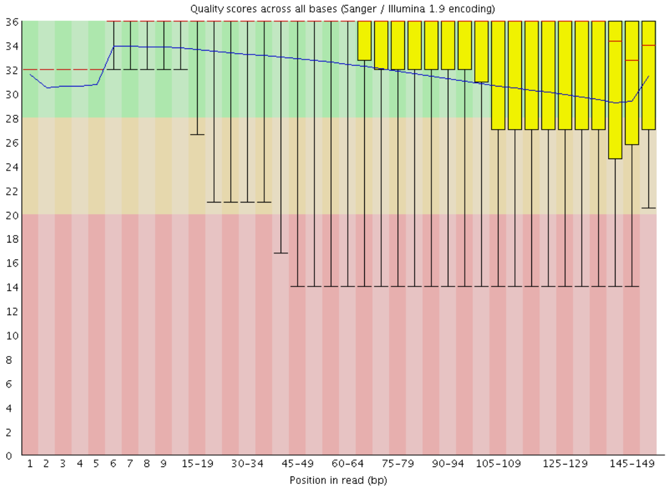


Dam I:2

Forward reads

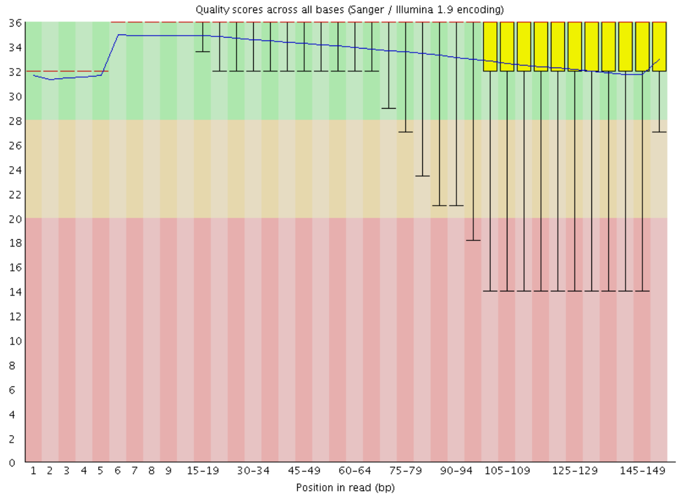
Reverse reads



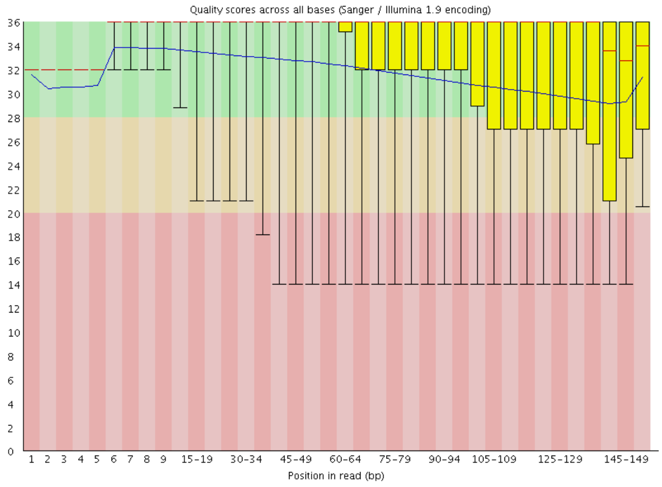


Offspring II:1

Forward reads

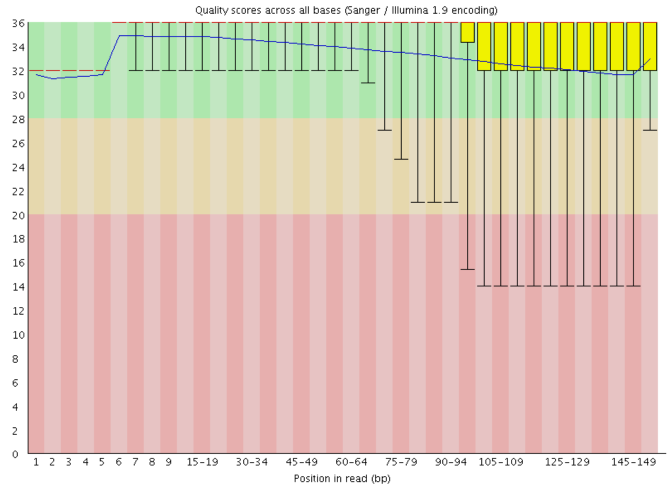


Reverse reads

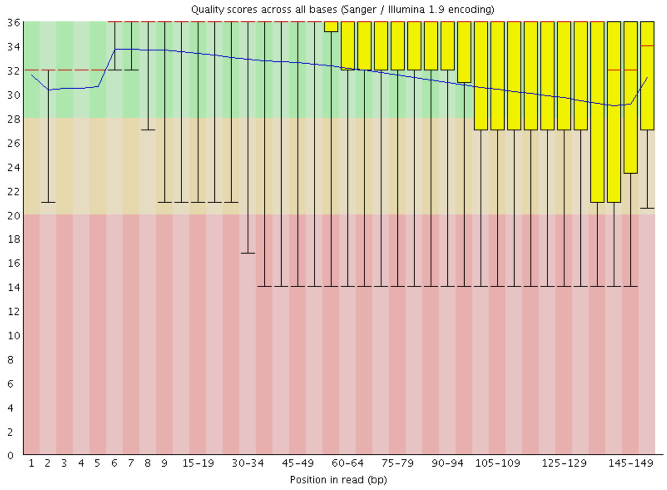


Offspring II:2

Forward reads



Reverse reads



4. Variant filtration

4.1 Candidate SNPs and Indels from run 1

	SNPs							
	Trio 1				Trio 2			
	Total	Right genotypes	Other genotypes	Missing genotype/s	Total	Right genotypes	Other genotypes	Missing genotype/s
Variant function								
exonic	2,597	65	2,260	272	2,597	30	2,265	302
splicing	22	1	17	4	22	1	18	3
ncRNA	-	-	-	-	-	-	-	-
UTR5	185	1	159	25	185	-	149	36
UTR3	770	19	674	77	770	12	660	98
intronic	89,698	1,702	78,415	9,581	89,698	1,967	76,686	11,045
upstream	4,667	46	3,927	694	4,667	66	3,803	798
downstream	3,961	75	3,409	477	3,961	56	3,385	520
intergenic	4,941,638	101,188	4,361,049	479,401	4,941,638	102,587	4,286,204	552,847
Total	5,043,792				5,043,792			
Exonic variant function								
frameshift insertion	-	-	-	-	-	-	-	-
frameshift deletion	-	-	-	-	-	-	-	-
frameshift block substitution	-	-	-	-	-	-	-	-
stopgain	1	-	1	-	1	-	1	-
stoploss	-	-	-	-	-	-	-	-
nonframeshift insertion	-	-	-	-	-	-	-	-
nonframeshift deletion	-	-	-	-	-	-	-	-
nonframeshift block substitution	-	-	-	-	-	-	-	-
nonsynonymous SNV	815	18	714	83	815	8	727	80
synonymous SNV	1,323	39	1,169	121	1,323	19	1,151	153
unknown	458	8	382	68	458	3	386	69
Total	2,597				2,597			

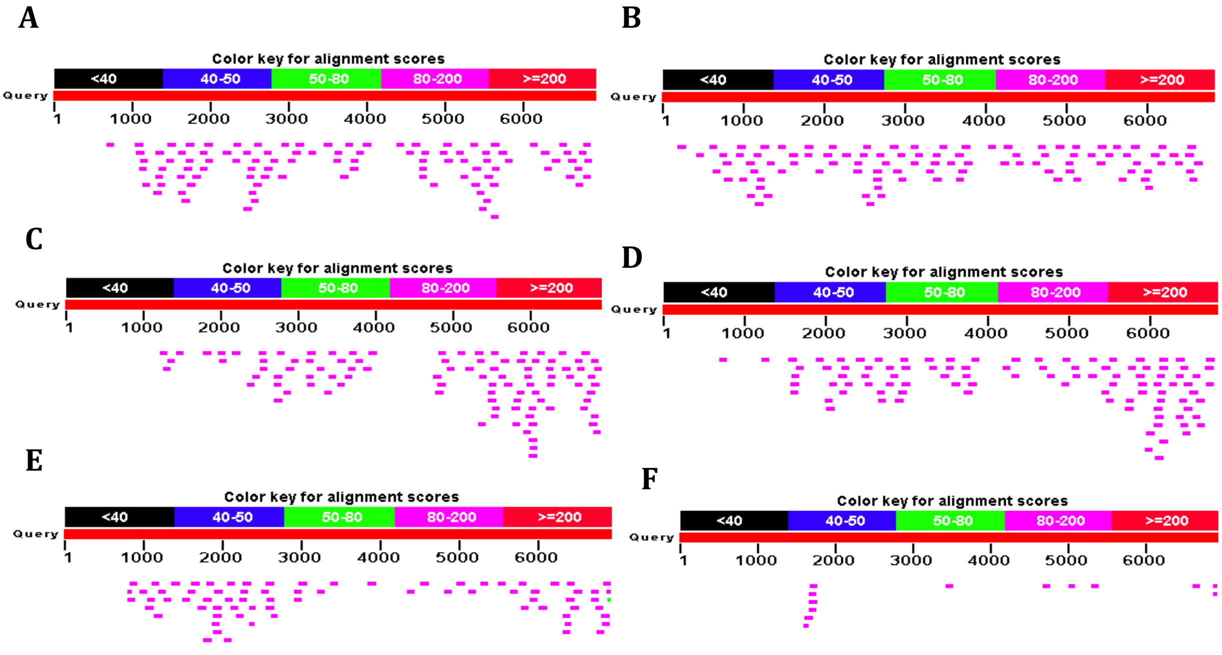
	INDELs							
	Trio 1				Trio 2			
	Total	Right genotypes	Other genotypes	Missing genotype/s	Total	Right genotypes	Other genotypes	Missing genotype/s
Variant function								
exonic	124	1	95	28	124	1	88	35
splicing	36	-	32	4	36	-	32	4
ncRNA	-	-	-	-	-	-	-	-
UTR5	48	-	38	10	48	1	41	6
UTR3	226	1	187	38	226	-	190	36
intronic	27,589	291	21,866	5,432	27,589	394	21,244	5,951
upstream	1,318	6	990	322	1,318	14	993	311
downstream	1,048	5	849	194	1,048	14	810	224
intergenic	1,416,368	18,730	1,140,039	257,599	1,416,368	18,938	1,108,771	288,659
Total	1,446,788				1,446,788			
Exonic variant function								
frameshift insertion	19	-	16	3	19	-	13	6
frameshift deletion	13	-	11	2	13	-	12	1
frameshift block substitution	-	-	-	-	-	-	-	-
stopgain	-	-	-	-	-	-	-	-
stoploss	-	-	-	-	-	-	-	-
nonframeshift insertion	12	-	10	2	12	-	11	1
nonframeshift deletion	17	-	14	3	17	-	13	4
nonframeshift block substitution	-	-	-	-	-	-	-	-
nonsynonymous SNV	-	-	-	-	-	-	-	-
synonymous SNV	-	-	-	-	-	-	-	-
unknown	63	1	44	18	63	1	39	23
Total	124				124			

4.2 Candidate SNPs and Indels from the merged reads

	SNPs							
	Trio 1				Trio 2			
	Total	Right genotypes	Other genotypes	Missing genotype/s	Total	Right genotypes	Other genotypes	Missing genotype/s
Variant function								
exonic	3,186	120	2,983	83	3,186	74	3,036	76
splicing	31	1	23	7	31	1	26	4
ncRNA	-	-	-	-	-	-	-	-
UTR5	212	4	204	4	212	5	192	15
UTR3	889	39	840	10	889	19	854	16
intronic	102,530	3,709	96,067	2,754	102,530	3,836	95,649	3,045
upstream	5,621	144	5,139	338	5,621	169	5,116	336
downstream	4,719	169	4,447	103	4,719	136	4,403	180
intergenic	5,666,220	192,148	5,339,183	134,889	5,666,220	192,192	5,322,502	151,526
Total	5,783,790				5,783,790			
Exonic variant function								
frameshift insertion	-	-	-	-	-	-	-	-
frameshift deletion	-	-	-	-	-	-	-	-
frameshift block substitution	-	-	-	-	-	-	-	-
stopgain	-	-	-	-	-	-	-	-
stoploss	-	-	-	-	-	-	-	-
nonframeshift insertion	-	-	-	-	-	-	-	-
nonframeshift deletion	-	-	-	-	-	-	-	-
nonframeshift block substitution	-	-	-	-	-	-	-	-
nonsynonymous SNV	996	37	941	18	996	16	963	17
synonymous SNV	1,645	68	1,547	30	1,645	51	1,569	25
unknown	545	15	495	35	545	7	504	34
Total	3,186				3,186			

	INDELs							
	Trio 1				Trio 2			
	Total	Right genotypes	Other genotypes	Missing genotype/s	Total	Right genotypes	Other genotypes	Missing genotype/s
Variant function								
exonic	160	5	142	13	160	3	147	10
splicing	54	1	49	4	54	-	47	7
ncRNA	-				-	-	-	-
UTR5	60	1	53	6	60	1	55	4
UTR3	300	5	277	18	300	2	283	15
intronic	35,974	875	32,355	2,744	35,974	978	31,897	3,099
upstream	1,766	38	1,523	205	1,766	48	1,497	221
downstream	1,429	31	1,281	117	1,429	32	1,260	137
intergenic	1,836,191	48,110	1,654,064	134,017	1,836,191	48,757	1,636,303	151,131
Total	1,875,989							
Exonic variant function								
frameshift insertion	24	1	18	5	24	1	20	3
frameshift deletion	20	-	19	1	20	-	19	1
frameshift block substitution	-	-	-	-	-	-	-	-
stopgain	-	-	-	-	-	-	-	-
stoploss	-	-	-	-	-	-	-	-
nonframeshift insertion	19	-	17	2	19	-	18	1
nonframeshift deletion	23	-	22	1	23	1	20	2
nonframeshift block substitution	-	-	-	-	-	-	-	-
nonsynonymous SNV	-	-	-	-	-	-	-	-
synonymous SNV	-	-	-	-	-	-	-	-
unknown	74	4	66	4	74	1	70	3
Total	160				160			

5. Expression *ABCA4*



Alignment of *ABCA4* cDNA against different RNA-seq output data from different tissues. The cDNA sequence from *ABCA4* (ENSCAFT00000032029) presented 6,812 bp and was blasted with blastn (Camacho et al. 2009) against few dog tissues . **A.** Alignment with brain tissue (SRX11063). **B.** Alignment with kidney tissue (SRX11061). **C.** Alignment with ovary tissue (SRX11066). **D.** Alignment with testis tissue (SRX111069). **E.** Alignment with blood (SRX11070). **F.** Alignment with skin tissue (SRX11064).