Genetic Basis for Inherited Eye Diseases in Dogs: A Case Study of Pigmentary Chorioretinopathy in Chinese Crested Dogs

Merina Shrestha

Norwegian University of Life Sciences
Department of Animal and Aquacultural Sciences
Master Thesis 30 credits 2012

Genetic Basis for Inherited Eye Diseases in Dogs: A Case Study of Pigmentary Chorioretinopathy in Chinese Crested Dogs

Merina Shrestha

Registration number: 860203-P240 (SLU)/977321 (UMB)

Thesis Animal Breeding and Genetics

Course Code: EX0556 (SLU)/ M30-IHA (UMB)

18th June, 2012

Department of Animal Breeding and Genetics

SUPERVISORS

Tomas Bergström (SLU)

Göran Andersson (SLU)

Kristina Narfström (SLU)

Dag Inge Våge (UMB)

Education and Culture

Erasmus Mundus

ACKNOWLEDGEMENT

Firstly, I am very grateful to all the people involved in **EMABG program** for this great opportunity and funding my study. It has been an honour to be a member of this program which has not only broadened my knowledge in academic section but equally assisted me to mature in every other aspect.

I would like to express my gratitude to my main supervisor **Tomas Bergström** for motivating and supporting me in all those times when I was immersed in a confused state of mind, with his friendly, calm and optimistic attitude. Again, I would like to thank Tomas and other supervisors **Göran Andersson, Kristina Narfström** and **Dag Inge Våge** for guidance and encouragement throughout the period.

I would also like to thank **Marcin Kierczak** for providing me with all the knowledge regarding R software and statistical concepts. Also, I would like to acknowledge all the people in **animal science lab** who took their time and assisted me with my lab work.

I would like to appreciate the time and effort **Tormod Ådnøy** (associate professor, University of Life Sciences-UMB) spent to assist me in my first step into the field of animal breeding and genetics. Also, I would like to thank my friends who filled my journey with fun and interesting events.

ABSTRACT

A novel inherited retinal disease, pigmentary chorioretinopathy, has been observed in one of the oldest purebred breed, Chinese crested. Two forms of progressive retinal atrophy (PRA), Progressive rod con degeneration (prcd) and one other form with unknown genetic basis, has already been observed in the breed. This novel retinopathy showed clinical features of progressive nature with bilateral degeneration, migration of lesions from tapetal to non-tapetal and central area in fundus and end stage leading to blindness. A distinct feature of primary defect in retinal pigment epithelium differentiates this disease from PRA. A genome-wide association study with 170K Illumina CanineHD SNP chip was performed using 19 cases and 21 controls. After correcting for population stratification, applying a combined approach of Mixed model and Structured association in R (GenABEL package), an association analysis using 14 cases and 21 controls resulted in a stronge association with two SNPs (P-raw: 4.59e-06 and 5.74e-06) located around 300 kb apart, on chromosome 8. After further analysis in Ensembl (CanFam 2.0), we found that one of the SNPs was located in an intronic sequence of the gene *MDGA2* (MAM domain containing glycosylphosphatidylinostitol anchor 2 - 371.42 kb) and the other was located downstream to this gene. *MDGA2* is a member of the immunoglobulin superfamily class (IgCAM) and is involved in cell adhesion, migration and recruitment to inflammatory sites. Sequencing of the coding region of *MDGA2* revealed a non-synonymous mutation that caused an amino acid shift from “threonine” to “serine”. The sequence analysis was inconclusive; hence more individuals need to be sequenced for a definite conclusion. Also, a manual gene annotation of *MDGA2* is required for a mutation to be concluded as not causative. Hence, additional studies need to be performed to unravel the genetic basis of the disease which will guide us to the possible preventing strategy in near future.

Keywords: Retinopathy, Retinal disease, Genome wide association study, Chinese crested, Pigmentary chorioretinopathy, Retinal pigment epithelium (RPE)

TABLE OF CONTENTS

INTRODUCTION

Man's best friend, the domestic dog has proven to be a valuable model for investigating the genetic basis underlying canine and human diseases (Karlsson and Lindblad-Toh, 2008). Humans and dogs share common physiological traits, have often similar clinical responses, similar disease presentation and also live in the same environment as humans do. The domestic dog has gone through several bottlenecks. The first bottleneck occurred during the domestication from wolves around 7,000 -50,000 generations ago (Lindblad-Toh et al., 2005). The establishment of dog breeds in the mid 1800s has led to further loss of genetic variation in each breed and has resulted in unique features and breed-specific phenotypes. This has caused increased risk for developing a variety of genetic diseases, some that appears to be breed-specific.

Canine Inherited Eye Diseases:

The number of reported hereditary disorders in domestic dogs exceeds 400 (Giger et al., 2006). With the advancement in genetic approaches and the development of molecular genetic tools and approaches to identify underlying causative loci for a disease, this number has further been increasing. To this date, 24 mutations in 18 genes underlying retinal diseases (retinopathy) in at least 58 dog breeds have been identified (Miyadera et al., 2012a).

Retina in dogs is undeveloped at birth and matures at 3-6 weeks after birth, but the structure of developed retina in dogs is similar to retina in human (Miyadera et al., 2012b). The retina is a neuronal portion of the eye. The retina consist of five different neuronal cell layers: photoreceptors (rods and cones) located in the outer nuclear layer (ONL), bipolar cells, horizontal cells and amacrine cells located in the inner nuclear layer (INL) and the innermost cell layer of ganglion cells (Doh et al., 2010).

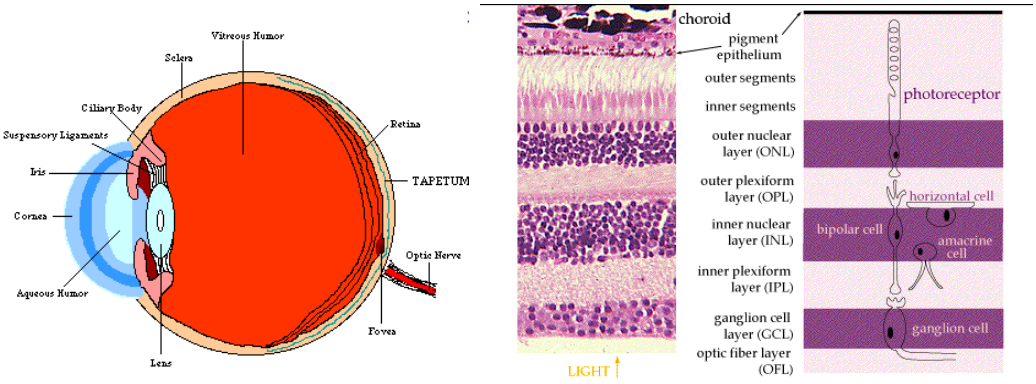


Figure 1 Retinal layers (from http://eyesee-eyetalk.blogspot.se/2007/12/231-worn-and-torn-retina.html). To the left is the sketch of dog eye with general anatomy. The right picture gives a view of different cell layers that describes retina.

The photoreceptor cells are: rods (vision for dark and dim light) and cones (vision for day light and colour). The rods are present in large numbers compared to cones in both humans and dogs. The photoreceptor layer is further divided into outer segment (OS), which lies close to the pigment epithelium and consist of a membranous disk, and inner segment (IS) that consists of cell

nuclei. The bipolar, horizontal and amacrine cells are retinal neurons responsible for lateral interactions from photoreceptor cells to the ganglion cells (Purves et al., 2001). The information is transferred from photoreceptor cells to bipolar cells to ganglion cells which transfer the information to the optic nerve.

The retinal pigment epithelium (RPE) is a cell monolayer that contains melanin that reduces the backscattering of the lights and acts as a barrier between choroid and the photoreceptor cells (Purves et al., 2001). RPE cells are involved in retinoid cycle in regeneration of the light absorbing chromophore (11-*cis*-retinal). Also, RPE cells participates in degradation and recycling of rods and cones in outer segments through phagocytosis (Kuksa et al., 2003).

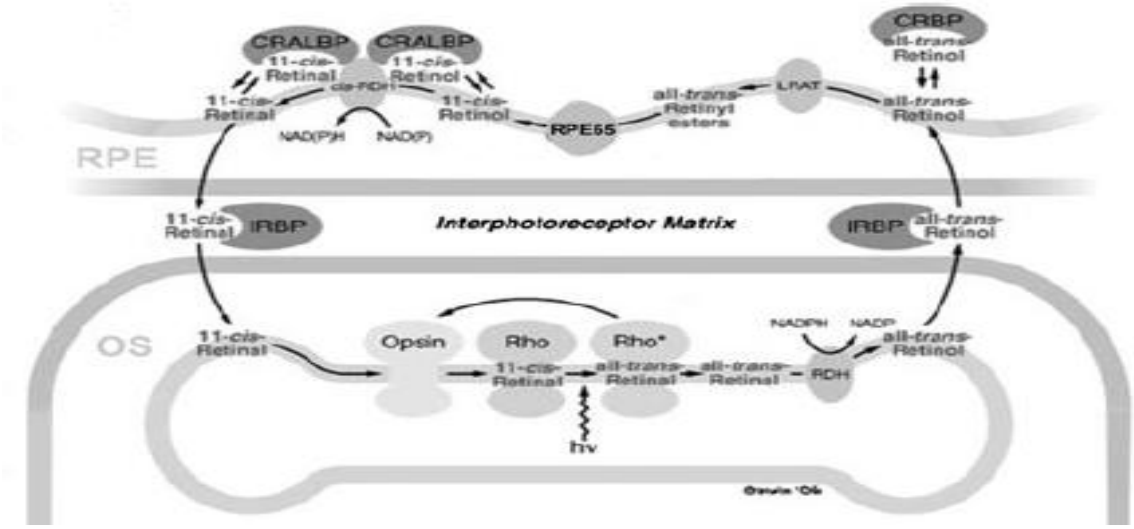


Figure 2. Retinoid cycle (Miyadera et al., 2012a). Outer segment (OS) of photoreceptor cells contain Opsin (rhodopsin-rods and photopsin-cones) that is bound to 11-cis-retinal (light absorbing chromophore). 11-cis-retinal is converted to all-trans-retinal by light (photon) that initiates phototransduction cycle (light stimuli is converted into chemical signals). All-trans-retinal is converted to all-trans-retinol and transported to RPE. Esterification and oxidization process convert all-trans-retinol to 11-cis-retinal that binds to opsin in outer segment ready to begin next cycle.

Inherited retinal diseases can be classified as progressive, stationary and developmental (Miyadera et al., 2012a). Inherited progressive disorders primarily affect photoreceptor cells, show progressive ophthalmoscopic degenerative changes with variable rate and is breed specific. Progressive retinal atrophy (PRA) and Cone rod dystrophy (CRD) are examples of inherited progressive disorders. Stationary retinal disorder shows very slow progression or non progressive abnormalities after the initial diagnosis of the disease. Achromatopsia (cone degeneration, hemeralopia), canine Leber congenital amaurosis LCA (congenital stationary night blindness-csnb) and retinal dystrophy are included in stationary retinal disorders. The developmental (non degenerative) retinal disorders include syndromic retinal dysplasia (defect in retinal differentiation): oculoskeletal dysplasia 1 (in Labrador retrievers) and 2 (in Samoyed), and asyndromic retinal dysplasia (Bedlington terriers and Labrador retrievers) with unknown genetic causes. It also includes Collie eye anomaly (CEA) that affects the retina-choroid-sclera complex, the main clinical feature being choroidal hypoplasia (under-development of choroid).

Among many inherited retinal disorders mentioned above, PRA is a heterogeneous disease that has been described in more than 100 breeds (Vilboux et al., 2008, Aguirre and Acland, 2006).

PRA is a group of inherited retinopathies that share general clinical ocular abnormalities and the primary feature being a loss of rods followed by loss of cone photoreceptors. The bilateral nature of PRA progresses with similar retinal changes of increase in severity on both eyes in a parallel way and end stage retinal atrophy leading to blindness.

Pigmentary Chorioretinopathy:

The focus of this study is to find the genetic basis of a novel inherited retinal disease that has been termed pigmentary chorioretinopathy in Chinese crested dog (CCD). The breed is one of the oldest purebred dogs and in recent years has become increasingly popular in Sweden (among the top 10 most popular breed in Sweden in 2010). Two different varieties of CCD can be observed: a hairy phenotype known as "Powder puff" or wild type and hairless (mutant) phenotype. The hairless phenotype in CCD has a crest of fur on head, tail and fur around the feet and this phenotype arise due to a monogenic autosomal semidominant trait, Canine ectodermal dysplasia (CED) (Narfström et al.).

Earlier, progressive rod cone degeneration (prcd) mutation has been found in CCD as one of the two forms of PRA. The genetic basis for the other type is not known (Optigen). However, CCD involved in this study showed negative results for the prcd mutation (Zangerl, Goldstein et al. 2006).

The clinical features of pigmentary chorioretinopathy include bilateral nature of degeneration, appearance of lesions/pigmentation in the periphery of tapetal fundus in earlier stage that migrate to central and non-tapetal fundus as the disease progresses leading to blindness (Narfström et al.). Difference in electroretinographic responses and defects in arterio-venous phase were more visible in later stages of the disease. The foremost changes are observed in RPE cells that appeared thick and contained pigment granules. On histology, this monolayer appeared as degenerating multilayer in some areas. RPE cells were detached and migrated towards the outer segment. The phagocytosis cells and pigment granules were observed along with RPE cells in the outer segment. Such abnormal changes in RPE were observed before degeneration of rods and cones which distinguishes current retinopathy in CCD from classical PRA.

Primary abnormal change in RPE is also an essential feature in Canine multifocal retinopathy (CMR) observed in a number of dog breeds such as English Mastiff, Bullmastiff, Great Pyreness and Coton de Tulear. Two different mutations in *VMD2* gene: 73C>T stop mutation in cmr1 and 482G>A missense mutation in cmr2 are responsible for CMR (Guziewicz et al., 2007). Inherited retinal dystrophy also shows primary changes in RPE layer. A homozygous 4bp deletion: 485delAAGA in exon 5 of *Rpe65* gene (encodes a microsomal protein in RPE) leadingto frame shift mutation has shown to cause retinal dystrophy in Swedish briard/briard-beagle dogs (Veske et al., 1999). The primary defect in RPE is also seen in Retinal pigment epithelial dystrophy (RPED) with questionable genetic basis as the features of this disease is also seen in the dogs with vitamin E deficiency.

Genome Wide Association Studies (GWAS):

The objective of this study is to identify potential genomic loci associated with pigmentary chorioretinopathy with a Genome-Wide Association Study (GWAS) approach. In general, GWAS is used to find common causal genetic variants underlying disease through use of high density SNP genotyping arrays. It is based on the concept of common disease common variant (CDCV). Balding (2006) explains CDCV as the hypothesis that genetic variants underlying complex diseases are common and can be detected with current population association study designs. GWA study includes the benefits of linkage studies which do not need prior knowledge of potential region of susceptibility, and advantages of association methods that is more powerful at detecting genes of small effect (Dick, 2008).

The correlation between a marker allele (single nucleotide polymorphism- SNP) and a certain disease is evaluated by comparison of allele or genotype frequencies among individuals in different kinds of study designs such as case-control, multistage and family-based design. Case-control design is generally powerful for detecting association (Dick, 2008). This evaluation involves comparison of marker allele frequency between cases and controls using statistical methods. However, the main problem in case-control design is rise of spurious association (Pearson and Manolio, 2008), a significant association without any genetic effect. Many confounding factors such as population structure, cryptic relatedness/family structure, multiple testing, differential bias give rise to this problem.

Population structure is the major confounding factor in case-control design for spurious association. An example of spurious association was observed between the trait height with a SNP in a gene LCT (lactase gene) which varies widely across European population (Campbell et al., 2005). When the population under study includes subgroups such that individuals in subgroups are more related to one another on average than to other individuals in a population, such population is referred to as structured (Balding, 2006). Stratification may arise due to systematic ancestry differences and may have large impact when cases and controls are sampled from two different subpopulations. To overcome such circumstances the following assumptions need to be applied: unrelated cases and unrelated but matched controls should be drawn from same population ('unrelated' refers to the unknown relationships which is presumed to be distant), affected individuals are representative of all cases of the disease, genomic and epidemiologic data for cases and control should be collected similarly. Observed differences in allele frequencies is due to the outcome of certain trait of interest and not due to differences in background population between cases and controls (Pearson and Manolio, 2008).

Another important issue in case-control studies is cryptic relatedness, a hidden kinship among cases or controls (Voight and Pritchard, 2005). Devlin and Roeder (1999) consider cryptic relatedness to have serious impact compared to population structure. Voight and Pritchard (2005) provided only certain scenarios where cryptic relatedness might have a greater impact. A greater impact will be observed when a small study population has been derived from the population which had a recent and rapid growth. Also, in a population with extensive inbreeding the relatedness among individuals in the study population will make GWAS approach less likely to be successful. In such cases, the genetic factors controlling the trait may have reached fixation in the population.

The presence of confounding effects in the analysis is evaluated by calculating variance inflation factor "λ". It is calculated as the median chi-square association statistic across SNPs divided by its theoretical median under null distribution (Price et al., 2010). $\lambda > 1$ indicates stratification or

presence of any other confounding effects, $\lambda \sim 1$ (approximately) indicates absence of any confounding factors. Also, λ is proportional to the sample size.

The basic association test to search for the potential genomic loci associated with the disease includes a chi-square test based on 2*2 contingency table of allele counts or 2*3 contingency table of genotype counts for case and control group. The relative frequency of allele or genotype is expected to be same in the case and control groups under the null hypothesis (no association). The independence of rows and columns in a contingency table is evaluated by chi-square test (Clarke et al., 2011).

Approaches such as structured association and principal component analysis (PCA) infer genetic ancestry and accounts for population stratification. In structured association, cluster-based algorithm is applied to group participants and association statistic is calculated by stratifying the sample population by clusters. Computer software like STRUCTURE is used for allocation of individuals into different subpopulation and association test is performed via program as STRAT (Price et al., 2010). The number of clusters to be assigned to a population has always been a major limitation (Setakis et al., 2006). Unlike, computationally challenging structured association, PCA corrects for stratification using top principal components (eigen vectors) as the covariates (Price et al., 2010). In both approaches, the markers with strongly differentiated allele frequencies are corrected with a greater value. According to Price, Zaitlen et al. (2010), above mentioned approaches do not correct for other confounding factors such as family structure or cryptic relatedness.

An alternative method is the "genomic control" which corrects for both stratification and cryptic relatedness (Devlin and Roeder, 1999). As it is based on a Bayesian probability model, it does not apply the traditional methods of correction for multiple testing (e.g. Bonferroni correction). An association statistic Y^2 is adjusted at each position by uniform overall inflation factor, λ. "Y^2/λ" is used at individual marker instead of Y^2. Such correction might not be sufficient for markers that have greater difference in allele frequencies across the ancestral population whereas the correction might be unnecessary for markers that have very limited difference in allele frequency which leads to loss of power (Price et al., 2006).

Genomic control, Structured based association and PCA approach are based on the following general linear model,

$$Y = XB + e$$

Where, Y is the phenotype, X is the genotype at the candidate marker. Additional co-variates can also be added to X. B denotes the coefficient of fixed effects (markers and other co-variates) and 'e', the residual denotes the variation in Y unexplained by the fixed effects taken into consideration.

In a PCA approach, the principal components are included in X as the co-variates that take into account the fixed effects of genetic ancestry (Price et al., 2010).

Mixed Linear Model (MLM) approach can successfully model population structure, family structure and cryptic relatedness [(Price et al., 2010),(Yu et al., 2006)].

$$Y = Wv + XB + Zu + e$$

Y denotes a vector of phenotype, v and B denotes the unknown fixed effects of marker and non-marker effect respectively. 'u' denotes vector of unknown random polygenic effects and 'e' denotes a vector of residual effect. W, X and Z are the incidence matrices. Here, variance of the random polygenic effects, $Var(u) = \sigma^2 K$,

where, 'K' denotes a kinship matrix which is calculated on the basis of pedigree information or the genetic markers. To derive kinship matrix from the genetic markers, pairwise identity by state (IBS) is calculated among individuals. IBS is computationally tractable without prior knowledge of allele frequencies or pedigree information with less impact of population stratification (Roberson and Pevsner, 2009). IBS for a pair of individuals 'i' and 'j' is computed by:

$$f_{i,j} = \Sigma_k \frac{(x_{i,k} - p_k) * (x_{j,k} - p_k)}{(p_k * (1 - p_k))}$$

where, k=number of SNPs, $x_{i,k}$ is the genotype of ith individual at k^{th} SNP and $x_{j,k}$ is the genotype of jth individual at k^{th} SNP coded as 0, ½ and 1, p_k is the frequency of the "+" allele. 'e' is a vector of random residual effect which has zero mean and covariance $R = I\sigma_e^2$ where, I is the identity matrix and σ_e^2 is the unknown residual variance. The evaluation is done on each marker with the null hypothesis of v=0 and alternative hypothesis of v≠0. The maximum likelihood estimation is performed and the test of null hypothesis is done using F-statistic or chi-square test.

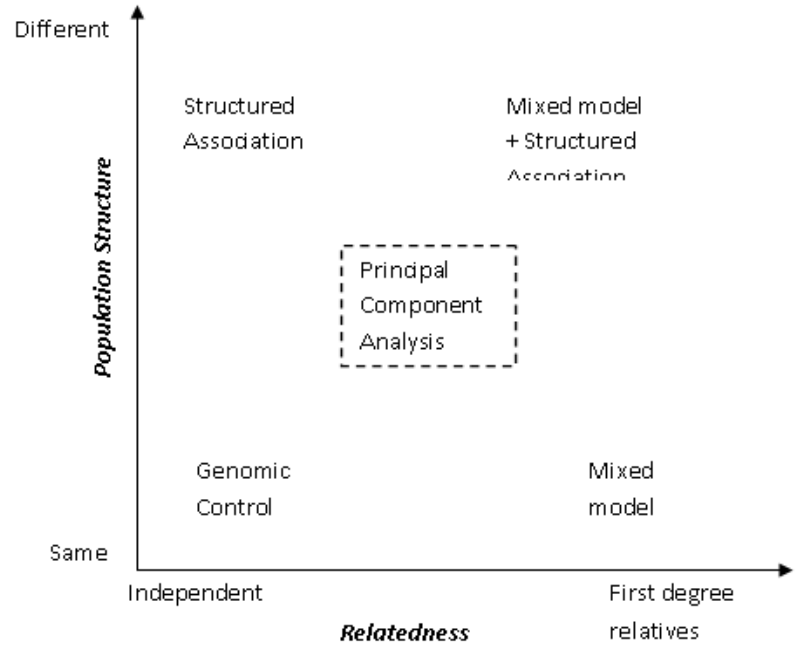


Figure 3. Application of different approaches during presence of different levels of population structure and familial relatedness in the study sample (Aulchenko, 2010). Combined use of mixed model and structured association is appropriate when the study population has both high level of relatedness and population structure. Mixed model approach is suitable when the study population is less effected by population structure but has high level of relatedness. In contrast, use of structured association is effective when the study population has less relatedness but high population structure. Genomic control is effective when there is presence of less relatedness and population structure.

Another problem in GWA study is multiple testing which imposes a risk of large number of the markers being associated with the disease phenotype falsely or by chance alone. In order to

overcome such scenarios various correction methods are applied but till date Bonferroni correction is the most commonly used method (Balding, 2006). The level of significance is divided by the number of multiple tests performed calculating a value. The p-value for a test should be less than the level of significance obtained after Bonferroni correction for any association to be taken as significant. Another procedure, permutation testing is also used to correct for multiple testing. Here, the case-control labels among the study samples are changed randomly for a specified times and the original tests is repeated for specified times. The empirical P-values thus obtained are compared with the original P-value. Although computationally intensive, this approach is considered as a "gold standard" for accurate correction (Clarke et al., 2011).

Various genetic analysis tools have been developed for association studies. Each tool comprises different approaches which can be used on the basis of presence of confounding factors in the study samples. PLINK, STRUCTURE, EIGENSTRAT, EMMAX are some of the tools set for GWAS (Clarke et al., 2011, Price et al., 2010).

'GenABEL' (package for genome-wide association analysis between quantitative or binary traits and SNPs) in R-program (www.r-project.org) open statistical analysis software has been used for GWA in the case control design study of CCD. Different methods of association analysis such as Basic association test, Genomic control, structured association, PCA and Mixed models are included in the GenABEL-package (Kierczak et al., 2011). For GWAS in case-control design of CCD, association analyses have been performed using all the models mentioned above.

MATERIALS AND METHODS

Materials

Sample collection:

The sample comprised of 40 dogs (Table 1). The information regarding age of onset of disease, gender and age of controls have been collected from Swedish Kennel Club- SKK (www.skk.se). Age at diagnosis in the 19 cases varies from ~3 to ~10 years. The age at examination for the 21 controls ranges from ~4 to ~11 years. 26 individuals belonged to Powder Puff variant and 14 others couldn't be identified. The majority of individuals (25) were Swedish, second largest group (10) are from Finland and remaining individuals are from Germany and USA. To ensure consistent phenotype classification, the majority of dogs were examined by the same specialist in Veterinary Opthalmology (prof Kristina Narfström). Blood samples were collected from both cases and controls into EDTA tubes and genomic DNA was extracted from peripheral blood leukocytes using a QIAsymphony SP/AS instrument (Qiagen, Hilden, Germany).

Table 1 The sample population divided on the basis of phenotypic status, gender and country of origin.

Country of origin	Cases			Controls			Total
	Male	Female	Total	Male	Female	Total	
Sweden	6	5	11	10	4	14	25
Finland	3	2	5	3	2	5	10
Germany	1	-	1	-	2	2	3
USA	-	2	2	-	-	-	2
Total	10	9	19	13	8	21	40

SNP Array:

The long Linkage disequilibrium (LD) regions in dogs at around 500 Kb to 1 Mb (Karlsson et al., 2007) gives an opportunity to use few SNPs for whole genome coverage for association analysis. CanineHD Beadchip from illumina was used for the study. The array includes SNPs derived from 2.5 million SNP set of Dog Genome Project (CanFam 2.0) and 1, 696 SNPs that have been identified using hybridization-based targeted re-sequencing method. CanFam 2.0 assembly is built on the basis of a whole genome shotgun (WGS) sequencing of the Boxer and partial sequence of a Poodle (REF) and 100 000 sequence reads from nine different breeds. More than 170 000 evenly spaced SNPs with an average of > 70 SNPs per megabase (Mb) are present on the array.

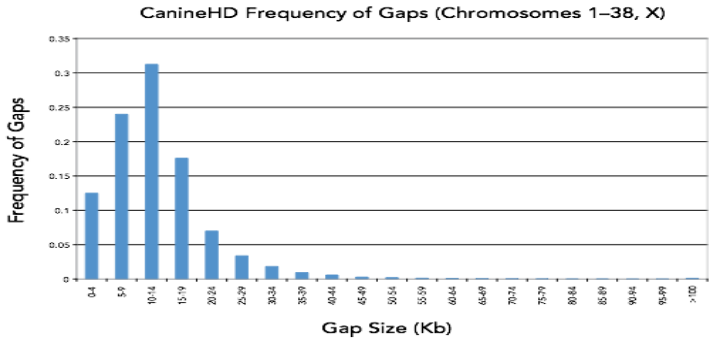


Figure 4 The frequency of different length of distance between SNP markers. The distance of 10-14 kb between markers is observed in high frequency in the array.

Pedigree

The records from SKK were used to build a pedigree of. The registration ID of each individual in the study sample was entered in SKK website and the relationships were obtained. The ancestors of the individuals were traced back for two to three generations. The age of onset for the cases and age at which the controls had last been examined for eye disease is noted in the pedigrees. Firstly, the relationship among the cases and controls were investigated through pedigree. Later the cases were further followed (traced forward to find the offsprings) to understand the mode of inheritance of the disease.

GWA Study:

40 CCD dogs were genotyped with the CanineHD Beadchip at GeneSeek (http://www.neogen.com/GeneSeek/). The genotyping results from GeneSeek were in Plink format (ped- and map files) and were then converted into a format required by GenABEL. A separate file "phenotype.dat" was created to provide phenotypic information: sex and disease status of studied individuals.

Quality control:

174357 markers were genotyped/ called for 40 individuals. The quality of data plays important role and have a significant impact on association analysis hence, should be edited correctly. Therefore, in a preliminary quality control procedure both markers and individuals were analysed on the basis of genotyping rate (< 95% removed) In the next step, individuals were analysed on the basis of autosomal heterozygosity (False Discovery rate < 1%) and value of IBS (>= 0.95 removed) between individuals. The markers were further analysed for minor allele frequency (MAF) where markers with a threshold of $< 1\times10^{-08}$ were considered as non-informative and were removed. P-value for Hardy-Weinberg Equilibrium (HWE) is the probability that the deviation of the marker from HWE is explained by the chance (Purcell et al., 2007). The markers deviating from HWE with p-value of $< 1\times10^{-08}$ were removed carefully as the selection, inbreeding, population stratification and also disease association could have lead to the deviation (Balding, 2006). 42564 markers were removed on the basis of call rate and MAF threshold. 480 markers were removed on the basis of HWE criterion. 130308 remaining markers and all 40 individuals passed the criteria.

Classical Multi-dimensional Scaling Plot (MDS):

To build a kinship matrix based on the genetic markers, only the autosomal markers were used to calculate IBS between the individuals. A 40*40 genomic kinship matrix can also be used as a quality control step where very closely related individuals could be removed based on the value of IBS. A distance matrix is built from the genomic kinship matrix that was used to make a classical Multi-Dimensional Scaling (MDS) plot. A MDS plot is constructed to map all pair wise distances into k-dimensional space such that the distance between the individuals in the plot is closest to the real distance in the distance matrix. The primary purpose of MDS plot is to visualize population substructure (Purcell et al., 2007). However, k-dimensional representation of the substructure obtained from MDS can also be used as a covariate in association analysis. Also, the classical MDS obtained using Euclidean distance metric is numerically identical to PCA.

5 out of 19 cases considered as outliers on the basis of MDS plot were removed and further steps were performed with 14 cases and 21 controls (35 individuals).

Final quality control:

The individuals and markers were further analysed on basis of genotyping rate (< 95% removed) and markers with a minor allele frequency (MAF) of less than 5% were removed. HWE was applied only to the controls. FDR of 0.2 is set due to negative value in HWE. 19 596 markers were removed on the basis of genotyping rate and MAF threshold. 358 markers were removed on the basis of HWE threshold. 110 409 markers passed the final quality controls out of which 107 862 autosomal markers were used for MDS plot.

Clustering (K-means):

The individuals were clustered through K-means clustering method (Kierczak et al., 2011). 10 different clusters were built. To determine the optimal number of subpopulations that the total population could be divided in, within group sum of squares (WSS) was plotted against the number of clusters. Guided by the plot (WSS vs K), the population was divided into 3, 4 and 5 clusters and association analyses were performed taking these cluster into account. The clustering of individuals was performed through the nearest mean clustering method. In the classical MDS plot, the distance is calculated using Euclidean distance method.

Association Analysis:

Fisher's exact test was performed in GenABEL to analyse if there is significant association of "sex" variable with disease phenotype. P-value of 0.175 was observed for 35 individuals, suggesting a non significant association between the variable "sex" and "disease" phenotypes. Hence, this variable was not included in the association analysis as a covariate.

GWA study was carried out using six different models:

- Basic association test
- Genomic control
- Structured association: grouping individuals into 3, 4 and 5 different clusters.
- Principal component analysis
- Mixed model approach, and
- A combined approach of mixed model and structured association: 3, 4 and 5 different clusters.

The level of significance for multiple testing was obtained with Bonferroni correction by dividing a p-value=0.05 with the total number of multiple tests performed. The significant value for a test after Bonferroni correction was $4.52x10^{-07}$.
Regardless of model used, the strongest association of the disease phenotype was found to SNPs on a region of chromosome 8. Two SNPs "BICF2P713861" and "TIGRP2P110467" were located at position 26769490 and 27113849 on chromosome 8, respectively. This held true despite that fact that different values of inflation factor λ was obtained from different models.

Visualization of Linkage Disequilibrium (LD) on chromosome 8:
Haploview, an open source program written in Java (Barrett, 2009), helps to visualise the haplotype block built on the basis of pair wise LD between the markers. This plot helps to identify if associated SNPs are present in the haplotype block such that association analysis taking haplotype into account can also be performed.

The pair wise LD among the markers on chromosome 8 was calculated from Haploview (version 4.0). Linkage disequilibrium denotes a non-random association between alleles such that those alleles appear more often than by chance. The files were loaded in linkage format: "ped" file and "info" file. The markers on chromosome 8 were filtered according to the threshold of p-value=1e-06 for HWE, <75% genotyping rate and MAF=$1x10^{-03}$. Pairwise LD was measured as "r^2" for the markers within a distance of 500kb."r^2" is based on the correlation of marker alleles. The value extends from 0 to 1 ($0 < r^2 < 1$).

The genotypes of the individuals for the markers from position "26675235" to "27139726" was analysed to search for the haplotype block segregating differently in cases and controls which helps to narrow down the region to look for the casual variant (Downs et al., 2011).

Odds Ratio:
Allelic odds ratio are generally used to analyse the association of the marker with disease phenotype in association studies with SNPs (Sato et al., 2006). With this concept, the odds ratio for the potential risk allele was calculated using VassarStats (http://faculty.vassar.edu/lowry/VassarStats.html). Yates values were used and were corrected for continuity.

Sequencing of coding region:
The positions of the markers with strongest association were investigated in the Ensembl genome database (http://www.ensembl.org/) derived from CanFam 2.0 assembly that linked to a gene MDGA2/ MAMDC1 (MAM-domain containing glycosylphosphatidylinositol anchor 2) of 317.42 kb. All 15 exons of MDGA2 gene were sequenced. This procedure was performed to search for any non-synonymous mutations in MDGA2 gene between cases and controls in the study sample.

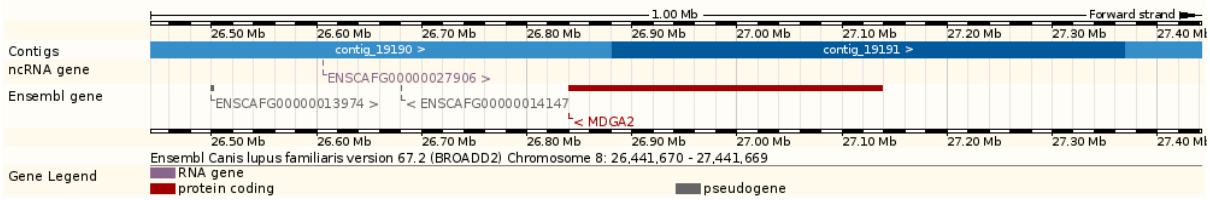


Figure 5 The image exported from Ensembl (CanFam 2.0) gives an overview of the location of gene MDGA2 on chromosome 8.

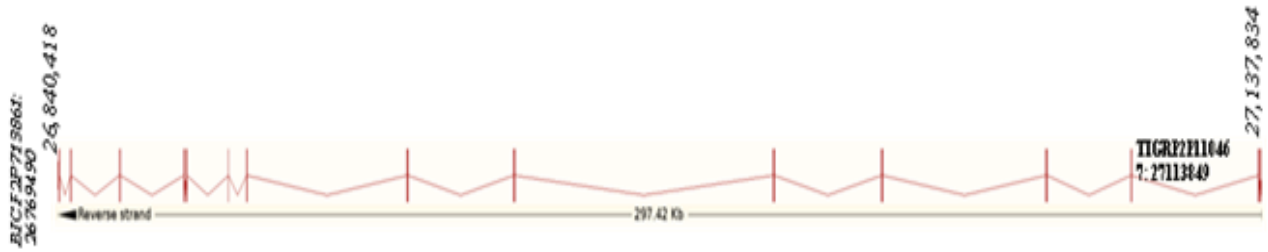


Figure 6 The transcript of MDGA2, that contains 15 exons. The position "26,840,418" and "27,137,834" are the last and first position in base pair, respectively of the transcript that is 297.42 kb long. Out of two strongly associated SNPs, "TIGRP2P110467" is located on Intron 2-3 and another SNP "BICF2P713861" is located outside the transcript ~70 kb downstream.

The primers were designed through PrimerZ (http://genepipe.ncgm.sinica.edu.tw/primerz), a web application that designs exon-primers (Tsai et al., 2007) based on the annotation of a specific gene in Ensembl gene ID (ENSCAFG00000014164). PrimerZ generated 20 different primer pairs that covered all exons 14 exons of MDGA2 and a xx bp portion of the the 3'-UTR and 5'-UTR regions. All primers were designed with M13 forward (-21) or reverse (-29) tails included for direct sequencing. The primers pairs are listed under Primer section of Appendix.

Polymerase Chain Reaction (PCR) and Sequencing:

Big Dye Direct Cycle Sequencing Kit (www.appliedbiosystems.com) was used for PCR amplification and sequencing of PCR product.

5 µl of PCR amplification master mix, 1.5 µl of ~0.8 µM of M13 tailed primer pairs (forward and reverse), 1 µl of ~4ng/ µl concentration of genomic DNA and 2.5 µl of deionised water were mixed for PCR amplification. This final volume of 10 µl was amplified with PCR program: 95^{o}C for 15 min followed by 25 cycles of initial denaturation of 95^{o}C for 1 min, annealing of 58^{o}C for 1 min and elongation of 72^{o}C for 1 min, followed by final elongation temperature of 72^{o}C for 7 min.

Gel electrophoresis was performed by loading ~1 µl of PCR product and ~2 µl of Loading Buffer into 2% of agarose gel. The size of the bands was confirmed by use of a ladder of 50 – 2,000 bp .

Sequencing of PCR products was performed by adding 1 µl of Big Dye Direct M13 forward or reverse primer and 2 µl of Big Dye Direct sequencing master mix into each well. Sequencing program was performed with GeneAmp^{TM} PCR System 9700 thermal cycler (Applied Biosystems) and included three holding steps: 37^{o}C for 15 min, 80^{o}C for 2min, 96 ^{o}C for 1min followed by 25 cycles of initial denaturation temperature 96^{o}C for 10sec, annealing temperature 50^{o}C for 5sec and elongation temperature 60^{o}C for 4 min. The sequenced products were held at 4^{o}C until purification that was done using Big-dye X Terminator^{TM} purification kit. A mixture of 45 µl of SAM solution and 10 µl of BAM solution was added to each sequenced product (well). The mixture is vortexed using IKA MS3 Digital Vortexer at speed of 2000 rpm for ~20 mins. The solutions was centrifuged using a swinging bucket centrifuge for 2 mins at 1000*g and analysed using capillary electrophoresis.

Capillary electrophoresis was performed on 3500XL Genetic Analyzer (Applied Biosystems) with Dye Set Z, POP-7^{TM} polymer, 50 cm array and Big dye direct seq_3.1_pop_7X_term as run module. Initially, only one case and one control were sequenced using all 20 primer pairs. The difference was observed in some areas. Hence, other four samples from control and four samples from cases were sequenced for region of promoter, exon 6 and 15.

Sequence Analysis:

The sequence alignment software Codon Code Aligner (version 4.0.2, CodonCode Corporation) was used to analyse the chromatogram files (sequences) of exons to find for any changes in the nucleotides between cases and controls. The sequence from Ensembl was taken as the consensus /reference sequence based on which both forward and reversed sequences of cases and controls were analysed. The exons of the case and control were translated into amino acid sequence to identify if the changes at some positions lead to non-synonymous mutation. The translation was performed using an online tool: ExPasy (http://web.expasy.org/translate/). The translated sequences were then aligned using online tool: CLUSTALW (http://www.ebi.ac.uk/Tools/msa/clustalw2/).

RESULTS AND DISCUSSION

The genome wide association study of a novel retinal disease that has been called pigmentary chorioretinopathy in Chinese crested dogs was mapped to a region on CFA 8. The two SNP's with the strongest association were found in or in close proximity to the *MDGA2* gene. This implies a possibility of identifying a causal mutation in the *MDGA2* gene. All the results along with discussion are described below:

Mode of inheritance:

Initially, the pedigree was constructed to take account the relationship among the study samples that might lead to the population substructure which acts as one of the confounding factor in association studies. The Pedigrees of four families (Appendix section) were constructed by tracing back each individual to two or three generations toward the ancestor and also tracing forward the same individuals to its offspring. If individuals are traced further back then there is a chance of these families being related.

The pedigree analysis of the Chinese crested case-control population implies autosomal recessive mode of inheritance as some of the key features observed in the pedigree are: presence of equal number of males and females affected in the population, the phenotype is skipping a generation for example if an individual "B1" in pedigree "B" is focussed, the disease is appearing in third generation where individual "B7" has a disease phenotype where as it's parents are healthy and no disease is observed in that litter. Some matings that produce diseased offspring have at least one parent affected as can be observed in pedigree "A", "A2" individual mated with a normal individual give many offspring with disease phenotype. Similarly, in pedigree B, B3 individual mated to normal give offspring B4 and B5 with disease phenotype.
In pedigree D, the disease is occurring in each generation. Here, the D1 individual mated to 5 yrs old normal male give offspring D2 with disease phenotype. D2 when mated with a normal male (1 year at age of examination) gives a diseased female diagnosed around 3 years of age. One half of each sex among offspring is affected. This family gives a picture as if disease is dominant and rare. However, many offspring with unknown phenotype are present in the pedigree. If they are normal then the chance of dominant mode of inheritance will decrease.

The affected males mated with normal females transmit the disease to all the females but not males. This criterion for X-linked dominance is not fulfilled in the pedigree hence this mode of inheritance can be excluded. A2 male mated with normal female has female offspring in 2006 that is not affected by disease. Similarly, C2 in pedigree C mated to normal females give daughters with normal phenotype.

The scenario of two affected parents is not observed in these pedigrees. In pedigree C, C1 female mated with a normal male give a normal male offspring. Similarly, in pedigree B, B1 affected female mated to an individual with unknown status give healthy male offspring in 2001 and 2002. This also helps to rule out the fact that the disease is X-linked recessive. Further, males and females are affected in almost equal numbers.

From these conditions, the disease is segregating in a population at some point skipping generations. Hence, it can be concluded that the disease is genetic. Available pedigree information suggests that the disease is inherited according to a simple Mendelian autosomal recessive mode of inheritance.

Occurrences of eye disease in dogs have also been associated with mutations in mitochondrial DNA. There are some evidences that mutations in the mitochondrial genome would lead to functional defect in eyes. Leber's hereditary optic neuropathy (LHON), retinitis pigmentosum (NARP-neuropathy, ataxia and retinitis pigmentosa) are some of the diseases that occur due to defect in function of the mitochondrial genome (Appleyard et al., 2006). In case of NARP there is a point mutation in MTATP6 gene that encodes for the subunit6 of F1FO-ATPase (Claude, 2012). However, the occurrence of such mitochondrial disease is transmitted from affected females to all the offspring. However, such scenario is not fulfilled in our pedigree and consequently this non Mendelian mode of inheritance could be ruled out.

Many of the individual's veterinary data are missing and hence their disease status is unknown. Only the diseased individuals have been traced to their offspring. The scenario might be different if we have more individuals with known disease status and also track down those healthy individuals in a litter that contains a diseased individual. This step might give a better picture and capture a wider population compared to the current situation. Also, to repeat the eye check of the normal individuals may also help, as many individuals are young when examined and the age of onset of disease varies a lot, starting early from ~3 years to ~10 years.

The mode of inheritance generally helps to interpret the family history, potential risk of being susceptible to this disease and to make better use of the available genetic tests. The results from pedigree analysis can in this case be used to further develop the result obtained from GWA study. Such analysis provides an idea of what can be expected from the study being performed. The individual with diseased parents (recessive inheritance mode) is expected to have positive disease status. The result of GWAS for that individual is expected to have genotype of case. However, lots of potential problems such as incomplete penetrance of disease, new mutations, pedigree errors, variable age of onset (as seen in our case), misclassifications etc. might reduce this benefit. But there can be situations, where pedigree information might help us to confirm occurrence of penetrance phenomenon.

Genome Wide Association Study (GWAS):

Multi Dimensional Scaling (MDS) plot:

GWAS of 40 Chinese crested dogs (19 cases and 21 controls) was initially performed. To investigate the degree of stratification and relatedness among the sampled population an MDS plot was constructed using all 40 individuals. The classical multidimensional scaling, also referred to as Principal Coordinates Analysis, is primarily used to improve the visualisation of the relatedness among the individuals based on the genomic data. The relative distance among individuals in a distance matrix of 40 dimensions (axes) is converted to the points in certain desired number of scaling dimensions for simple interpretation. The Euclidean distance (straight line) between the scaling dimensions can give rise to the original distance matrix. The distance between the points representing individuals in the MDS plot is a straight line distance. MDS plots showing the genetic distance among the individuals in two dimensions can be observed in the

following figure 5a (left plot). In figures 5a, the cluster of four cases on the bottom right corner and three cases on the top left corner can be observed.

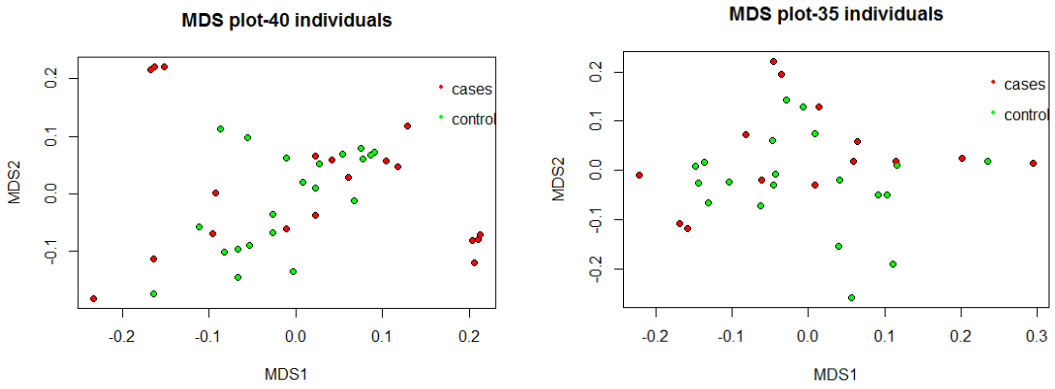


Figure 7a and 7b. MDS plot of 40 and 35 individuals (removing 5 cases) where each individual is represented by a circle, cases are coloured in red and controls in green.

The individuals in the lower right corner of the plot were identified to be the full siblings. Similarly, two individuals out of three on the top leftmost corner of the plot were identified to be father and daughter. The relationship of third individual from Finland with other two couldn't be identified due to missing veterinary data. In figures 5b Individuals seemed to be spread out without formation of any clusters. This indicates implying no indication of stratification. However, MDS plot can help us to observe any stratification to some extent only.

Optimal number of cluster:

Within group sum of squares (WSS) plotted against the number of clusters was used to group the study population into an adequate number of clusters. The location of a sharp bend in the plot can be used to estimate the optimal number of clusters (Kierczak et al., 2011). Through careful observation, a sharp bend can be seen on 4 no. of clusters in figure 6.

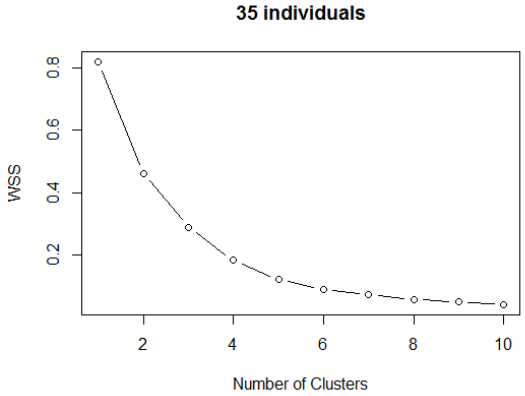


Figure 8 Within cluster sum of squares obtained from each cluster is plotted against the number of clusters.

An optimum number of clusters to be considered is still not clear. In order to ensure optimum number of clusters to be used in association analysis through structured association approach the study sample was clustered into three different numbers of clusters: 3, 4 and 5.

Clustering:

Figure 9 below shows the individuals in the study samples clustered into different groups as we increase the number of clusters. In the plot, cluster 3, individuals divided into three clusters are represented by different colours as red, blue and green. These individuals are plotted on the basis of the pair wise identity by state (IBS) values. As, the cluster number is increased to 4, the individuals on the bottom part of the plot form a different cluster of green colour (figure: Cluster =4), inferring that they differ genetically from the rest of the members in the same cluster. Further, in figure cluster = 5, three individuals on the right of the plot forms a different cluster represented by black colour in the plot. This suggest that compared to three individuals who grouped into a different cluster (green cluster in figure: cluster =4) these three individuals in black clusters are genetically similar. This implies presence of stratification or relatedness in the study sample. However, the members in other two clusters represented as blue and red in figure: cluster =3, doesn't divide further into any groups.

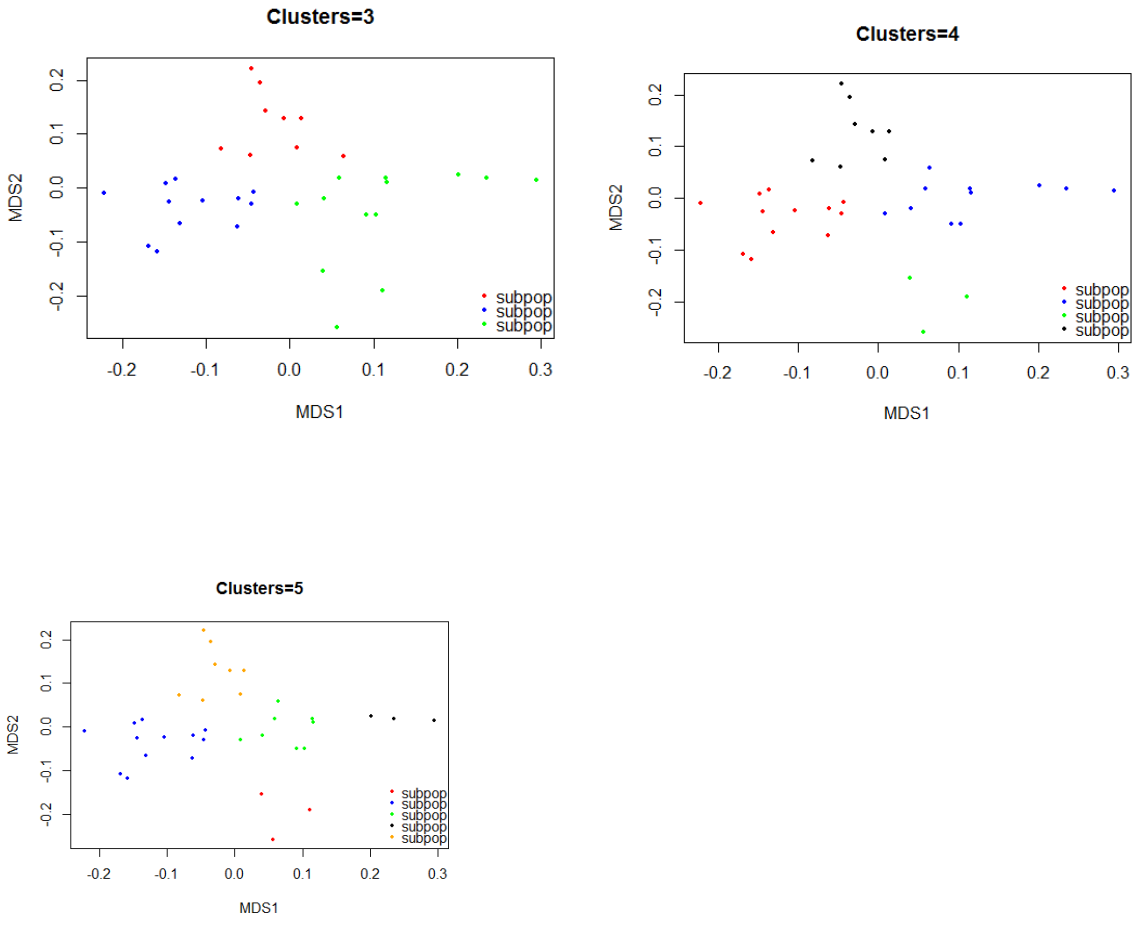


Figure 9 The population of 35 individuals divided into three, four and five subpopulations based on IBS-values. Each subpopulation is denoted by a specific colour in all three plots.

Association Analysis:

Association analyses were performed with different models to ascertain that the associated SNP that we observe is not due to spurious association. The effect of variable "sex" was not significant to the disease phenotype status hence for approaches based on linear and mixed model, the fixed effects were only the genotype of individuals. The information of country of origin is also not taken into account because the practise of providing a registration number (identity number) to a dog is not very specific. In our sample, some Swedish dogs have parents from Finland and as the dog was brought here, Swedish registration number is provided. Due to such scenario, the information of country of origin may further mislead.

Table 2 The table describes all the methods used for association analyses. The top three SNPs along with the chromosome number, position (base pair) of SNP in the chromosome, chi-square test statistic value, score provided to each SNP (P1df), corrected score (Pc1df) and Genomic inflation factor from each approach are given in the table.

Methods:	SNP name	Chr.	Position	chi-square	P1df	Pc1df	Inflation
Basic association	**BICF2P713861**	8	26769490	21	4.59E-06	2.29E-05	**1.170997**
	TIGRP2P110467	8	27113849	19.53955	9.85E-06	4.41E-05	
	BICF2S232422	5	32983613	18.95833	1.34E-05	5.73E-05	
Genomic control	**TIGRP2P110467**	8	27113849	22.75492	1.84E-06	6.19E-06	**1.113844**
	BICF2P713861	8	26769490	22.28774	2.35E-06	7.70E-06	
	BICF2S232422	5	32983613	19.8935	8.19E-06	2.38E-05	
Structured association	**BICF2P713861**	8	26769490	24.12909	9.01E-07	5.61E-06	**1.170345**
3 clusters:	**TIGRP2P110467**	8	27113849	24.11187	9.09E-07	5.65E-06	
	BICF2P1044496	8	27972553	20.31726	6.56E-06	3.09E-05	
4 clusters:	**BICF2P713861**	8	26769490	22.65861	1.93E-06	8.79E-06	**1.146857**
	TIGRP2P110467	8	27113849	22.63706	1.96E-06	8.88E-06	
	TIGRP2P110728	8	27933541	19.42601	1.05E-05	3.86E-05	
	TIGRP2P110467	8	27113849	23.1011	1.54E-06	1.11E-05	**1.196298**
5 clusters:	**BICF2P713861**	8	26769490	22.93346	1.68E-06	1.20E-05	
	TIGRP2P110728	8	27933541	19.26553	1.14E-05	5.99E-05	
Principal component analysis:	**TIGRP2P110467**	8	27113849	25.77222	3.84E-07	1.26E-05	**1.351601**
	BICF2P713861	8	26769490	25.50727	4.41E-07	1.40E-05	
	BICF2S232422	5	32983613	22.69999	1.89E-06	4.16E-05	

Mixed Model:	**TIGRP2P110467**	8	27113849	21.50683	3.53E-06	3.51E-06	**0.9996414**
	BICF2P713861	8	26769490	21.39995	3.73E-06	3.71E-06	
	BICF2S232422	5	32983613	19.05455	1.27E-05	1.27E-05	
Mixed model and Structure association:	**SNP name**	**Chr.**	**Position**	**chi-square**	**P1df**	**Pc1df**	**Inflation**
3 clusters:	**BICF2P713861**	8	26769490	22.24331	2.40E-06	3.09E-06	**1.022345**
	TIGRP2P110467	8	27113849	21.90407	2.87E-06	3.68E-06	
	BICF2S232422	5	32983613	18.15447	2.04E-05	2.51E-05	
4 clusters:	**BICF2P713861**	8	26769490	21.44949	3.63E-06	4.59E-06	**1.021269**
	TIGRP2P110467	8	27113849	20.97703	4.65E-06	5.84E-06	
	BICF2P752103	8	26675235	18.34723	1.84E-05	2.25E-05	
5 clusters:	**BICF2P713861**	8	26769490	20.7553	5.22E-06	7.00E-06	**1.027856**
	TIGRP2P110467	8	27113849	20.4185	6.22E-06	8.31E-06	
	BICF2P752103	8	26675235	19.14179	1.21E-05	1.59E-05	

Different models have their own efficiency depending on the level of substructure in a population and relatedness among individuals (Aulchenko, 2010). Use of various models might also give an idea of the level of population stratification present and also the model that would better fit by correcting for the stratification. Table 2 gives an overview of the results obtained from all the models. The least value of λ 0.99 is observed with mixed model approach on 35 CCD and the maximum value of λ 1.35 is observed for PCA method. PCA is efficient method to be applied when the effect of substructure and relatedness among individuals is at medium level. But, as the value of inflation factor, λ >1, the study sample is affected of substructure or relatedness. The value of λ 1.14 is obtained when structured association method is applied for 35 individuals whereas the value of λ is 0.99 for mixed model. The value of λ > 1 in structured association gives a slight hint of high level of relatedness among individuals in the study sample which is taken into account by the mixed model approach. Similarly, value of λ is 1.02 when a combined approach of mixed model and structured association is applied on the population of 35 individuals by grouping population into four clusters. Both methods mentioned above seem to correct better for population structure and relatedness. Zhang et al. (2010) mentions that the statistical approach which leads to a distribution of observed negative logarithmic p-values close to the expected diagonal line corrects better for the false positive association. This concept could be applied to evaluate QQ plot from both above mentioned methods and to choose one method for further analysis.

Quantile Quantile (QQ) plot:

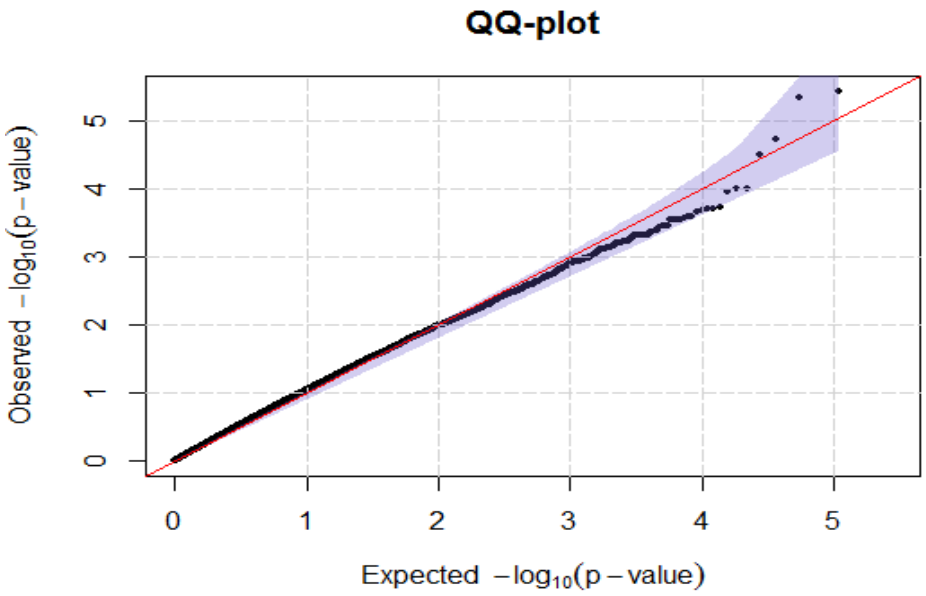


Figure 10a QQ-plot for a combined approach (mixed model and structured association)

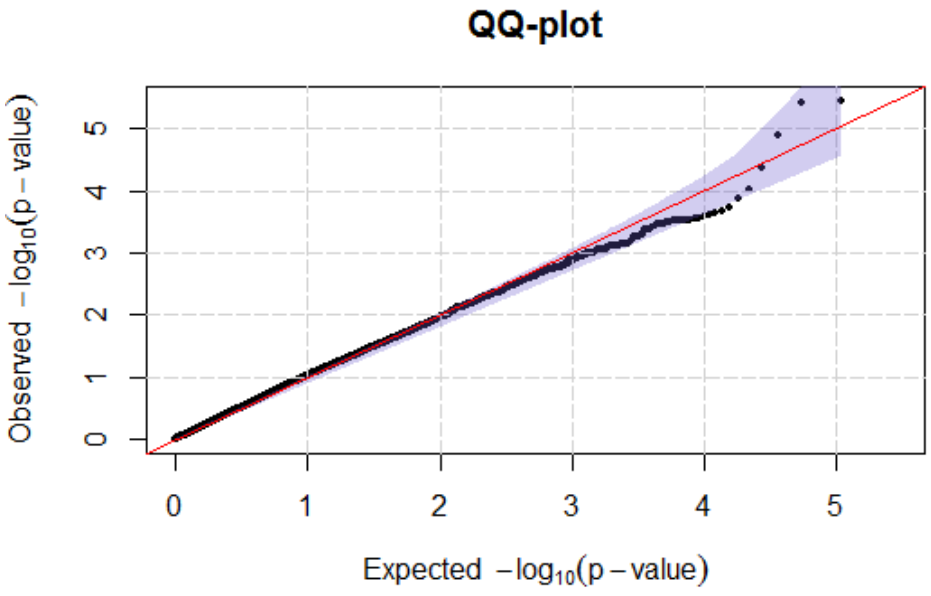


Figure 10b QQ-plot for mixed model approach

Figure 10a and 10b.Quantile-Quantile (QQ) plot of the distribution of negative logarithmic corrected p-values (black solid line) for the SNPs obtained from association analysis against the distribution of expected value (red diagonal line) under null hypothesis of no association. The grayish blue shade refers to the area bounded by confidence interval. The black solid line follows

diagonal red line to some extent and further along the tail, the values are deviating downward from the diagonal and only some markers deviate upward from the diagonal.

The quantile quantile (QQ) plot is a standard tool to diagnose the population structure in GWAS (Pearson and Manolio, 2008) . The distribution of test statistics (p-values or chi-square) obtained from the association tests are compared with the distribution of expected test statistics under null hypothesis of no association. According to Pearson and Manolio (2008), the strong deviation from the null hypothesis of no association may occur due to presence of highly associated and heavily genotyped locus. It may also occur due to presence of high level of confounding factors (stratification, relatedness). Also, McCarthy et al. (2008) mentions that the deviation of p-values across the entire distribution is a result of population substructure or relatedness and the deviation at the significant end of the range is the one brought by the large-effect susceptibility loci.

Here the QQ plot, figure 10a, of combined approach and of mixed model (figure 10b) appears to be very similar, but when observed carefully on negative logarithmic of p-value "4", QQ plot of a combine approach seems to have values much closer to the diagonal line compared to mixed model approach. Hence a combined approach is selected and further analyses are performed.

The scenario of the black solid line overlapping the red diagonal to some range fulfils pre-assumption of GWAS that is many SNPs are not associated with disease phenotype. However, deviation of solid line downward the diagonal in both plots might be due to the effects of population structure and relatedness. The two points on top highly deviating from the diagonal suggest the association to disease phenotype. The points are still below the upper boundary of confidence interval which suggests that the association is not significant.

Manhattan plot:

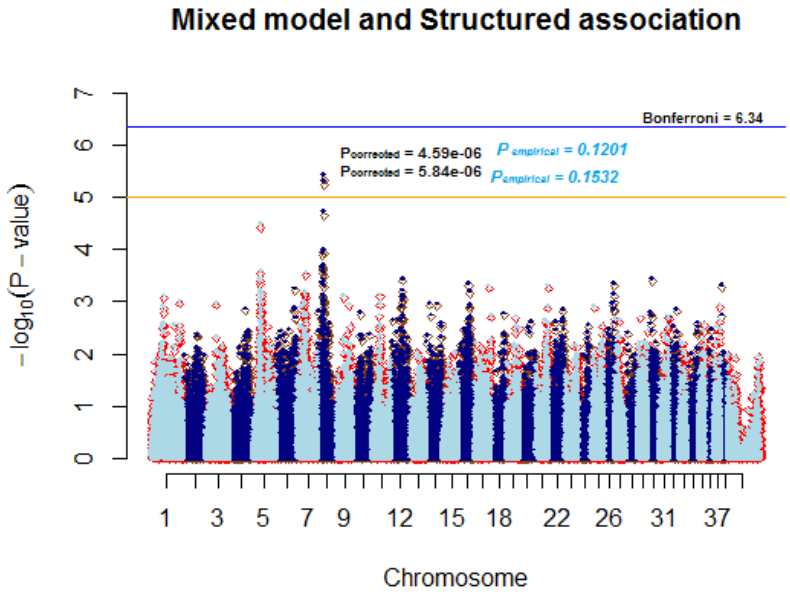


Figure 11Manhattan Plot for the association analysis using population of 35 individuals through a combined approach of Mixed model and Structured association taking into account four different clusters of the population

In the figure 9, the negative logarithm of bonferroni corrected p-value (4.52e-07) =6.34 is represented by horizontal blue line. The horizontal orange line below is drawn for ease of visualization of top two SNPs on chromosome 8. The blue dots indicate raw p-value for the SNPs whereas the dots with brown and red border lines represent the corrected p-values for the same SNPs and are observed below the bonferroni corrected value. Also, the $P_{empirical}$ values for the top two SNPs on chromosome 8 obtained after correction for multiple testing using Permutation method (10,000) times in GenABEL, can be viewed in blue colour (italic) next to the corrected values in the plot. The empirical p-value for the first SNP is 0.1201 > significance level (0.05). This infers that the association is not significant. Balding (2006) mentions that Bonferroni correction is regarded as overly conservative because the test of associations of markers with disease phenotype are considered as independent tests whereas in reality, due to linkage disequilibrium among markers different number of tests cannot be regarded as completely independent. This dependency of information is not taken into account in a testing procedure (p-value based) of GWAS which results into the loss of efficiency (Wei et al., 2009). For further analysis, approaches such as Hidden Markov Model can be applied that takes into consideration the dependency information and improves the accuracy of multiple testing. Moreover, an addition of individuals for the association study will also lead to increment of the associated score.

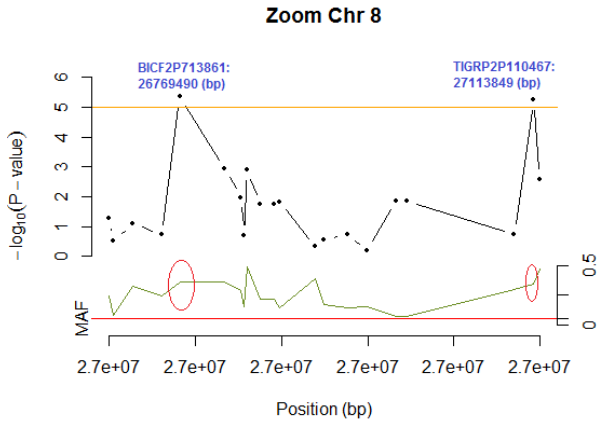


Figure 12 Manhattan Plot of negative logarithmic p-values of the markers on zoomed region of chromosome 8 along with the minor allele frequency (MAF) of the markers on the lower part. The red line in the MAF region denotes 5% threshold of MAF for the markers.

In the plot, the orange line is drawn in order to visualise the two SNPs easily. This plot suggests that the SNPs are not in the region of selective sweep. The region in a genome that displays reduced heterozygosity extended to certain region creating a long haplotype block fixed in a population describes selective sweep (Vaysse et al., 2011). It appears as a consequence of the selection of breed and genetic drift. In our scenario of case and control, the region of selective sweep denotes the region that is fixed for the whole population, hence no association will be seen in that region. The occurrence of the marker is highly unlikely in the region of selective sweep. As can be observed, both strongly associated markers are not present in the region of the selective sweep.

Pairwise LD plot:

Table 3 The top 10 SNPs in CFA 8 obtained on the basis of corrected p-value (Pc1df) are described in the table along with the position in chromosome. The markers are sorted on the basis of position left to right (increase in location number). The distance row gives distance between adjacent markers.

SNP	BICF2S23649439	BICF2P1391407	TIGRP2P109371	BICF2P299933	BICF2P638018	BICF2P1065138	BICF2P752103	BICF2P713861	TIGRP2P110467	BICF2P1044496
Chr.	8	8	8	8	8	8	8	8	8	8
Position	21298642	22082613	23439519	23613269	23630740	23650115	26675235	26769490	27113849	27972553
Distance:	-	783971	1356906	173750	17471	19375	3025120	94255	344359	858704
Pc1df	2.36E--04	1.32E-04	1.15E-04	2.18E-04	2.29E-04	2.29E-04	2.25E-05	4.59E-06	5.84E-06	1.19E-04

Figure 13 gives an image of how top 10 SNPs (mentioned in table 3) are located at CFA 8. Out of 10 markers, 3 form one block which also contain one of the two strongly associated markers "TIGRP2P110467". The total distance between the first and last SNP in the image is ~74, 000 bp. LD blocks can be observed as a black triangle that appears in the middle, but does not extend for too long due to recombination events that suggests to narrow down the region to the two strongly associated loci (from right to left, pointed by red arrows). The zoomed region for the two loci can be viewed in figure 14.

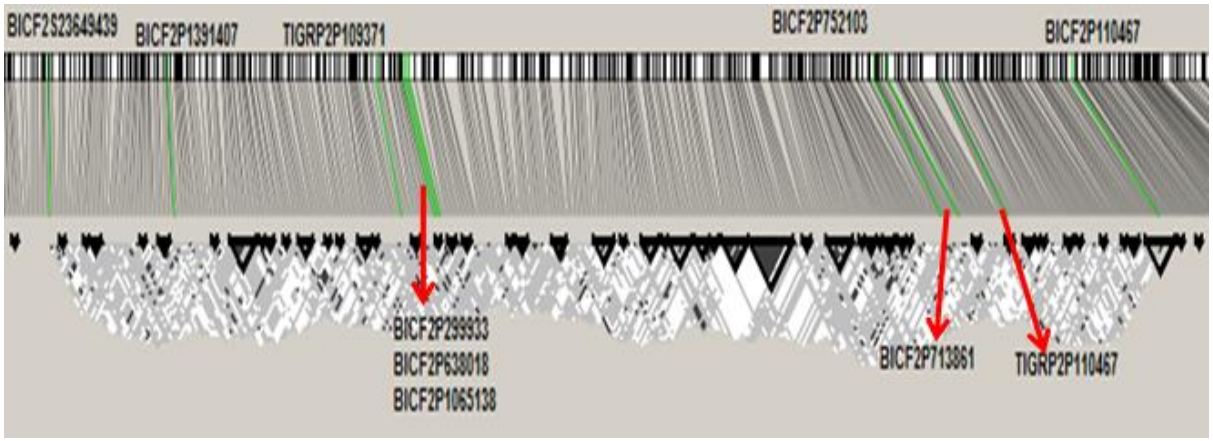


Figure 13 Zoomed image of pair wise LD plot on the region of chromosome 8 (from ~21298642 *to ~27972553)obtained from Haploview 4.2. The green line denotes the location of top 10 markers on chromosome 8 based on corrected p-values (scores) obtained from a combined approach of the mixed model and structured association. The designation of SNPs in the respective position next to the green line can be observed. For convenience the red arrows are used to point to the locations of some SNPs. Top two strongly associated are from the right pointed by red arrow below.*

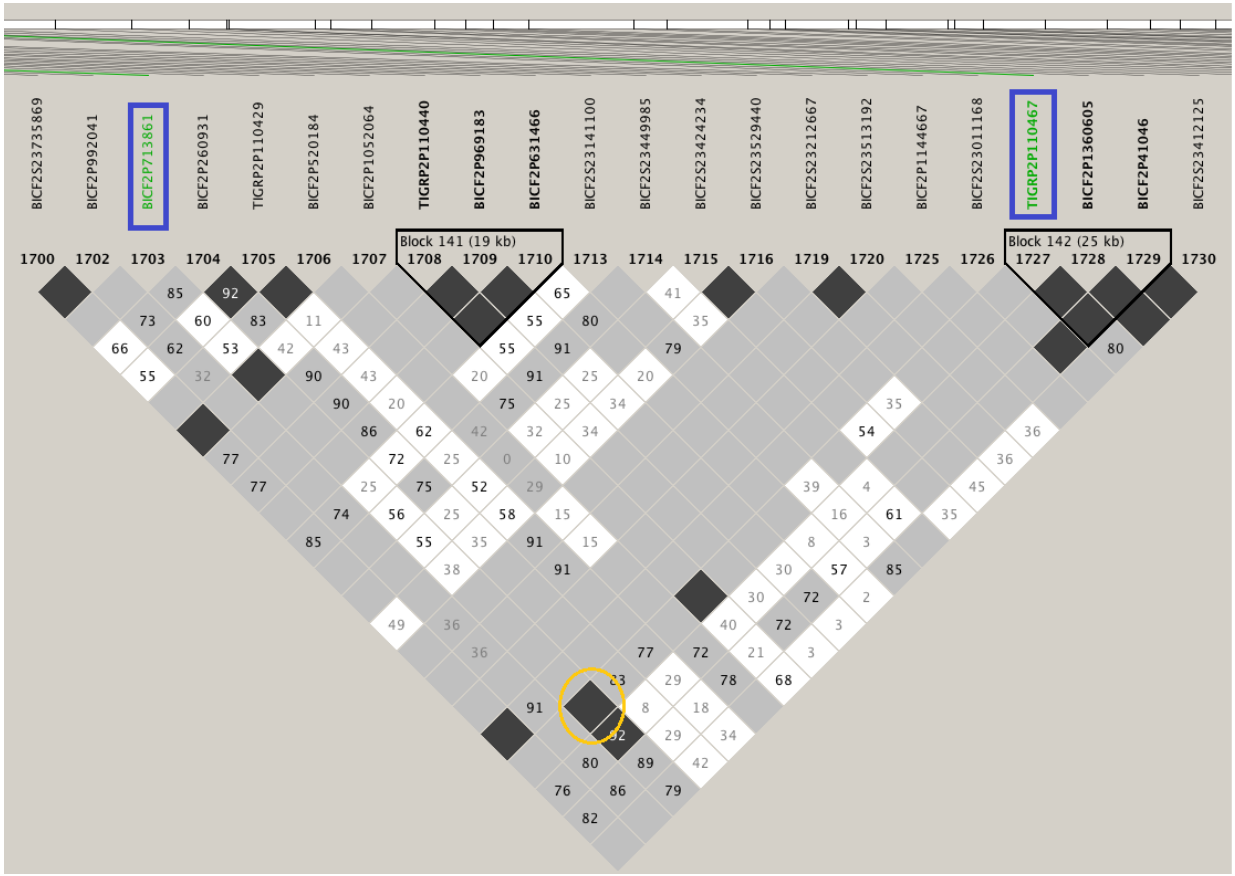


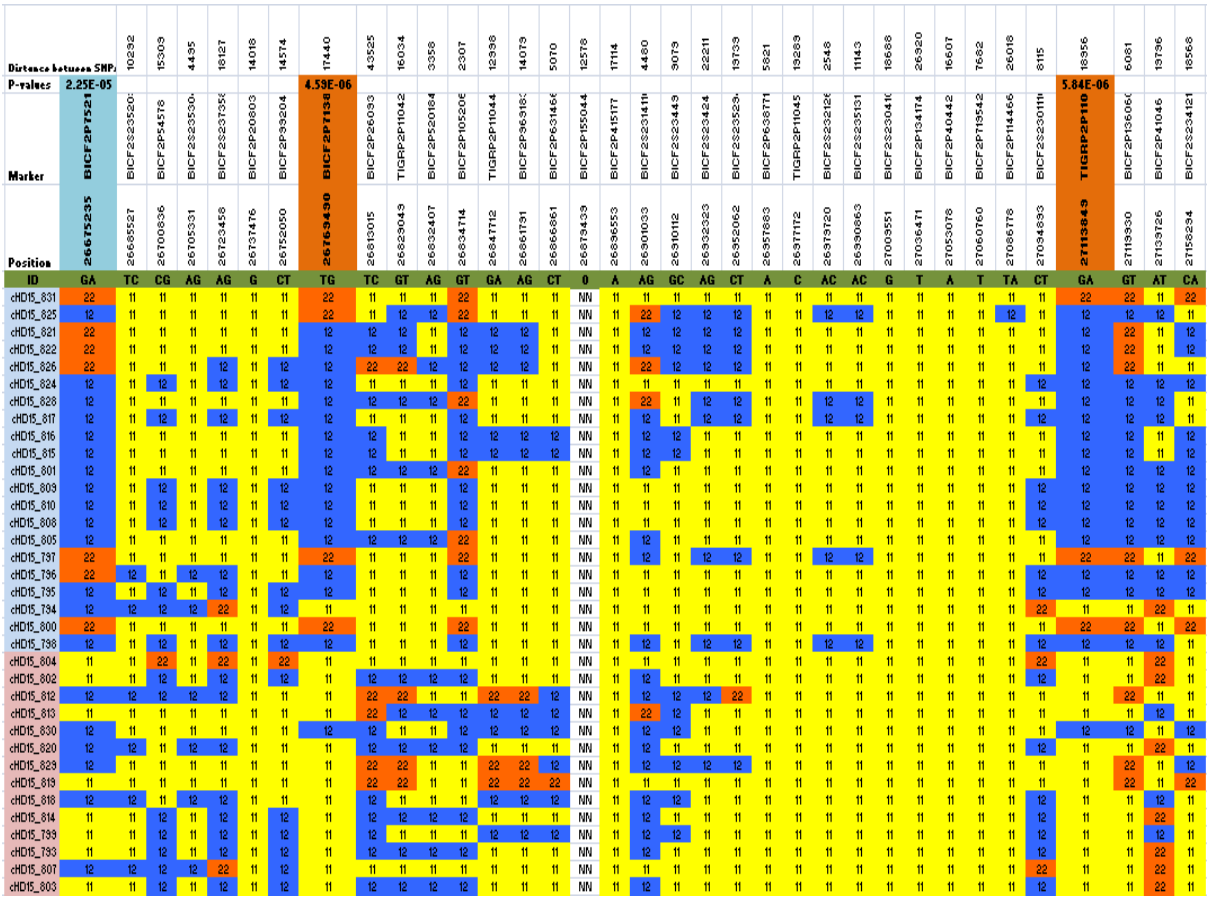
Figure 14 LD plot focussed to a region from ~"26675235" to ~ "27972553" comprising around 300 Kb. The two strongly associated SNPs are visualised by a bounded by blue border line.

In this clear picture of LD plot, pair wise LD values among markers are displayed on Confidence Bounds Color Scheme. Here, "Dark grey" refers to the evidence of strong LD (r^2=1), "light grey" refers to uninformative LD ($0 < r^2 < 1$) and "white" refers to the strong evidence of recombination (r^2 =0). The yellow circle on the dark grey area is the evidence of strong LD between the two strongly associated markers.

Two strongly associated markers in figure 14 appear to be linked tightly with r^2 =1. Only one of the two markers "TIGRP2P110467" is in a block of 25, 877 bp. The possible explanation might be that the threshold to exclude a marker on the basis of MAF is very low ($1x10^{-03}$) such that a marker without enough heterozygosity is also included, hence it acts as uninformative giving value of pair wise LD not enough to define a block. However, the patterns of the block formation can be modified with the change in parameters (Barrett, 2009). For further analysis, each block considered as a haplotype can be used for association analysis. In GWAs the distribution and structure of haplotype blocks can help to identify complex disease genes (Wang et al., 2002).

We can observe the genotypes of the individuals for the markers in figure 13 that might further help to explain the scenario in LD plot. The leftmost column contains "ID" of 35 individuals and the top part shaded with blue colour represents the control group and the lower part with pink shade represents the cases. The top two SNPs with strong association with disease phenotype is coloured in dark orange and third SNP in blue colour. The corrected p-values are mentioned as rows on top of the markers. The distance between these top two SNPs is 361,799 bp. Genotypes in the columns, "11" and "22" represents that individuals are homozygous for two different alleles

and is coloured in yellow and orange, respectively. Similarly, heterozygous genotype "12" is coloured in blue for ease of visualization. "NN" as seen in the white column denotes that the marker is not typed. The genotypes are extracted from the file without any prior quality control.



Distance between SNPs		10292	15309	4435	18127	14018	14574	17440	43525	16034	3358	2307	12938	14073	5070	12578	17114	4480	9079	22211	19739	5821	19283	2548	11143	16688	26320	16607	7682	26018	8115	16356	6081	19736	16568
P-values	2.25E-05							4.59E-06																								5.84E-06			
Marker	BICF2P75211	BICF2S23520	BICF2P54578	BICF2S23530	BICF2S23735	BICF2P20803	BICF2P33204	BICF2P713861	BICF2P26093	TIGRP2P11042	BICF2P520164	BICF2P105206	TIGRP2P11044	BICF2P36316	BICF2P631466	BICF2P155044	BICF2P415177	BICF2S23141	BICF2S23443	BICF2S23424	BICF2S23523	BICF2P638771	TIGRP2P11045	BICF2S23212	BICF2S23513	BICF2S23041	BICF2P134174	BICF2P40442	BICF2P119542	BICF2P114466	BICF2S23011	TIGRP2P110	BICF2P136061	BICF2P41046	BICF2S23412
Position	26675235	26685527	26700836	26705331	26723458	26737476	26752050	26763430	26813015	26823043	26832407	26834714	26847712	26861731	26866861	26873439	26836553	26901033	26910112	26932323	26952062	26957883	26977172	26979720	26990863	27003551	27036471	27053078	27060760	27086778	27094383	27113843	27119330	27139726	27158234
ID	GA	TC	CG	AG	AG	G	CT	TG	TC	GT	AG	GT	GA	AG	CT	0	A	AG	GC	AG	CT	A	C	AC	AC	G	T	A	T	TA	CT	GA	GT	AT	CA
cHD15_831	22	11	11	11	11	11	11	22	11	11	11	22	11	11	11	NN	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	22	22	11	22
cHD15_825	12	11	11	11	11	11	11	22	11	12	12	22	11	11	11	NN	11	22	12	12	12	11	11	12	12	11	11	11	11	12	11	12	12	12	11
cHD15_821	22	11	11	11	11	11	11	12	12	12	11	12	12	12	11	NN	11	12	12	12	12	11	11	11	11	11	11	11	11	11	11	12	22	11	12
cHD15_822	22	11	11	11	11	11	11	12	12	12	11	12	12	12	11	NN	11	12	12	12	12	11	11	11	11	11	11	11	11	11	11	12	22	11	12
cHD15_826	22	11	11	11	12	11	12	12	22	22	12	12	12	12	11	NN	11	22	12	12	12	11	11	11	11	11	11	11	11	11	11	12	22	11	11
cHD15_824	12	11	12	11	12	11	12	12	11	11	11	12	11	11	11	NN	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	12	12	12	12
cHD15_828	12	11	11	11	11	11	11	12	12	12	12	22	11	11	11	NN	11	22	11	12	12	11	11	12	12	11	11	11	11	11	11	12	12	12	11
cHD15_817	12	11	12	11	12	11	12	12	11	11	11	12	11	11	11	NN	11	12	11	12	12	11	11	12	12	11	11	11	11	11	12	12	12	12	11
cHD15_816	12	11	11	11	11	11	11	12	12	11	11	12	12	12	12	NN	11	12	12	11	11	11	11	11	11	11	11	11	11	11	11	12	12	11	12
cHD15_815	12	11	11	11	11	11	11	12	12	11	11	12	12	12	12	NN	11	12	12	11	11	11	11	11	11	11	11	11	11	11	11	12	12	11	12
cHD15_801	12	11	11	11	11	11	11	12	12	12	12	22	11	11	11	NN	11	12	11	11	11	11	11	11	11	11	11	11	11	11	11	12	12	12	12
cHD15_809	12	11	12	11	12	11	12	12	11	11	11	12	11	11	11	NN	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	12	12	12	12
cHD15_810	12	11	12	11	12	11	12	12	11	11	11	12	11	11	11	NN	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	12	12	12	12
cHD15_808	12	11	12	11	12	11	12	12	11	11	11	12	11	11	11	NN	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	12	12	12	12
cHD15_805	12	11	11	11	11	11	11	12	12	12	12	22	11	11	11	NN	11	12	11	11	11	11	11	11	11	11	11	11	11	11	11	12	12	12	12
cHD15_797	22	11	11	11	11	11	11	22	11	11	11	22	11	11	11	NN	11	12	11	12	12	11	11	12	12	11	11	11	11	11	11	22	22	11	22
cHD15_796	22	12	11	12	12	11	11	12	11	11	11	12	11	11	11	NN	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	12	12	12	12
cHD15_795	12	11	12	11	12	11	12	12	11	11	11	12	11	11	11	NN	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	12	12	12	12
cHD15_794	12	12	12	12	22	11	12	11	11	11	11	11	11	11	11	NN	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	22	11	11	22	11
cHD15_800	22	11	11	11	11	11	11	22	11	11	11	22	11	11	11	NN	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	22	22	11	22
cHD15_798	12	11	12	11	12	11	12	12	11	11	11	12	11	11	11	NN	11	12	11	12	12	11	11	12	12	11	11	11	11	11	11	12	12	12	12	11
cHD15_804	11	11	22	11	22	11	22	11	11	11	11	11	11	11	11	NN	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	22	11	11	22	11
cHD15_802	11	11	12	11	12	11	12	11	12	12	12	12	11	11	11	NN	11	12	11	11	11	11	11	11	11	11	11	11	11	11	11	12	11	11	22	11
cHD15_812	12	12	12	12	12	11	11	11	22	22	11	11	22	22	12	NN	11	12	12	12	22	11	11	11	11	11	11	11	11	11	11	11	11	22	11	11
cHD15_813	11	11	11	11	11	11	11	11	22	12	12	12	12	12	12	NN	11	22	12	11	11	11	11	11	11	11	11	11	11	11	11	11	11	12	11	11
cHD15_830	12	11	11	11	11	11	11	12	12	11	11	12	12	12	12	NN	11	12	12	11	11	11	11	11	11	11	11	11	11	11	11	12	12	11	12	11
cHD15_820	12	12	11	12	12	11	11	11	12	12	12	12	11	11	11	NN	11	12	11	11	11	11	11	11	11	11	11	11	11	11	11	12	11	11	22	11
cHD15_829	12	11	11	11	11	11	11	11	22	22	11	11	22	22	12	NN	11	12	12	12	12	11	11	11	11	11	11	11	11	11	11	11	22	11	12	11
cHD15_819	11	11	11	11	11	11	11	11	22	22	11	11	22	22	22	NN	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	22	11	22	11
cHD15_818	12	12	11	12	12	11	11	11	12	11	11	11	12	12	12	NN	11	12	12	11	11	11	11	11	11	11	11	11	11	11	11	12	11	11	12	11
cHD15_814	11	11	12	11	12	11	12	11	12	12	12	11	11	11	11	NN	11	12	11	11	11	11	11	11	11	11	11	11	11	11	11	12	11	11	22	11
cHD15_799	11	11	12	11	12	11	12	11	12	11	11	11	12	12	12	NN	11	12	12	11	11	11	11	11	11	11	11	11	11	11	11	12	11	11	12	11
cHD15_793	11	11	12	11	12	11	12	11	12	12	12	12	11	11	11	NN	11	12	11	11	11	11	11	11	11	11	11	11	11	11	11	12	11	11	22	11
cHD15_807	12	12	12	12	22	11	12	11	11	11	11	11	11	11	11	NN	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	22	11	11	22	11
cHD15_803	11	11	12	11	12	11	12	11	12	12	12	12	11	11	11	NN	11	12	11	11	11	11	11	11	11	11	11	11	11	11	11	12	11	11	22	11

Figure 15 The genotypes of individuals for SNPs on CFA 8 from position 26,675,235 to 27,158,294 bp.

The genotypes in the columns for two strongly associated SNPs infer a potential association with the disease, with the control mostly heterozygous and the cases mostly homozygous. The haplotype block with SNP "TIGRP2P110467" can be viewed from the right side, starting from 2nd to 4th SNP. An extended homozygosity block that can separate between cases and controls could not be determined.

Odds ratio:

With the concept of LD, if the SNPs are in strong LD with the causative variant then odds ratio for the SNPs can also be related to the odds ratio of the causative variant. The Odds ratio (OR) of 41.33 of the T-allele for SNP "**BICF2P713861"** with corrected P-value: 4.59E-06 indicates strong association of the allele with the disease phenotype as similar is the case with "G" allele of SNP "**TICRP2P110467**" (corrected p-value: 5.84E-06) with an OR of 37.52. The detailed result can be viewed in the table below:

Table 4 The OR for both alleles of two markers along with the confidence interval, p-value for test statistic, MAF distribution is described. For marker "BICF2P713861" G is the minor allele and has frequency of 0.35. For marker "TIGRP2P110467", A is the minor allele with frequency of 0.338. p-value <0.0001 denotes significance of test statistic.

SNP	**For 35 indvs:**				**95% Conf. Interval**			
	%Geno	**MAF**	**Allele:**	**Odds-ratio**	**Lower limit**	**Upper limit**	**Yates value**	**p-value**
BICF2P713861	100	0.35 (G)	T	41.33	5.1482	331.855	21.34	<0.0001
			G	0.0242	0.003	0.1942	21.34	<0.0001
					95% Conf. Interval			
	%Geno	MAF	Allele:	Odds-ratio	Lower limit	Upper limit	Yates value	p-value
TIGRP2P110467	100	0.338 (A)	G	37.526	4.678	300.98	19.81	<0.0001
			A	0.0266	0.0033	0.2137	19.81	<0.0001

Sequencing of the coding region:

Analysis of the two markers ~300 kb apart in chromosome 8 revealed that SNPs lie in close proximity linked to a gene *MDGA2/MAMDC1* (MAM-domain containing glycosylphosphatidylinositol anchor 2) in a genome browser Ensembl derived from CanFam 2.0 assembly. According to Ensembl, *MAMDC1* spans a region of 317.42 kb on reverse strand of CFA 8 (chromosome 14 in humans), consists of 15 exons and codes for a protein of 844 residues. There is >97% amino acid identity among mouse, rat, dog, chimpanzee, horse and human which implies an important function of this gene (Hellquist et al., 2009). This protein is predicted to be a new member of immunoglobulin superfamily (IgCAM) and plays role in cell adhesion, migration and recruitment to inflammatory sites. According to Genecards (www.genecards.org –version 3.08), a database of human genes (known and predicted) the expression of MDGA2 is observed in retinal tissues of human. However, in Canine, the annotation of gene has been derived from human. According to UniProtKB (www.uniprot.org), version 10 (a resource for protein sequence and functional information), canine MDGA2 protein is not characterized. To date, MDGA2 proteins have only been manually annotated for four species: human, rat, mouse and Cynomolgus monkey.

Following results on table 5 were obtained for the sequencing of 1, 441 bp long promoter and 15 exons for the cases and controls:

Table 5 The genotypes that differed for the cases and controls at different positions of promoter (denoted as Pro with numbers: 60, 234, 281, 895, 1023, 1138 and 1333) region, exon 6 (66th) and exon 15(228th) are described. "-" represent identical nucleotide compared to the reference sequence at that position.

	Positions	**Pro-60**	**Pro-234**	**Pro-281**	**Pro-895**	**Pro-1023**	**Pro-1138**	**Pro-1333**	**Exon 6: 66**	**Exon 15: 228**
	Reference (Ensembl)	**A**	**C**	**T**	**G**	**A**	**A**	**Deletion**	**T**	**C**
CASES	3	-	-	-	-	-	-	-	-	Y(C/T)
	15	R (G/A)	Y (C/T)	Y	R	R	R	W (A/T)	W (A/T)	M (A/C)
	33	R	Y	Y	R	R	R	W	W	Y
	36	R	Y	Y	R	R	R	T	T	C
	48	-	-	-	-	-	-	-	-	-
CONTROLS	8	G	T	C	A	G	G	T	A	T
	14	G	Y	C	A	R	R	T	W	M
	27	G	T	C	A	G	G	T	W	Y
	28	G	T	C	A	G	G	T	W	Y
	30	R	Y	Y	R	R	R	-	W	T

There are seven positions in the promoter region where some cases and controls differed. Two of the cases (CCR 03 and CCR 48) are homozygous "A" for position 60 where as other cases are heterozygous "R: G/A". All controls are homozygous "G" except for CCR30 which is heterozygous "R". Similarly, genotypes of the cases and controls can be observed for other six positions mentioned in the region. At position 1333, all the controls have an insertion "T" except for CCR30. In cases, CCR15, CCR 33 and CCR36 have insertion. CCR 15 and CCR33 are heterozygous "W: A/T" whereas CCR36 is homozygous "T" for insertion. CCR03 and CCR 48 don't have any insertion in that position.

For exon 6, cases CCR03, CCR 36 and CCR 48 are homozygous "T" for the position 66, where as other two cases CCR 15 and CCR 33 are heterozygous "W:A/T". All the controls are heterozygous "W" except for CCR 08 which is homozygous "A". For exon 15, two cases CCR 36 and CCR 48 are homozygous "C" for position 228, CCR 15 is heterozygous "M:A/C" whereas CCR 03 and CCR 33 are heterozygous "Y: C/T". In controls, CCR 08 and CCR 30 are homozygous "T", CCR 27 and CCR 28 are heterozygous "Y: C/T" and CCR 14 is heterozygous "M: A/C".

With above analysis, only one case CCR 48 and CCR 08 are homozygous in every position and differ from one another. Also, CCR 48 is identical to the reference sequence where as CCR08 is not. This result is not conclusive and a definite conclusion can be derived with addition of more individuals in the sequence analysis.

The multiple alignment of translated sequence (844 residues) for one case (CCR48) and one control (CCR08) along with the reference amino acid residue from Ensembl determined a non-synonymous change on position 348 where the essential amino acid "T"-Threonine on reference sequence is changed to non-essential amino acid "S"-Serine in the control sample. Both amino acids are polar (hydrophilic) in nature. Multiple aligned sequences can be observed in the appendix section.

The changes in different nucleotides for the cases and controls need to be justified as the reference sequence used is based on predicted annotation. New canine assembly CanFam 3.1 has been developed (personal information). The predicted annotation for MDGA2 gene can be evaluated for any incomplete explanation of the region. If new features are available for example change in exon annotation then, Exome sequencing can be performed for those exons. Other possible steps can be sequencing of a whole region of MDGA2, around 300 kb in length, to identify causal variant for disease phenotype. Another approach could be to isolate mRNA from the retinal tissue and compare the expression profile of the specific gene in the tissue (Shimizu-Matsumoto et al., 1997). Mutations are not always non-synonymous. Hence, an approach that has a capacity to detect as many potential ways of causing mutations should be applied for mutational screening.

In GWAS the association observed could be direct, indirect or spurious. The association is direct when the associated marker is itself involved in causing a disease. An indirect association occurs when the associated allele is not directly involved but is in LD with the alleles causing disease. The occurrence of two SNPs on CFA 8 in almost all models with least value for the chance of false positive error strengthens a possibility of obtained SNPs to be truly associated with disease phenotype and not a spurious association although this association is not significant after correcting for the multiple testing using Bonferroni correction and Permutation methods. However, the potential association to chromosome 8 can be further evaluated. GWAS is susceptible to many errors and biases. The spurious association might be related to statistical fluctuations (occur by chance and generate low p-values), bias due to study design and due to some technical errors (Hirschhorn and Daly, 2005). Hence to evaluate the association from GWA finding, replication plays a crucial role (McCarthy et al., 2008). Replication helps to identify the associations that are reproducible by evaluating the associations in other independent samples hence validating a true association. To avoid spurious association one of the essential criteria is to have a distinct definition for a disease that we are analysing. However, in this case, age of onset among cases vary from ~3 yrs to ~11 yrs. Again, epistasis, interaction of gene and environment influences such factors and complicates a specific definition for a disease.

Karlsson et al. (20007) mentions that due to a genome structure of dogs and variation within and across different breeds of dogs, the region associated with certain trait could be identified initially within a breed using a sparse marker set (presence of long haplotype blocks due to extended LD) and further, the region could be confirmed by using a dense marker array taking multiple breeds (short haplotype blocks across breeds). For e.g: Mapping a coat colour locus in white boxer and bull terrier initially by independent association analysis within both breeds and further association analysis including both breeds to fine map the region of association. This application of this approach seems to be a bit complicated in case of CCD. Firstly, the disease is rare in a population of CCD that collection of unrelated samples to avoid stratification is quite complicated. Such

features of a disease have not been reported in any other breeds apart from CCD. Hence, performing a GWAS across breeds seems to be complicated in case of CCD.

After revision of all the results, the two SNPs on CFA 8 identified with strong association to the disease phenotype can be taken as the potential SNPs. Further work related with this study can be directed in many ways. In order to confirm it as a significant association, association analysis with additional dogs need to be performed. As the SNPs appear to be potential marker, a gene MDGA2 linked to their loci also seems as a potential candidate gene, however further analysis need to be done for a definite conclusion. Also, MDGA2 molecule is highly expressed in peripheral and central nervous system and is involved in cell adhesion (Joset et al., 2011). A primary feature of this retinal disease includes detachment of RPE cells and migration toward the outer segment of photoreceptor cells and retina is a part of the central nervous system with different neuronal layers involved in transferring stimuli from photoreceptor cells to the optic nerve. These scenarios also suggest MDGA2 to be a potential candidate for the disease. Hence, MDGA2 gene should be further evaluated based on the annotation obtained from a new CanFam 3.1 assembly (which is currently unavailable). Hence, an identification of the genetic basis of any disease helps in developing the molecular tools to better diagnose the disease and better application in field of breeding to create healthy population.

REFERENCES

AGUIRRE, G. D. & ACLAND, G. M. 2006. Models, Mutants, and Man: Searching for Unique Phenotypes and Genes in the Dog Model of Inherited Retinal Degeneration. *In:* OSTRANDER, E. A., GIGER, U. & LINDBLAD-TOH, K. (eds.) *The Dog and Its Genome.* New York: Cold Spring Harbor Laboratory Press.

APPLEYARD, G. D., FORSYTH, G. W., KIEHLBAUCH, L. M., SIGFRID, K. N., HANIK, H. L. J., QUON, A., LOEWEN, M. E. & GRAHN, B. H. 2006. Differential Mitochondrial DNA and Gene Expression in Inherited Retinal Dysplasia in Miniature Schnauzer Dogs. *Investigative Ophthalmology & Visual Science,* 47.

AULCHENKO, Y. 2010. ABEL tutorial.

BALDING, D. J. 2006. A tutorial on statistical methods for population association studies. *NATURE REVIEWS,* 7**,** 781-791.

BARRETT, J. C. 2009. Haploview: Visualization and Analysis of SNP Genotype Data. *Cold Spring Harbor Protocols*.

CAMPBELL, C. D., OGBURN, E. L., LUNETTA, K. L., LYON, H. N., FREEDMAN, M. L., GROOP, L. C., ALTSHULER, D., ARDLIE, K. G. & HIRSCHHORN, J. N. 2005. Demonstrating stratification in a European American Population. *Nature Genetics,* 38.

CLARKE, G. M., ANDERSON, C. A., PETTERSSON, F. H., CARDON, L. R., P, A., MORRIS & ZONDERVAN, K. T. 2011. Basic statistical analysis in genetic case-control studies. *Nature Protocol,* 6**,** 121-133.

CLAUDE, F. 2012. *Yeast as a model and tool to study mitochondrial diseases (NARP)* [Online]. [Accessed 12 June 2012].

DEVLIN, B. & ROEDER, K. 1999. Genomic Control for Association Studies. *Biometrics,* 22**,** 997-1004.

DICK, D. M. 2008. Introduction to association. *In:* NEALE, B. M., FERREIRA, M. A., MEDLAND, S. E. & POSTHUMA, D. (eds.) *Statistical Genetics: Gene Mapping through Linkage and Association.* Taylor & Francis Group.

DOH, S. T., HAO, H., LOH, S. C., PATEL, T., TAWIL, H. Y., CHEN, D. K., PASHKOVA, A., SHEN, A., WANG, H. & CAI, L. 2010. Analysis of retinal cell development in chick embryo by immunohistochemistry and in ovo electroporation techniques. *BMC Developmental Biology,* 10.

DOWNS, L. M., WALLIN-HÅKANSSON, B., BOURSNELL, M., MARKLUND, S., HEDHAMMAR, Å., TRUVÉ, K., HÜBINETTE, L., LINDBLAD-TOH, K., BERGSTRÖM, T. & MELLERSH, C. S. 2011. A Frameshift Mutation in Golden Retriever Dogs with Progressive Retinal Atrophy Endorses SLC4A3 as a Candidate Gene for Human Retinal Degenerations. *PLoS Genetics,* 6.

GIGER, U., SARGAN, D. R. & MCNIEL, E. A. 2006. Breed-specific Hereditary Diseases and Genetic Screening. *In:* OSTRANDER, E. A., GIGER, U. & LINDBLAD-TOH, K. (eds.) *The Dog and Its Genome.* New York: Cold Spring Harbor Laboratory Press.

GUZIEWICZ, K. E., ZANGERL, B., LINDAUER, S. J., MULLINS, R. F., SANDMEYER, L. S., GRAHN, B. H., STONE, E. M., ACLAND, G. M. & AGUIRRE, G. D. 2007. Bestrophin Gene Mutations Cause Canine Multifocal Retinopathy: A Novel Anima Model for Best Disease. *Invest Oohthalmol Vis Sci,* 48**,** 1959-1967.

HELLQUIST, A., ZUCCHELLI, M., LINDGREN, C. M., SAARIALHO-KERE, U., JA¨RVINEN, T. M., KOSKENMIES, S., JULKUNEN, H., ONKAMO, P. I., SKOOG, T., PANELIUS, J., ISA¨NEN-SOKOLOWSKI, A. R., HASAN, T., WIDEN, E., GUNNARSON, I., SVENUNGSSON, E., PADYUKOV, L., ASSADI, G., BERGLIND, L., MA¨KELA¨, V.-V., KIVINEN, K., WONG, A., GRAHAM, D. S. C., VYSE, T. J., D'AMATO, M. & KERE, J. 2009. Identification of MAMDC1 as a Candidate Susceptibility Gene for Systemic Lupus Erythematosus (SLE). *PLoS one,* 4.

HIRSCHHORN, J. N. & DALY, M. J. 2005. Genome-wide Association Studies for Common Diseases and Complex Traits. *NATURE REVIEWS,* 6**,** 95-108.

JOSET, P., WACKER, A., BABEY, R., INGOLD, E. A., ANDERMATT, I., STOECKLI, E. T. & GESEMANN, M. 2011. Rostral growth of commissural axons requires the cell adhesion molecule MDGA2. *Neural Development,* 6.

KARLSSON, E. K., BARANOWSKA, I., WADE, C. M., HILLBERTZ, N. H. C. S., ZODY, M. C., ANDERSON, N., BIAGI, T. M., PATTERSON, N., PIELBERG, G. R., KULBOKAS, E. J., COMSTOCK, K. E., KELLER, E. T., MESIROV, J. P., EULER, H. V., KÄMPE, O., HEDHAMMAR, Å., LANDER, E. S., ANDERSSON, G., ANDERSSON, L. & LINDBLAD-TOH, K. 2007. Efficient mapping of mendelian traits in dogs through genome-wide association. *Nature Genetics,* 39**,** 1321-1328.

KARLSSON, E. K. & LINDBLAD-TOH, K. 2008. Leader of the pack: gene mapping in dogs and other model organisms. *NATURE REVIEWS,* 9**,** 713-725.

KIERCZAK, M., SHEN, X., TENGVALL, K., AULCHENKO, Y. & CARLBORG, Ö. 2011. GWAS Analysis in Stratified Populations - LUPA Workshop.

KUKSA, V., IMANISHI, Y., BATTEN, M., PALCZEWSKI, K. & MOISE, A. R. 2003. Retinoid cycle in the vertebrate retina: experimental approaches and mechanisms of isomerization. *Vision Research,* 43**,** 2959-2981.

LINDBLAD-TOH, K., WADE, C. M., MIKKELSEN, T. S., KARLSSON, E. K., JAFFE, D. B., KAMAL, M., CLAMP, M., CHANG, J. L., III, E. J. K., ZODY, M. C., MAUCELI, E., XIE, X., BREEN, M., WAYNE, R. K., OSTRANDER, E. A., PONTING, C. P., GALIBERT, F., SMITH, D. R., DEJONG, P. J., KIRKNESS, E., ALVAREZ, P., BIAGI, T., BROCKMAN, W., BUTLER, J., CHIN, C.-W., COOK, A., CUFF, J., DALY, M. J., DECAPRIO, D., GNERRE, S., GRABHERR, M., KELLIS, M., KLEBER, M., BARDELEBEN, C., GOODSTADT, L., HEGER, A., HITTE, C., KIM, L., KOEPFLI, K.-P., PARKER, H. G., POLLINGER, J. P., SEARLE, S. M. J., SUTTER, N. B., THOMAS, R., WEBBER, C., PLATFORM, B. I. G. S. & LANDER, E. S. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature,* 438.

MCCARTHY, M. I., ABECASIS, G. R., CARDON, L. R., GOLDSTEIN, D. B., LITTLE, J., IOANNIDIS, J. P. A. & HIRSCHHORN, J. N. 2008. Genome-wide association studies for complex traits:consensus, uncertainty and challenges. *Nature Review,* 9**,** 356-369.

MIYADERA, K., ACLAND, G. M. & AGUIRRE, G. D. 2012a. Genetic and phenotypic variations of inherited retinal diseases in dogs: the power of within- and across-breed studies. *Mamm Genome,* 23**,** 40-61.

MIYADERA, K., ACLAND, G. M. & AGUIRRE, G. D. 2012b. Genetic and phenotypic variations of inherited retinal diseases in dogs: the power of within- and across-breed studies. *Mammalian Genome,* 23**,** 40-61.

NARFSTRÖM, K., JALOMÄKI, S., MOWAT, F., SAMARDZIJA, M., CHAUDIEU, G., BERGSTRÖM, T. & GRIMM, C. Clinical assessment of a novel pigmentary chorioretinopathy in the Chinese Crested Dog.

OPTIGEN, L. *prcd-PRA Test* [Online]. Available: http://www.optigen.com/opt9_test_prachc.html [Accessed].

PEARSON, T. A. & MANOLIO, T. A. 2008. How to Interpret a Genome-wide Association Study. *JAMA,* 299**,** 1335-1344.

PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. & REICH, D. 2006. Principal component analysis corrects for stratification in genome-wide association studies. *Nature Genetics,* 38.

PRICE, A. L., ZAITLEN, N. A., REICH, D. & PATTERSON, N. 2010. New approaches to population stratification in genome-wide association studies. *Nature Review Genetics*.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A. R., BENDER, D., MALLER, J., SKLAR, P., BAKKER, P. I. W. D., DALY, M. J. & SHAM, P. C. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics,* 81**,** 559-575.

PURVES, D., AUGUSTINE, G. & FITZPATRICK, D. 2001. Neuroscience. *The Retina*. 2nd ed. Sunderland (MA): Sinauer Associates.

ROBERSON, E. D. O. & PEVSNER, J. 2009. Visualization of Shared Genomic Regions and Meiotic Recombination in High-Density SNP Data. *PLoS one,* 4.

SATO, Y., SUGANAMI, H., HAMADA, C., YOSHIMURA, I., SAKAMOTO, H., YOSHIDA, T. & YOSHIMURA, K. 2006. The confidence interval of allelic odds ratios under the Hardy–Weinberg disequilibrium. *J Hum Genet,* 51**,** 772-780.

SETAKIS, E., STIRNADEL, H. & BALDING, D. J. 2006. Logistic regression protects against population structure in genetic association studies. *Genome Research,* 16**,** 290-296.

SHIMIZU-MATSUMOTO, A., ADACHI, W., MIZUNO, K., INAZAWA, J., KOHJI NISHIDA, KINOSHITA, S., MATSUBARA, K. & OKUBO, K. 1997. An Expression Profile of Genes in Human Retina and Isolation of a Complementary DNA for a Novel Rod Photoreceptor Protein. *Investigative Ophthalmology & Visual Science,* 38**,** 2576-2585.

TSAI, M.-F., LIN, Y.-J., CHENG, Y.-C., LEE, K.-H., HUANG, C.-C., CHEN, Y.-T. & YAO, A. 2007. PrimerZ: streamlined primer design for promoters, exons and human SNPs. *Nucleic Acids Research,* 35.

VAYSSE, A., RATNAKUMAR, A., DERRIEN, T., AXELSSON, E., PIELBERG, G. R., SIGURDSSON, S., FALL, T., SEPPÄLÄ, E. H., HANSEN, M. S. T., LAWLEY, C. T., KARLSSON, E. K., CONSORTIUM, T. L., BANNASCH, D., VILÀ, C., LOHI, H., GALIBERT, F., FREDHOLM, M., HÄGGSTRÖM, J., HEDHAMMAR, Å., ANDRÉ, C., LINDBLAD-TOH, K., HITTE, C. & WEBSTER, M. T. 2011. Identification of Genomic Regions Associated with Phenotypic Variation between Dog Breeds using Selection Mapping. *PLoS Genetics,* 7.

VESKE, A., NILSSON, S. E. G., NARFSTRÖM, K. & GAL, A. 1999. Retinal Dystrophy of Swedish Briard/Briard–Beagle Dogs is due to a 4-bp Deletion in RPE65. *Genomics,* 57**,** 57-61.

VILBOUX, T., CHAUDIEU, G., JEANNIN, P., DELATTRE, D., HEDAN, B., BOURGAIN, C., QUENEY, G., GALIBERT, F., THOMAS, A. & ANDRE, C. 2008. Progressive Retinal Atrophy in the Border Collie: A new XLPRA. *BMC Veterinary Research,* 4.

VOIGHT, B. F. & PRITCHARD, J. K. 2005. Confounding from cryptic relatedness in case-control association studies. *PLoS Genetics,* 1**,** 302-311.

WANG, N., AKEY, J. M., ZHANG, K., CHAKRABORTY, R. & JIN, L. 2002. Distribution of Recombination Crossovers and the Origin of Haplotype Blocks: The Interplay of Population History, Recombination, and Mutation. *Am. J. Hum. Genet.,* 71**,** 1227-1234.

WEI, Z., SUN, W., WANG, K. & HAKONARSON, H. 2009. Multiple testing in genome-wide association studies via hidden Markov models. *Bioinformatics,* 25**,** 2802-2808.

YU, J., PRESSOIR, G., BRIGGS, W. H., BI, I. V., YAMASAKI, M., DOEBLEY, J. F., MCMULLEN, M. D., GAUT, B. S., NIELSEN, D. M., HOLLAND, J. B., KRESOVICH, S. & BUCKLER, E. S. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics,* 38**,** 203-208.

APPENDIX

Commands:

These are the commands used in GenABEL package in R.

To convert plink files to GenABEL format:

convert.snp.ped("data.ped", "data_edited.map", "genotype.raw")

Loading files:

load.gwaa.data(pheno="phenotype.dat", geno="genotype.raw", makemap=FALSE, sort=TRUE) -> data

First quality control:

check.marker(data, call=0.95, perid.call=0.95,extr.call=0.1, extr.perid.call=0.1,ibs.threshold=0.95,ibs.mrk="all",ibs.exclude="both",maf=1e-08, p.lev=1e-08, XXY.call=0.8, intermediateXF=c(0.5, 0.5)) -> firstqc

Transferring edited data to new data1:

data1 <- data[firstqc$idok, firstqc$snpok]

####### To check association of "sex" variable to the disease status:

tab <- table(phdata(data1)$status, phdata(data1)$sex)

fisher.test(tab)

Making genomic kinship matrix:

autosomalmarkers <- which(chromosome(data1) != "39")

length(autosomalmarkers)

autosomalmarkernames <- snpnames(data1)[autosomalmarkers]

data1genomickinship <- ibs(data1 [, autosomalmarkernames], weight ="freq")

data1distance <- as.dist(0.5 - data1genomickinship)

To make classical MDS plot and K=5 denotes the number of principal components shown:

data1mds <- cmdscale(data1distance, k=5)

Assigning case and control group to datamds:

controls <- which(phdata(data1)$status == 0)

controlsX <- data1mds [controls, 1]

controlsY <- data1mds [controls, 2]

cases <- which(phdata(data1)$status == 1)

```
casesX <- data1mds[cases, 1]
casesY <- data1mds[cases, 2]

###### Making MDS plot coloring Cases and Controls:
plot(data1mds, xlab="MDS1", ylab="MDS2")
points(controlsX, controlsY, pch=19, cex=0.7, col="green")
points(casesX, casesY, pch=19, cex=0.7, col="red")
legend(x="topright", c("cases", "controls"), col=c("red", "green"), pch=c(19, 19), ncol=1, bty="n", pt.cex= c(0.5, 0.5))


###### Identifying outliers and removing them (this process is done only while removing outliers):
#####outliers <- identify(data1mds)
####outliers

##### outliers and non_outliers names:
###outliers_names <- phdata(data1)[outliers,]$id
###non_outliers_names <- phdata(data1)[-outliers,]$id
###outliers_names

#### Transferring non outlier to data2:
###data2 <- data1[non_outliers_names,]
###nids(data2)

##### Second quality control without removing outliers:
check.marker(data1, hweids=(phdata(data1)$status == 0), call=0.95, perid.call=0.95, maf=0.05, fdrate=0.2) -> secondqc
data2 <- data1[secondqc$idok, secondqc$snpok]
nsnps(data2)

###### Fast case-control analysis computes chi-square test from 2x2 (allelic) or 2x3 (genotypic) tables:
ANfast_cc <- ccfast("status", data2, clambda=FALSE,propPs=1.0)
lambda(ANfast_cc)
descriptives.scan(ANfast_cc, top=30, sort="Pc1df")

bonferroni <- -log10(0.05/nsnps(data2))
```

Genomic Control:

ANgc <- qtscore(status, data2, trait.type="binomial", clambda=FALSE, propPs)

Firstly make genomic kinship matrix from data2 which passed second quality control test and then make second MDS plot from the new data and do *Classical multidimensional scaling...*

data2distance <- as.dist(0.5 - data2genomickinship)

data2mds <- cmdscale(data2distance)

To determine the structure in the population:

wss <- (nrow(data2mds) - 1) * sum(apply(data2mds, 2, var))

for (i in 2:10) wss[i] <- sum(kmeans(data2mds,centers=i, nstart=nids(data2))$withinss)

plot(1:10, wss, type="b", xlab="Number of Clusters", ylab="WSS")

Making three clusters:

km <- kmeans(data2mds, centers=3, nstart=nids(data2))

Determine clusters and coordinates of individuals for one cluster, similar procedure can be performed for other clusters:

cl1 <- which(km$cluster == 1)

cl1num <- as.numeric(cl1)

cl1x <- data2mds[cl1num,1]

cl1y <- data2mds[cl1num,2]

plot(data2mds,type="n", xlab="MDS1",ylab="MDS2", main="Subpopulations (K=3)")

points(cl1x, cl1y, pch=19, cex=.5, col="red")

legend(x="topright", c("subpop 1", "subpop 2", "subpop 3"),col=c("red","blue","yellow"), pch=c(19,19,19), ncol=1, bty="n", pt.cex=c(0.5,0.5,0.5))

####### To find the individuals that are in each specific clusters:

cl1names <- names(which(km$cluster == 1))

cl1names

Taking strata into account:

When strata is taken into account then scores are computed within strata and then added up:

ANgc_sa <- qtscore(status, data2, strata= cluster_3, trait.type="binomial", clambda=FALSE, propPs)

make a polygenic model for Mixed model:

h2a <- polygenic_hglm(status, data2, kin=data2genomickinship, trait="binomial")

Mixed model approach without permutation:

ANmixedmodel <- mmscore(h2a, data2)

###Mixed model approach and Structured association without permutation:

ANmmsa <- mmscore(h2a, data2, strata=cluster_3)

####Mixed model approach with 10000 permutations:

ANmixedmodel_perm <- mmscore(h2a, data2, times=10000)

Eigenscore (PCA):

Firstly, the diagonal of the genomic kinship matrix is replaced by the variance using hom() function:

diag(data2genomickinship) <- hom(data2[, autosomalmarkers])$Var

ANeigen <- egscore(status, data2, kinship=data2genomickinship)

The commands for QQ plot and Manhattan plot are not mentioned here.

Primers:

20 different M13 tailed primer pairs used for Exome sequencing of MDGA2 gene:

Oligoname	M13 tailed primer	Tm	Product size
MDGA2_Pro1_Frw	TGTAAAACGACGGCCAGTtgctggtatccctaagttcactg	60.54	565
MDGA2_Pro1_Rev	CAGGAAACAGCTATGACCtccaccaattcataatgcactc	59.83	
MDGA2_Pro2_Frw	TGTAAAACGACGGCCAGTtgggaaaatttgggattctg	59.73	659
MDGA2_Pro2_Rev	CAGGAAACAGCTATGACCaacaagctcttggttccatgt	58.71	
MDGA2_Pro3_Frw	TGTAAAACGACGGCCAGTtgatgaggaagtctaagagcaca	59.14	583
MDGA2_Pro3_Rev	CAGGAAACAGCTATGACCgttgcggagtcccagtaaaa	60.11	
MDGA2_Pro4_Frw	TGTAAAACGACGGCCAGTttgcaaaatgatcaccacaa	58.52	621
MDGA2_Pro4_Rev	CAGGAAACAGCTATGACCcactgggttaagtttttccttca	59.56	
MDGA2_E1_Frw	TGTAAAACGACGGCCAGTcaagttcatacaaagcccaaga	59.27	402
MDGA2_E1_Rev	CAGGAAACAGCTATGACCagccgaaacaatgaaacaca	59.17	

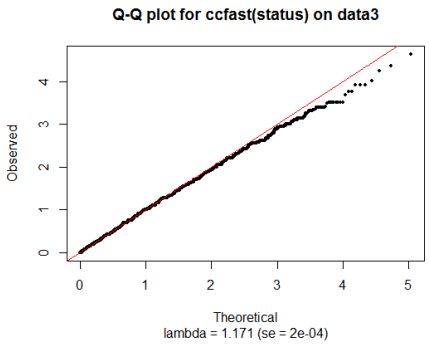
MDGA2_E2_Frw	TGTAAAACGACGGCCAGTtgaaaatgcacctttggtca	60.09	410
MDGA2_E2_Rev	CAGGAAACAGCTATGACCtcaaaggtcagatttggaaaga	58.81	
MDGA2_E3_Frw	TGTAAAACGACGGCCAGTaagacagtctgccacattttga	59.78	364
MDGA2_E3_Rev	CAGGAAACAGCTATGACCggggagcagagatgacttgt	59.26	
MDGA2_E4_Frw	TGTAAAACGACGGCCAGTtgattttggcaaaggaaaaga	59.68	500
MDGA2_E4_Rev	CAGGAAACAGCTATGACCtgcatgcaatatttgaggaaa	59.16	
MDGA2_E5_Frw	TGTAAAACGACGGCCAGTcatgttagcacagcacttgga	59.92	634
MDGA2_E5_Rev	CAGGAAACAGCTATGACCtttatcgggcaatcagaacg	60.96	
MDGA2_E6_Frw	TGTAAAACGACGGCCAGTcaaccagcagaaattgtttcc	59.6	490
MDGA2_E6_Rev	CAGGAAACAGCTATGACCccctaacccaatcacaaaatc	58.29	
MDGA2_E7_Frw	TGTAAAACGACGGCCAGTcggctattgttctccaacca	61.02	500
MDGA2_E7_Rev	CAGGAAACAGCTATGACCgacaagctactcaaagtatgttcca	58.94	
MDGA2_E8_Frw	TGTAAAACGACGGCCAGTtcgtctcttcctctcgcact	60.28	472
MDGA2_E8_Rev	CAGGAAACAGCTATGACCcctgggatttatgcaaaatgat	60.04	
MDGA2_E9_Frw	TGTAAAACGACGGCCAGTtttttgcccattttgttagaa	57.39	395
MDGA2_E9_Rev	CAGGAAACAGCTATGACCttggctgatgatccagaaga	59.32	
MDGA2_E10_Frw	TGTAAAACGACGGCCAGTatggtcacagtggggaagat	59.23	329
MDGA2_E10_Rev	CAGGAAACAGCTATGACCtgtctttgtacccatactcatcaaa	59.82	
MDGA2_E11_Frw	TGTAAAACGACGGCCAGTggcgccactattccaatct	60.05	384
MDGA2_E11_Rev	CAGGAAACAGCTATGACCagcattctgttgcatgtcatatt	58.67	
MDGA2_E12_Frw	TGTAAAACGACGGCCAGTtttcaatgagatcaaaattctcc	57.37	377
MDGA2_E12_Rev	CAGGAAACAGCTATGACCtcgacttcatgctcatgtttg	59.86	
MDGA2_E13_Frw	TGTAAAACGACGGCCAGTaacaccacagattcaagggatt	71.5	-
MDGA2_E13_Rev	CAGGAAACAGCTATGACCgagaattgactatccagcttcca	72.4	
MDGA2_E14_Frw	TGTAAAACGACGGCCAGTggtttggcttttcaatttctca	60.45	427
MDGA2_E14_Rev	CAGGAAACAGCTATGACCtgcttgccttacacagacca	60.45	
MDGA2_E15_e1Frw	TGTAAAACGACGGCCAGTtttccttttcacatgtcttgga	59.59	644
MDGA2_E15_e1Rev	CAGGAAACAGCTATGACCtgaaacaaaaccagccatga	60.09	
MDGA2_E15_e2Frw	TGTAAAACGACGGCCAGTaggcactggcatgaagaaag	60.4	426
MDGA2_E15_e2Rev	CAGGAAACAGCTATGACCatgccagctcttcacgtacc	60.29	

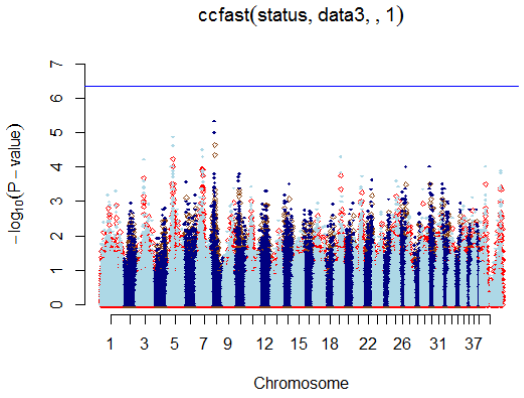
Top 10 SNPs for different models of association analysis for 35 individuals:

Basic Association analysis:

	Chromosome	Position	chi2.1df	P1df	Pc1df
BICF2P713861	8	26769490	21	4.59E-06	2.29E-05
TIGRP2P110467	8	27113849	19.53955	9.85E-06	4.41E-05
BICF2S232422	5	32983613	18.95833	1.34E-05	5.73E-05
BICF2G63035030	5	32197936	17.84257	2.40E-05	9.48E-05
BICF2G630555875	7	33168942	17.32323	3.15E-05	1.20E-04
BICF2P1341155	7	38798709	17.32323	3.15E-05	1.20E-04
BICF2P1195666	7	38810664	17.32323	3.15E-05	1.20E-04
BICF2G630558054	7	38778586	16.48263	4.91E-05	1.76E-04
BICF2P619504	19	38930835	16.48263	4.91E-05	1.76E-04
BICF2G630339850	3	43489192	16.12802	5.92E-05	2.06E-04
BICF2P739313	5	34211721	15.2381	9.48E-05	3.09E-04

Inflation factor	**1.170997**



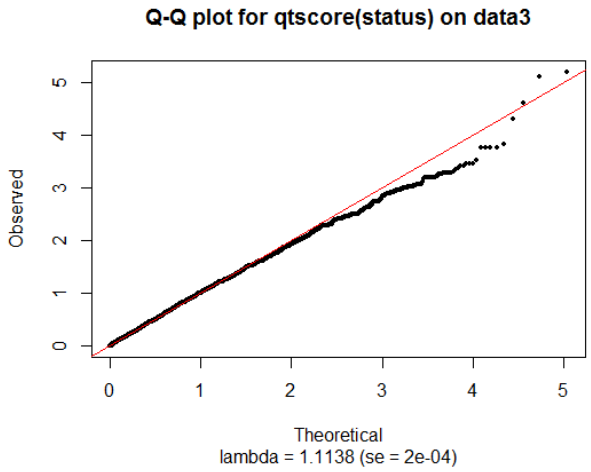


Genomic control:

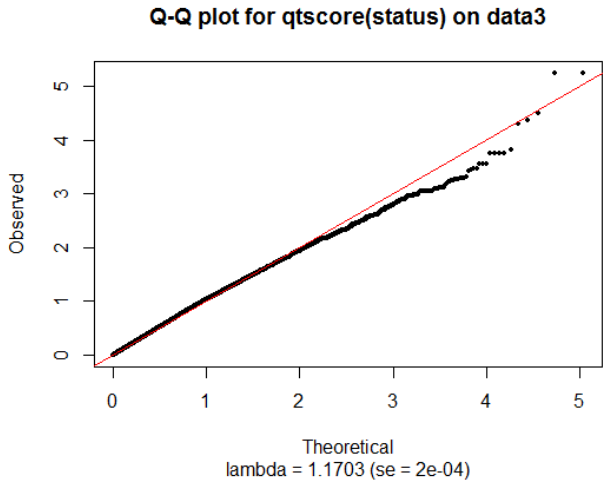
	Chromosome	Position	chi2.1df	P1df	Pc1df
TIGRP2P110467	8	27113849	22.75492	1.84E-06	6.19E-06
BICF2P713861	8	26769490	22.28774	2.35E-06	7.70E-06
BICF2S232422	5	32983613	19.8935	8.19E-06	2.38E-05
BICF2P1044496	8	27972553	18.32265	1.86E-05	5.00E-05
BICF2P752103	8	26675235	16.07506	6.09E-05	1.45E-04
TIGRP2P110728	8	27933541	15.72401	7.33E-05	1.72E-04
BICF2P1229493	8	28034747	15.72401	7.33E-05	1.72E-04
TIGRP2P110779	8	28046595	15.72401	7.33E-05	1.72E-04
BICF2P1271232	8	28382745	15.72401	7.33E-05	1.72E-04
BICF2S2308001	12	48682591	14.58333	1.34E-04	2.96E-04

Inflation factor

1.113844



Structured Association:



For strata =3

3 clusters:	Chromosome	Position	chi2.1df	P1df	Pc1df
BICF2P713861	8	26769490	24.12909	9.01E-07	5.61E-06
TIGRP2P110467	8	27113849	24.11187	9.09E-07	5.65E-06
BICF2P1044496	8	27972553	20.31726	6.56E-06	3.09E-05
BICF2S232422	5	32983613	19.61502	9.47E-06	4.24E-05
BICF2P752103	8	26675235	19.23058	1.16E-05	5.04E-05
BICF2S22946314	8	27895878	16.79685	4.16E-05	1.52E-04
TIGRP2P110728	8	27933541	16.5007	4.86E-05	1.73E-04
BICF2P1229493	8	28034747	16.5007	4.86E-05	1.73E-04
TIGRP2P110779	8	28046595	16.5007	4.86E-05	1.73E-04

BICF2P1271232	8	28382745	16.5007	4.86E-05	1.73E-04
Inflation factor	**1.170345**				

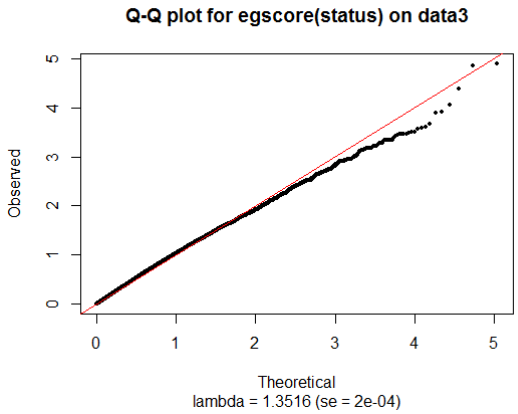
For 4 clusters:	**Chromosome**	**Position**	**chi2.1df**	**P1df**	**Pc1df**
BICF2P713861	8	26769490	22.65861	1.93E-06	8.79E-06
TIGRP2P110467	8	27113849	22.63706	1.96E-06	8.88E-06
TIGRP2P110728	8	27933541	19.42601	1.05E-05	3.86E-05
BICF2P1044496	8	27972553	19.42601	1.05E-05	3.86E-05
BICF2P1229493	8	28034747	19.42601	1.05E-05	3.86E-05
TIGRP2P110779	8	28046595	19.42601	1.05E-05	3.86E-05
BICF2P1271232	8	28382745	19.42601	1.05E-05	3.86E-05
BICF2P752103	8	26675235	18.13413	2.06E-05	7.00E-05
TIGRP2P109371	8	23439519	17.71218	2.57E-05	8.50E-05
BICF2S232422	5	32983613	17.44376	2.96E-05	9.62E-05
Inflation:	**1.146857**				

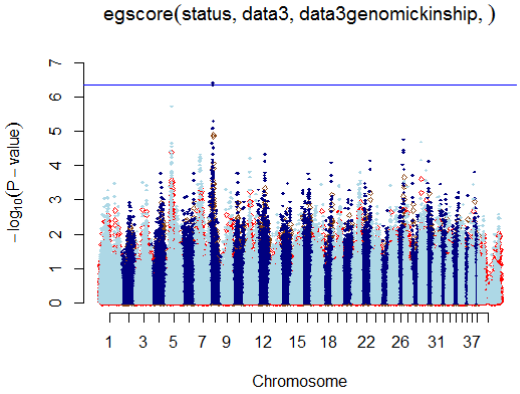
For 5 clusters:	**Chromosome**	**Position**	**chi2.1df**	**P1df**	**Pc1df**
TIGRP2P110467	8	27113849	23.1011	1.54E-06	1.11E-05
BICF2P713861	8	26769490	22.93346	1.68E-06	1.20E-05
TIGRP2P110728	8	27933541	19.26553	1.14E-05	5.99E-05
BICF2P1044496	8	27972553	19.26553	1.14E-05	5.99E-05
BICF2P1229493	8	28034747	19.26553	1.14E-05	5.99E-05
TIGRP2P110779	8	28046595	19.26553	1.14E-05	5.99E-05
BICF2P1271232	8	28382745	19.26553	1.14E-05	5.99E-05
BICF2P752103	8	26675235	19.21474	1.17E-05	6.13E-05
TIGRP2P109371	8	23439519	19.02244	1.29E-05	6.67E-05
BICF2S23649439	8	21298642	18	2.21E-05	1.05E-04
Inflation:	**1.196298**				

Principal component analysis:

	Chromosome	**Position**	**chi2.1df**	**P1df**	**Pc1df**
TIGRP2P110467	8	27113849	25.77222	3.84E-07	1.26E-05
BICF2P713861	8	26769490	25.50727	4.41E-07	1.40E-05
BICF2S232422	5	32983613	22.69999	1.89E-06	4.16E-05
BICF2P1044496	8	27972553	20.83687	5.00E-06	8.62E-05
BICF2P752103	8	26675235	19.94519	7.97E-06	1.22E-04
TIGRP2P109371	8	23439519	19.8398	8.42E-06	1.27E-04
BICF2G630194324	26	39882788	18.46203	1.73E-05	2.19E-04
BICF2P144600	29	15437619	18.17722	2.01E-05	2.45E-04

BICF2G63035789	5	32972070	17.99923	2.21E-05	2.63E-04
BICF2S23649439	8	21298642	17.96072	2.26E-05	2.67E-04
BICF2P692633	5	34675781	17.5517	2.80E-05	3.14E-04
inflation	**1.351601**				





Mixed model:

	Chromosome	Position	chi2.1df	P1df	Pc1df
TIGRP2P110467	8	27113849	21.50683	3.53E-06	3.51E-06
BICF2P713861	8	26769490	21.39995	3.73E-06	3.71E-06
BICF2S232422	5	32983613	19.05455	1.27E-05	1.27E-05
BICF2P752103	8	26675235	16.70483	4.37E-05	4.35E-05
BICF2P1044496	8	27972553	15.17054	9.82E-05	9.79E-05
BICF2G63035030	5	32197936	14.62165	1.31E-04	1.31E-04
BICF2G630558054	7	38778586	13.99363	1.83E-04	1.83E-04
BICF2S2308001	12	48682591	13.67068	2.18E-04	2.17E-04
BICF2G63031325	5	28659812	13.62211	2.24E-04	2.23E-04
TIGRP2P109371	8	23439519	13.39441	2.52E-04	2.52E-04
Inflation:		**0.9996414**			

Mixed model and structured association:

For 3 clusters:	Chromosome	Position	chi2.1df	P1df	Pc1df
BICF2P713861	8	26769490	22.24331	2.40E-06	3.09E-06
TIGRP2P110467	8	27113849	21.90407	2.87E-06	3.68E-06
BICF2S232422	5	32983613	18.15447	2.04E-05	2.51E-05
BICF2P752103	8	26675235	18.03192	2.17E-05	2.67E-05
BICF2P1044496	8	27972553	15.94452	6.52E-05	7.84E-05
BICF2G63031325	5	28659812	14.41084	1.47E-04	1.74E-04
BICF2G63035030	5	32197936	13.75645	2.08E-04	2.44E-04
BICF2S2308001	12	48682591	13.27211	2.69E-04	3.14E-04
TIGRP2P110728	8	27933541	13.21058	2.78E-04	3.25E-04

BICF2P1229493	8	28034747	13.21058	2.78E-04	3.25E-04
TIGRP2P110779	8	28046595	13.21058	2.78E-04	3.25E-04
Inflation:	**1.022345**				

For 5 clusters:	**Chromosome**	**Position**	**chi2.1df**	**P1df**	**Pc1df**
BICF2P713861	8	26769490	20.7553	5.22E-06	7.00E-06
TIGRP2P110467	8	27113849	20.4185	6.22E-06	8.31E-06
BICF2P752103	8	26675235	19.14179	1.21E-05	1.59E-05
BICF2S232422	5	32983613	16.86516	4.01E-05	5.11E-05
BICF2G630194640	26	39699790	16.28962	5.44E-05	6.86E-05
BICF2P638018	8	23630740	16.06417	6.12E-05	7.71E-05
BICF2P1065138	8	23650115	16.06417	6.12E-05	7.71E-05
TIGRP2P109371	8	23439519	15.64517	7.64E-05	9.56E-05
BICF2P1391407	8	22082613	15.22084	9.56E-05	1.19E-04
BICF2S23649439	8	21298642	14.66008	1.29E-04	1.59E-04
BICF2P1044496	8	27972553	14.55826	1.36E-04	1.68E-04
inflation	**1.027856**				

Multiple Alignment Sequence:

Yellow coloured area is the one where difference can be seen between case (CCR48) and control (CCR08).

```
CLUSTAL 2.1 multiple sequence alignment


dog_ensembl       LTQPFISTFQISKYNLLDDPVVTVHQSIGEAKEQFYYERTVFLRCVANSNPPVRYSWRRG 60
CCR48_frame1      LTQPFISTFQISKYNLLDDPVVTVHQSIGEAKEQFYYERTVFLRCVANSNPPVRYSWRRG 60
CCR08_frame1      LTQPFISTFQISKYNLLDDPVVTVHQSIGEAKEQFYYERTVFLRCVANSNPPVRYSWRRG 60
                  ************************************************************

dog_ensembl       QEVLLQGSDKGVEIYEPFFTQGETKILKLKNLRPQDYANYSCIASVRNVCNIPDKMVSFR 120
CCR48_frame1      QEVLLQGSDKGVEIYEPFFTQGETKILKLKNLRPQDYANYSCIASVRNVCNIPDKMVSFR 120
CCR08_frame1      QEVLLQGSDKGVEIYEPFFTQGETKILKLKNLRPQDYANYSCIASVRNVCNIPDKMVSFR 120
                  ************************************************************

dog_ensembl       LSNKTASPSIKLLVDDPIVVNPGEAITLVCVTTGGEPAPTLTWVRSFGTLPEKTVLNGGT 180
CCR48_frame1      LSNKTASPSIKLLVDDPIVVNPGEAITLVCVTTGGEPAPTLTWVRSFGTLPEKTVLNGGT 180
CCR08_frame1      LSNKTASPSIKLLVDDPIVVNPGEAITLVCVTTGGEPAPTLTWVRSFGTLPEKTVLNGGT 180
                  ************************************************************

dog_ensembl       LTIPAITSEDAGTYSCIANNNVGNPAKKSTNIIVRALKKGRFWITPDPYHKDDNIQIGRE 240
CCR48_frame1      LTIPAITSEDAGTYSCIANNNVGNPAKKSTNIIVRALKKGRFWITPDPYHKDDNIQIGRE 240
CCR08_frame1      LTIPAITSEDAGTYSCIANNNVGNPAKKSTNIIVRALKKGRFWITPDPYHKDDNIQIGRE 240
                  ************************************************************

dog_ensembl       VKISCQVEAVPSEELTFSWFKNGRPLRSSERMVITQTDPDVSPGTTNLDIIDLKFTDFGT 300
CCR48_frame1      VKISCQVEAVPSEELTFSWFKNGRPLRSSERMVITQTDPDVSPGTTNLDIIDLKFTDFGT 300
CCR08_frame1      VKISCQVEAVPSEELTFSWFKNGRPLRSSERMVITQTDPDVSPGTTNLDIIDLKFTDFGT 300
                  ************************************************************
```

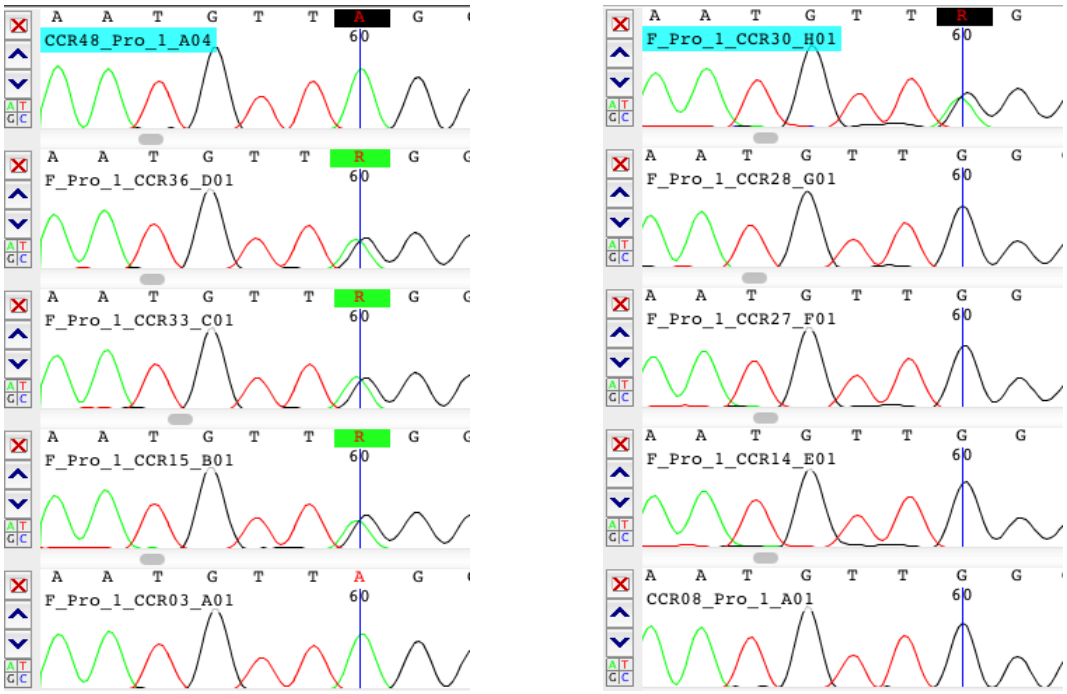
```
dog_ensembl       YTCVASLKGGGISDISIDVNISSSTVQVPPNLTVPQEKSPLVTREGDTIELQCQVTGKPK 360
CCR48_frame1      YTCVASLKGGGISDISIDVNISSSTVQVPPNLTVPQEKSPLVTREGDTIELQCQVTGKPK 360
CCR08_frame1      YTCVASLKGGGISDISIDVNISSSTVQVPPNLTVPQEKSPLVTREGDSIELQCQVTGKPK 360
                  ************************************************.************

dog_ensembl       PIILWSRADKEVAMPDGSMQMESYDGTLRIVNVSREMSGMYRCQTSQYNGFNVKPREALV 420
CCR48_frame1      PIILWSRADKEVAMPDGSMQMESYDGTLRIVNVSREMSGMYRCQTSQYNGFNVKPREALV 420
CCR08_frame1      PIILWSRADKEVAMPDGSMQMESYDGTLRIVNVSREMSGMYRCQTSQYNGFNVKPREALV 420
                  ************************************************************

dog_ensembl       QLIVQYPPAVEPAFLEIRQGQDRSVTMSCRVLRAYPIRVLTYEWRLGNKLLRTGQFDSQE 480
CCR48_frame1      QLIVQYPPAVEPAFLEIRQGQDRSVTMSCRVLRAYPIRVLTYEWRLGNKLLRTGQFDSQE 480
CCR08_frame1      QLIVQYPPAVEPAFLEIRQGQDRSVTMSCRVLRAYPIRVLTYEWRLGNKLLRTGQFDSQE 480
                  ************************************************************

dog_ensembl       YTEYPVKSLSNENYGVYNCSIINEAGAGRCSFLVTGKAYAPEFYYDTYNPVWQNRHRVYS 540
CCR48_frame1      YTEYPVKSLSNENYGVYNCSIINEAGAGRCSFLVTGKAYAPEFYYDTYNPVWQNRHRVYS 540
CCR08_frame1      YTEYPVKSLSNENYGVYNCSIINEAGAGRCSFLVTGKAYAPEFYYDTYNPVWQNRHRVYS 540
                  ************************************************************

dog_ensembl       YSLQWTQMNPDAVDRIVAYRLGIRQAGQQRWWEQEIKINGNIQKGELITYNLTELIKPEA 600
CCR48_frame1      YSLQWTQMNPDAVDRIVAYRLGIRQAGQQRWWEQEIKINGNIQKGELITYNLTELIKPEA 600
CCR08_frame1      YSLQWTQMNPDAVDRIVAYRLGIRQAGQQRWWEQEIKINGNIQKGELITYNLTELIKPEA 600
                  ************************************************************

dog_ensembl       YEVRLTPLTKFGEGDSTIRVIKYSAPVNPHLREFHCGFEDGNICLFTQDDTDNFDWTKQS 660
CCR48_frame1      YEVRLTPLTKFGEGDSTIRVIKYSAPVNPHLREFHCGFEDGNICLFTQDDTDNFDWTKQS 660
CCR08_frame1      YEVRLTPLTKFGEGDSTIRVIKYSAPVNPHLREFHCGFEDGNICLFTQDDTDNFDWTKQS 660
                  ************************************************************

dog_ensembl       TATRNTKYTPNTGPNADRSGSKEGFYMYIETSRPRLEGEKARLLSPVFSIAPKNPYGPTN 720
CCR48_frame1      TATRNTKYTPNTGPNADRSGSKEGFYMYIETSRPRLEGEKARLLSPVFSIAPKNPYGPTN 720
CCR08_frame1      TATRNTKYTPNTGPNADRSGSKEGFYMYIETSRPRLEGEKARLLSPVFSIAPKNPYGPTN 720
                  ************************************************************

dog_ensembl       TAYCFSFFYHMYGQHIGVLNVYLRLKGQTTIENPLWSSSGNKGQRWNEAHVNIYPITSFQ 780
CCR48_frame1      TAYCFSFFYHMYGQHIGVLNVYLRLKGQTTIENPLWSSSGNKGQRWNEAHVNIYPITSFQ 780
CCR08_frame1      TAYCFSFFYHMYGQHIGVLNVYLRLKGQTTIENPLWSSSGNKGQRWNEAHVNIYPITSFQ 780
                  ************************************************************

dog_ensembl       LIFEGIRGPGIEGDIAIDDVSIAEGECAKQDLTTKNSVDGAVGILVHLWLFPVIVLISIL 840
CCR48_frame1      LIFEGIRGPGIEGDIAIDDVSIAEGECAKQDLTTKNSVDGAVGILVHLWLFPVIVLISIL 840
CCR08_frame1      LIFEGIRGPGIEGDIAIDDVSIAEGECAKQDLTTKNSVDGAVGILVHLWLFPVIVLISIL 840
                  ************************************************************

dog_ensembl       SPRR- 844
CCR48_frame1      SPRR- 844
CCR08_frame1      SPRR- 844
                  ****
```

Chromatogram of Sequence Analysis of MDGA2:

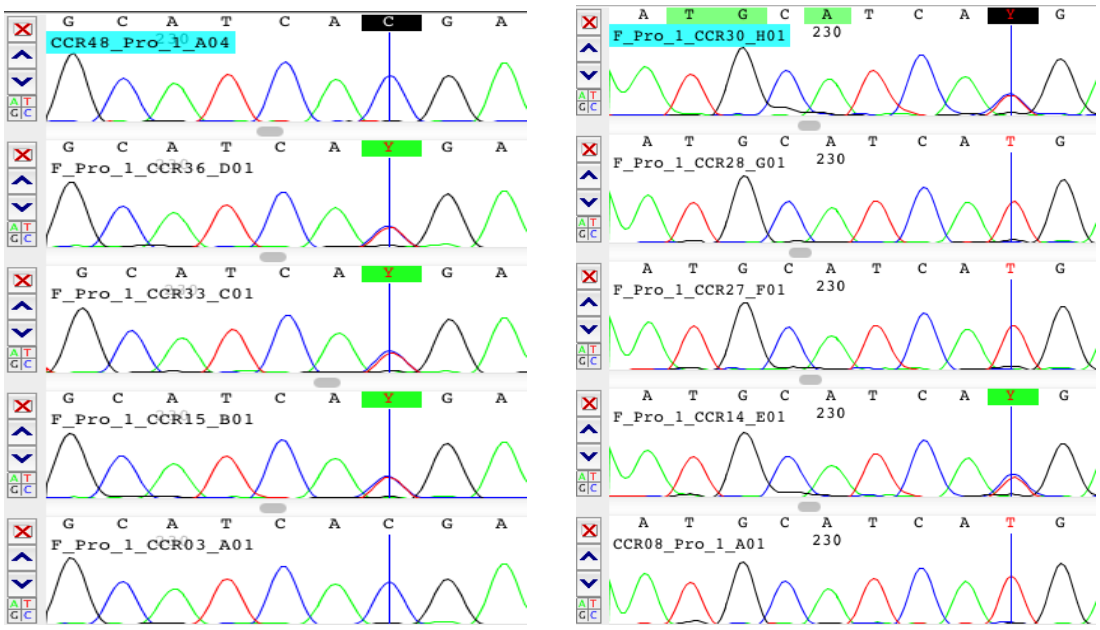
The difference in nucleotide between a case and a control can be observed in the chromatogram analysed through Codon code aligner (4.0.2). The vertical blue line is passed along the peak that is being analysed.

Promoter:

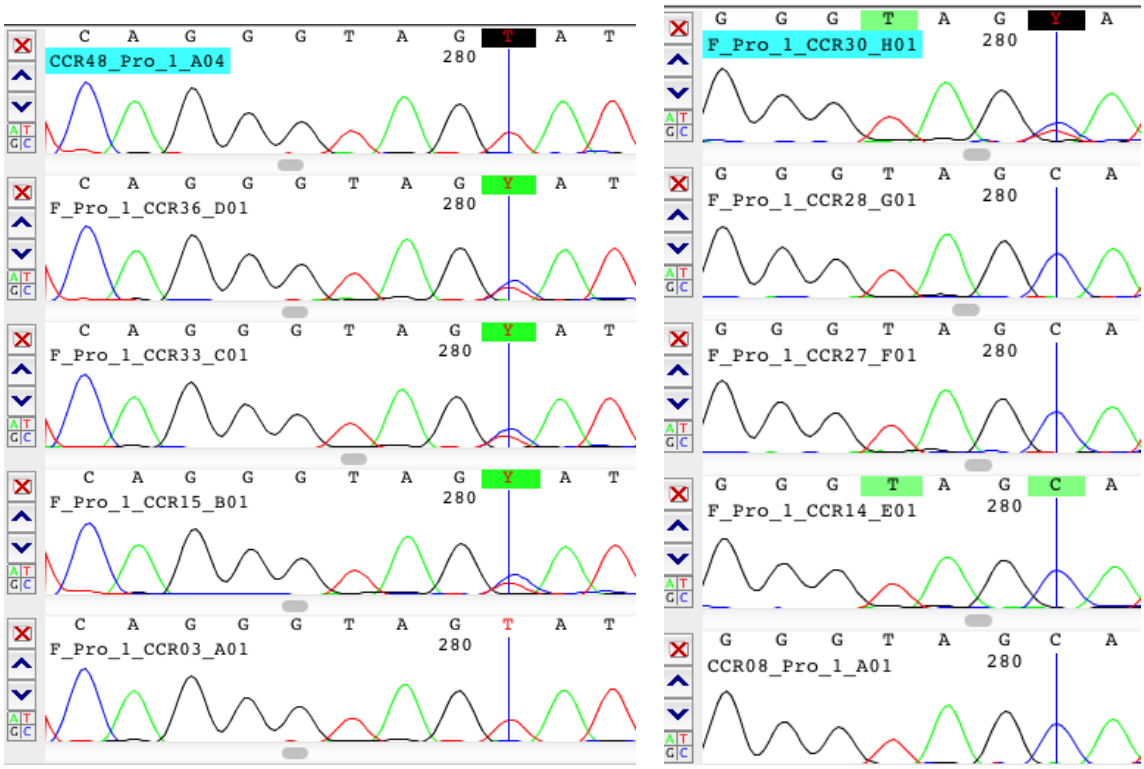
Position 60: The blue line below denotes the region of difference between cases and control. The cases are present on the left side and controls on the right side. Two of the cases (CCR 03 and CCR 48) are homozygous "A" for position 60 where as other cases are heterozygous "R: G/A". All controls are homozygous "G" except for CCR30 which is heterozygous "R".



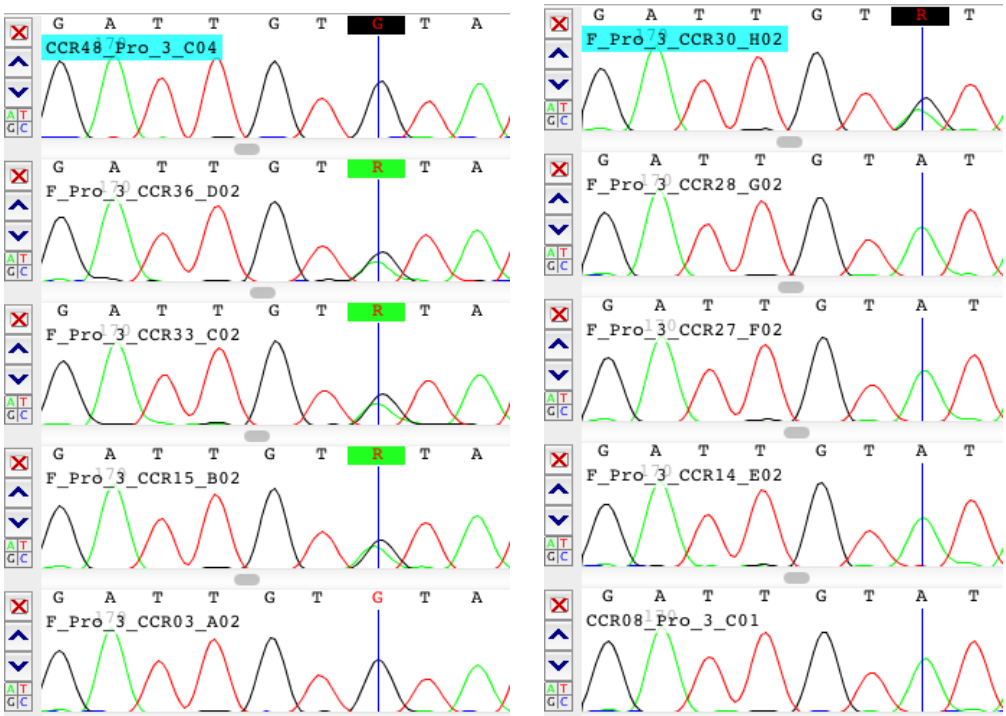
Position 234: The cases on left, CCR48 and 3 are homozygous "C" whereas all other cases are heterozygous "Y". The controls on right, CCR 30 and 14 are heterozygous "Y" whereas others are homozygous "T".



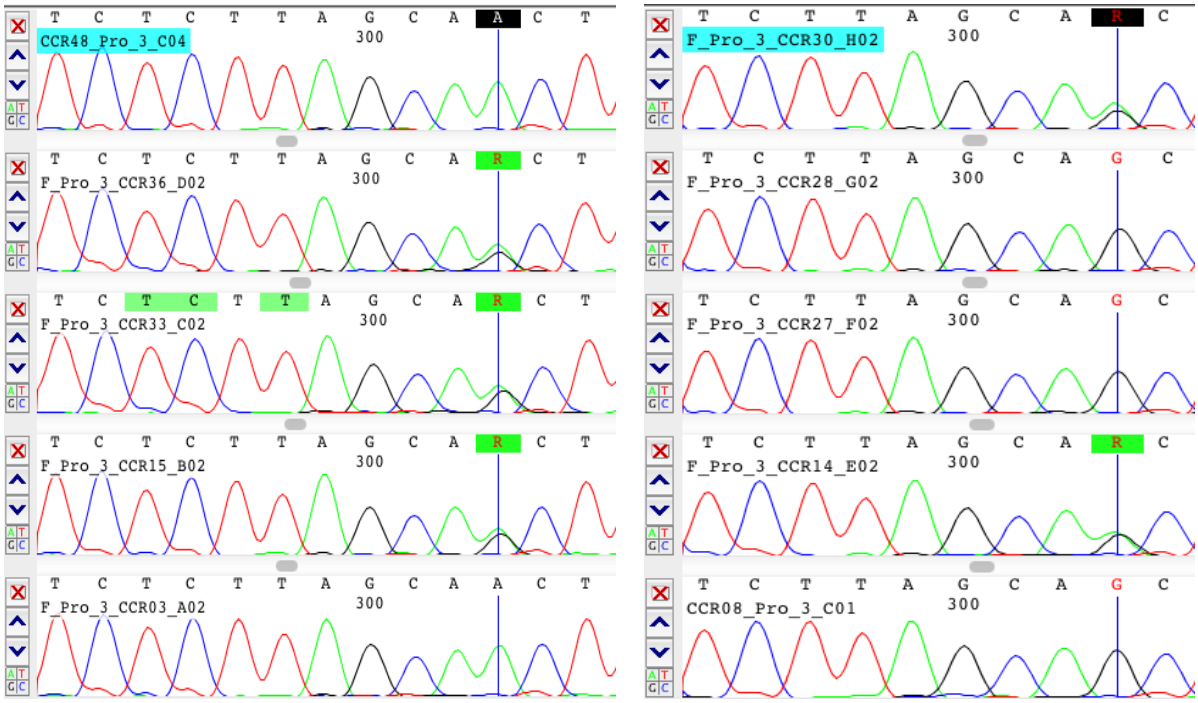
Position 281: The cases on left, CCR48 and 03 are homozygous "T" and other are heterozygous "Y". The controls on right, all controls are homozygous "C" except for CCR30 that is heterozygous "Y".



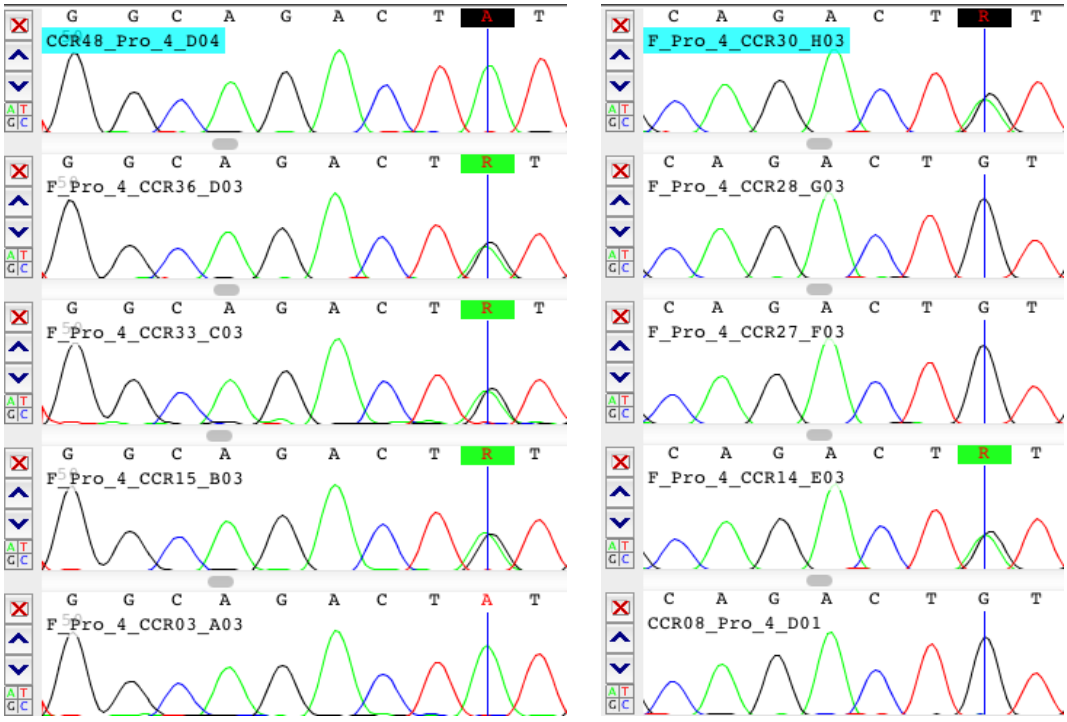
Position 895: The cases on left, CCR48 and 03 are homozygous "G" whereas others are heterozygous "R". The controls on right, all the controls are homozygous "A" except for CCR 30 that is heterozygous "R".



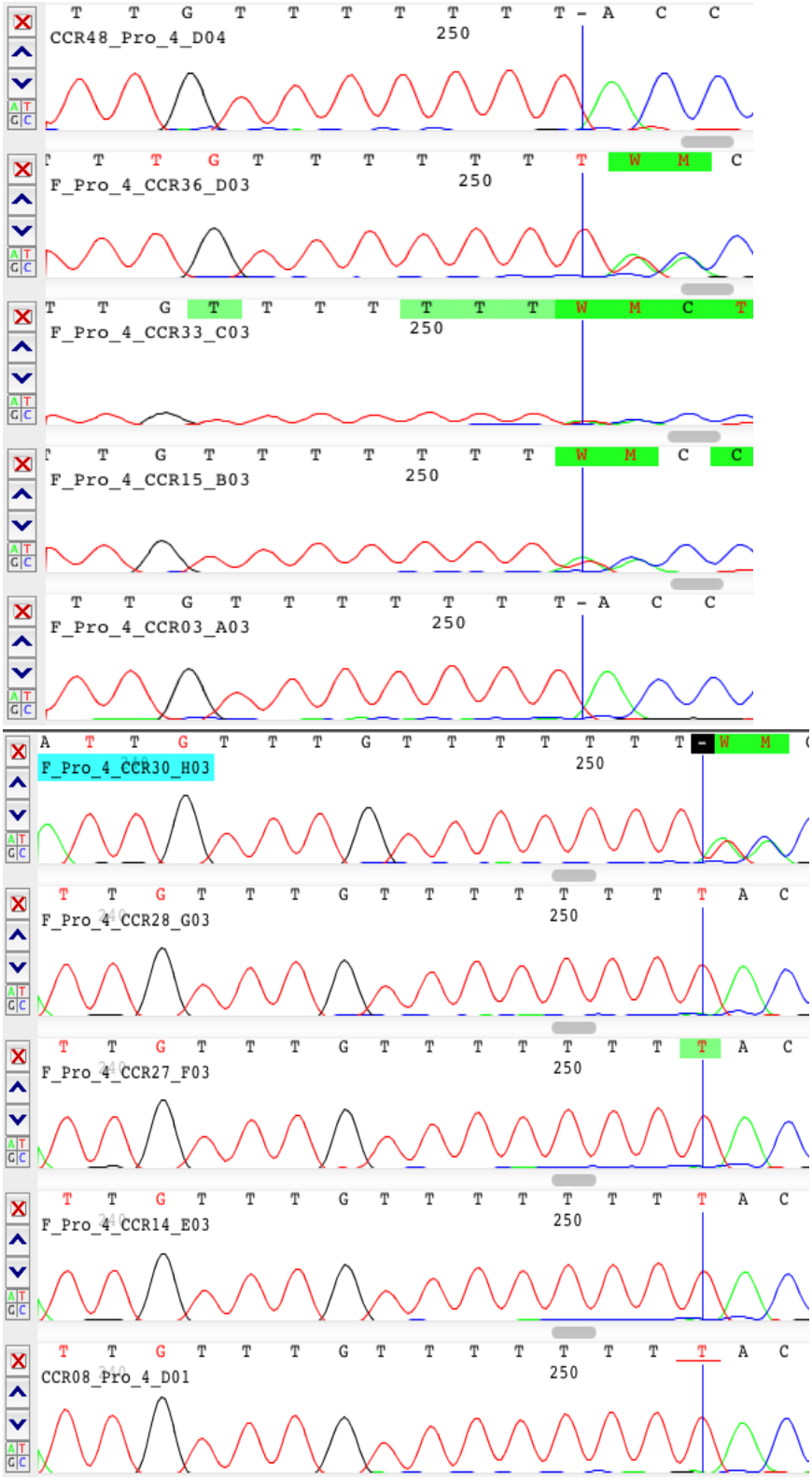
Position 1023: The cases on left, CCR 48 and 03 are homozygous "A" and other cases are heterozygous "R". The controls on right, controls CCR 30 and 14 are heterozygous "R" where as other controls are homozygous "G".



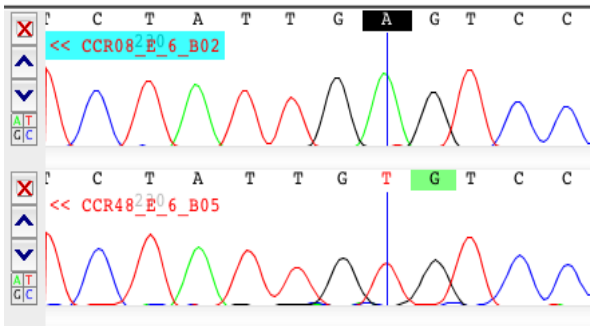
Position 1138: The cases on left, CCR48 and 03 are homozygous "A" and other cases are heterozygous "R". The controls on right, controls CCR 30 and CCR 14 are heterozygous "R" and other controls are homozygous "G".



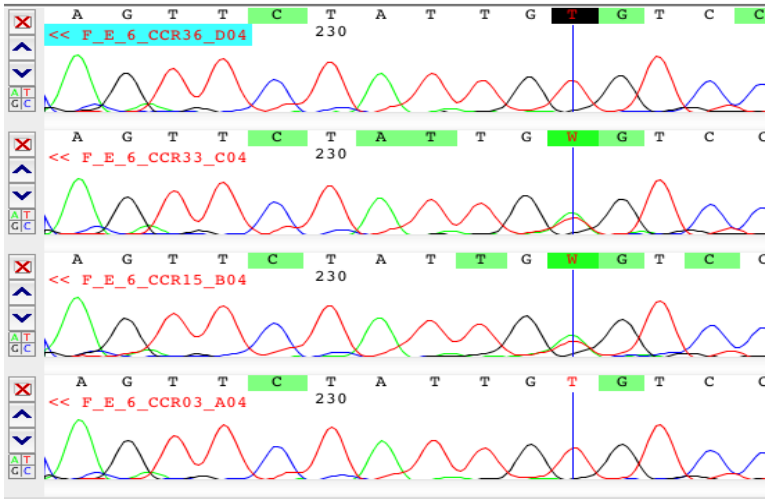
Position 1333: The controls on the bottom, all the controls have an insertion "T" except for CCR30. In cases on the top, CCR15, CCR 33 and CCR36 have insertion. CCR 15 and CCR33 are heterozygous "W: A/T" whereas CCR36 is homozygous "T" for insertion. CCR03 and CCR 48 don't have any insertion in that position.



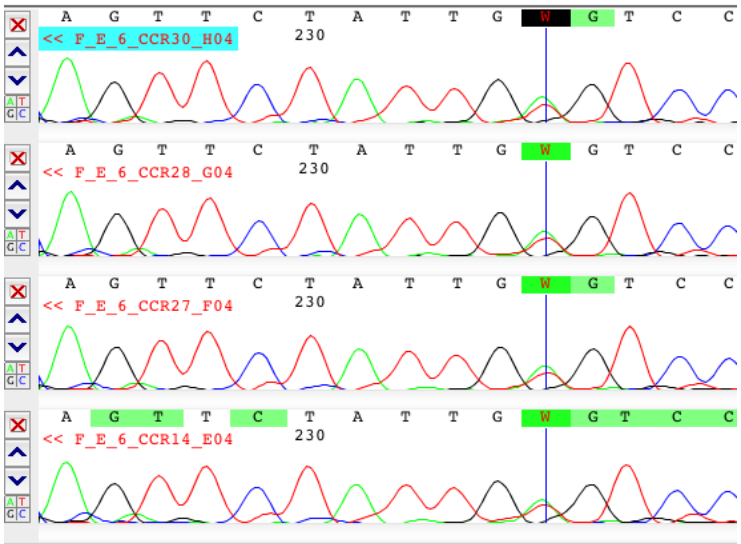
Exon 6:



In above chromatogram, the position that differed is marked by a blue line that runs through the peaks. A case (CCR48) is homozygous for an allele "T" where as the control "CCR08" is homozygous for an allele "A".

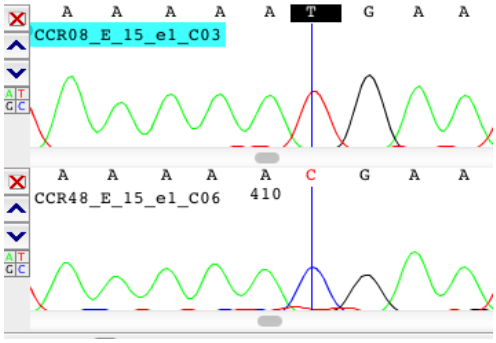


In above chromatogram, Exon 6 sequences of four cases CCR-36, 33, 15 and 3 are analysed. Two of the cases 36 and 3 are homozygous "T" for that position whereas other cases are heterozygous "W: A/T".

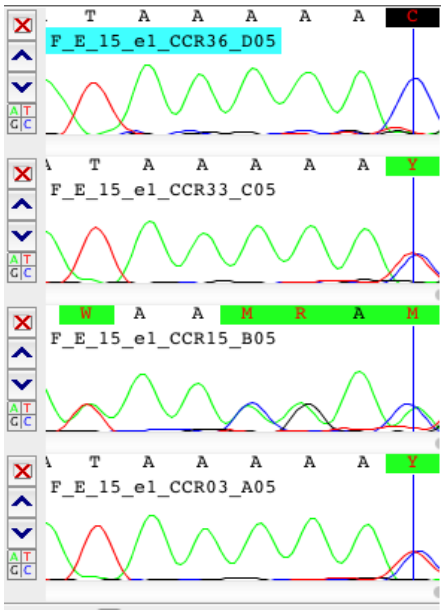


In above chromatogram, Exon 6 sequences of four controls CCR-33, 28, 27 and 14 are analysed. All of the controls are heterozygous “W: A/T” for that position.

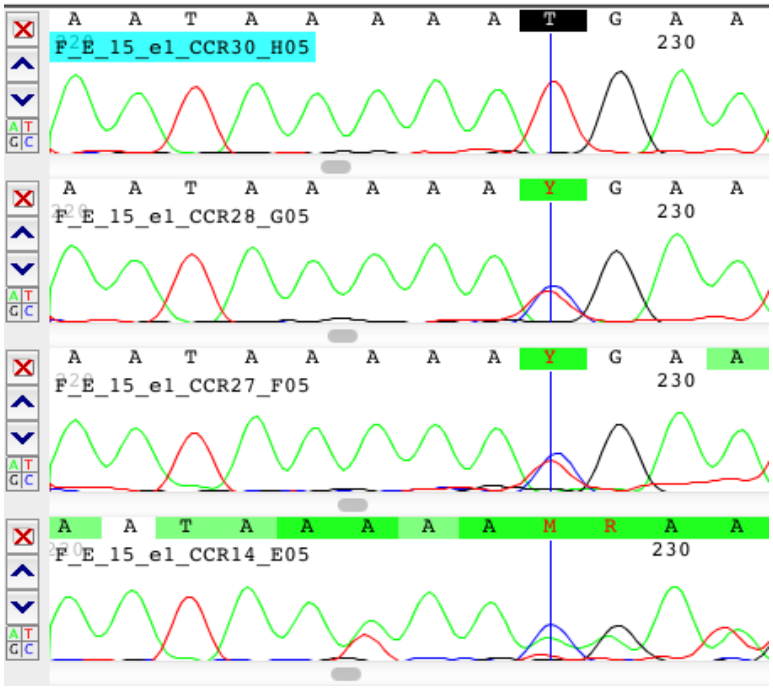
Exon 15:



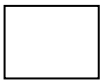
For Exon 15, in above chromatogram a case is homozygous for allele “C” and the control is homozygous for an allele “T”.



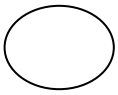
In above chromatogram, Exon 15 sequences of four cases CCR-36, 33, 15 and 3 are analysed. Two of the cases 33 and 3 are heterozygous “Y: T/C” for that position, other case CCR33 is heterozygous “M: A/C” and CCR36 is homozygous “C”.



In above chromatogram, Exon 15 sequences of four controls CCR-36, 33, 15 and 3 are analysed. Two of the cases 28 and 27 are heterozygous “Y: T/C” for that position, other case CCR14 is heterozygous “M: A/C” and CCR30 is homozygous “T”.



MALE : NORMAL

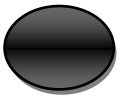


FEMALE : NORMAL

MALE : UNKNOWN VETERINARY DATA

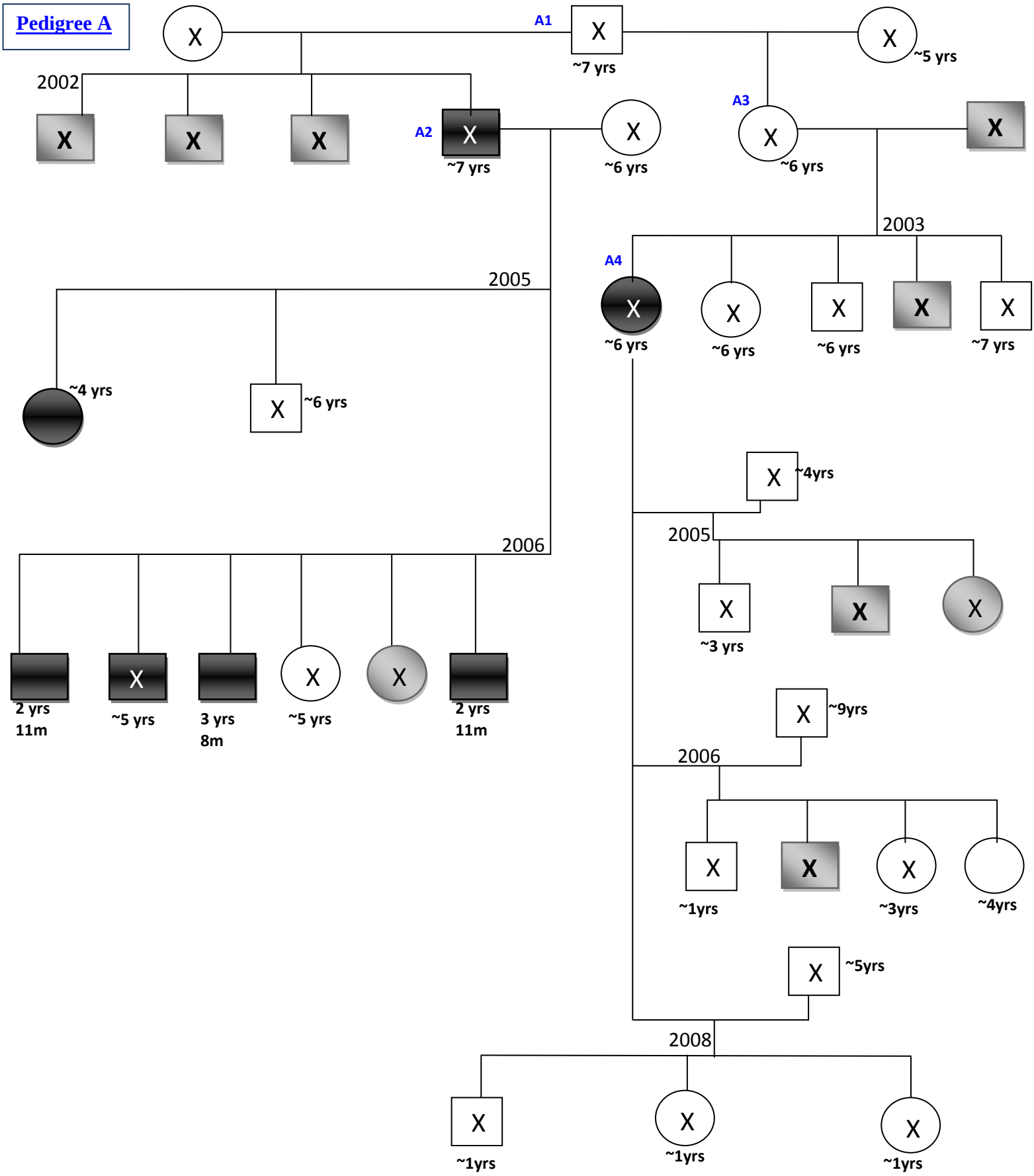
FEMALE : UNKNOWN VETERINARY DATA

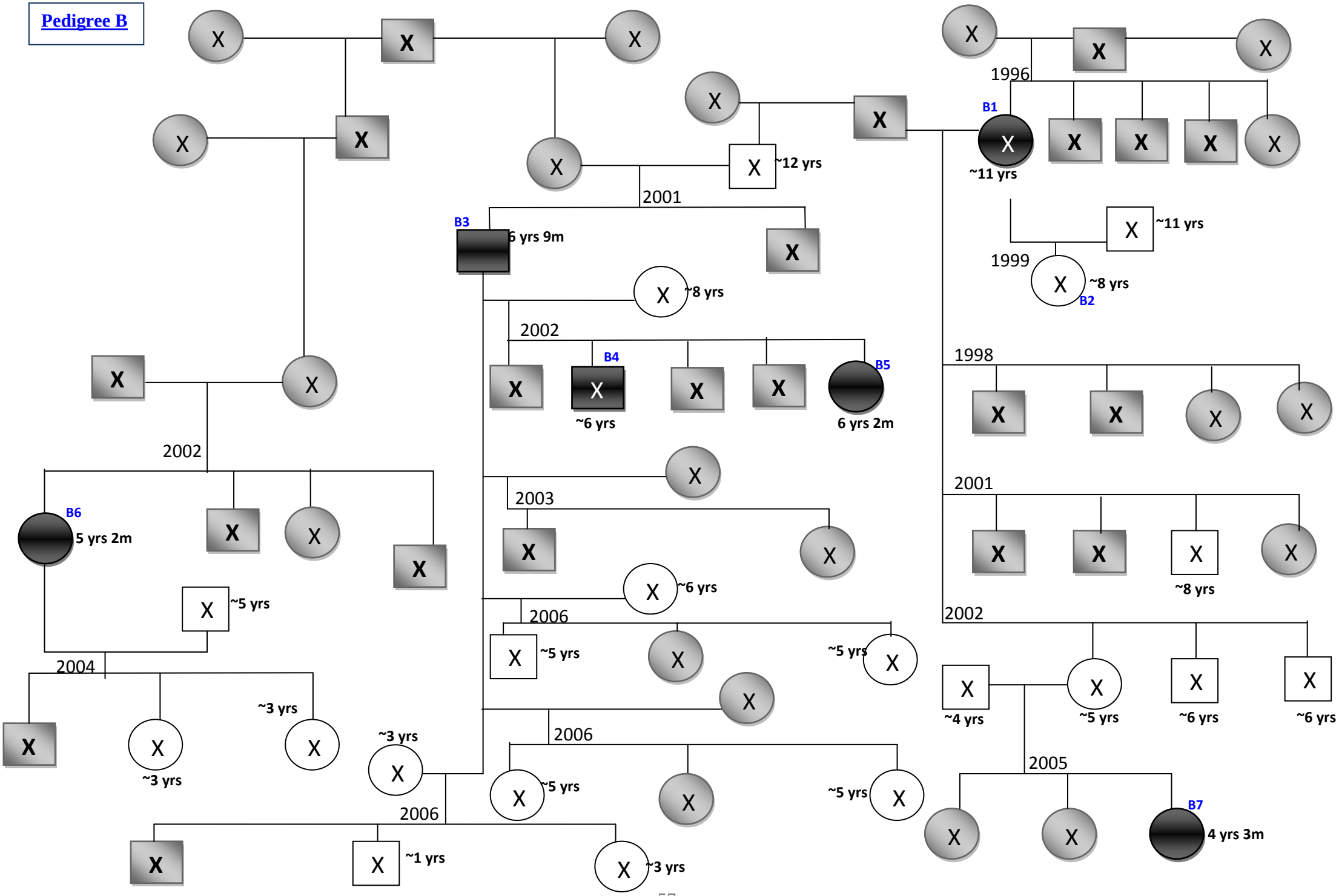
MALE : RETINOPATHY



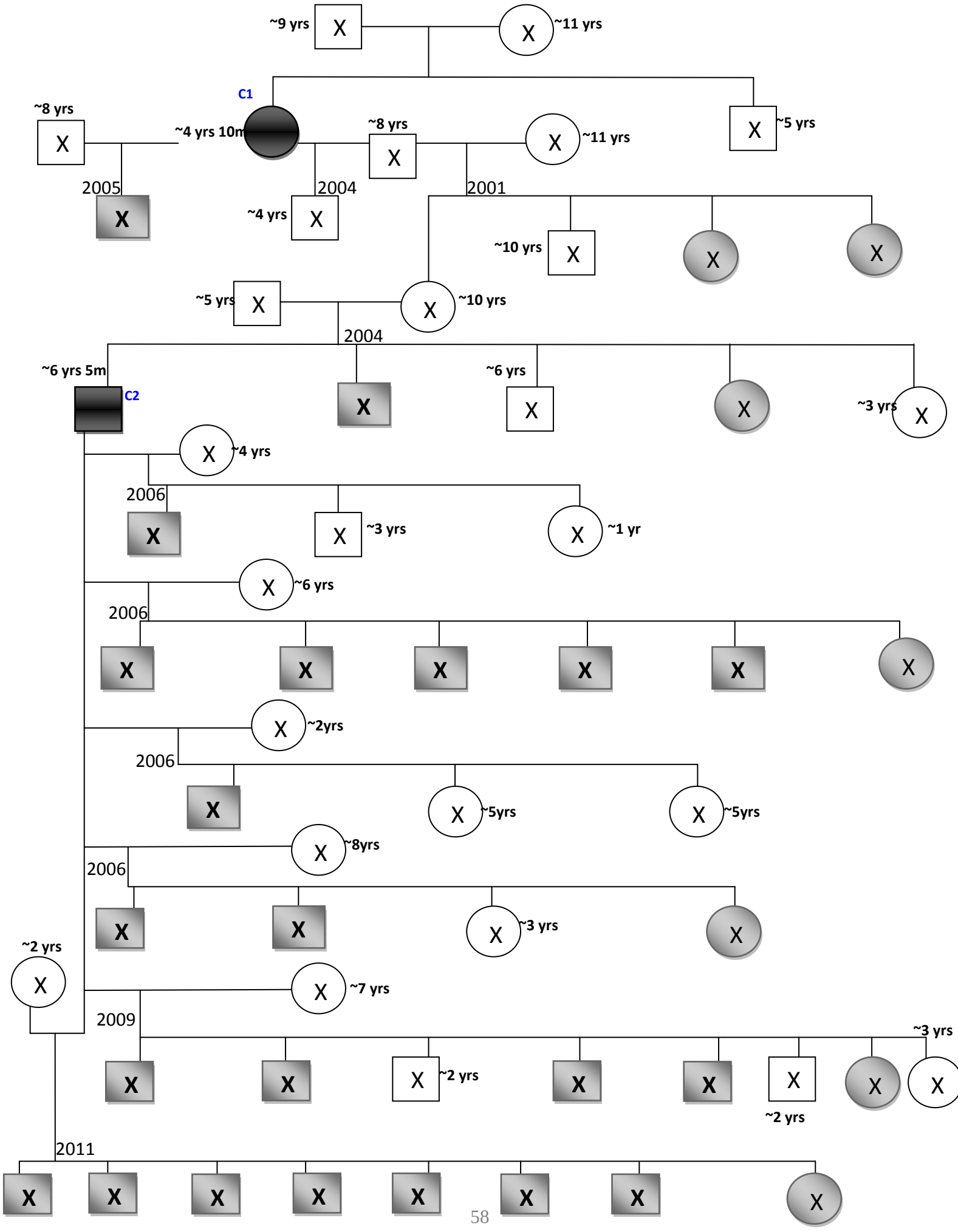
FEMALE : RETINOPATHY

X: INDIVIDUALS MARKED WITH 'X' DENOTES THAT WE DON'T HAVE THEM IN OUR SAMPLE LIST.

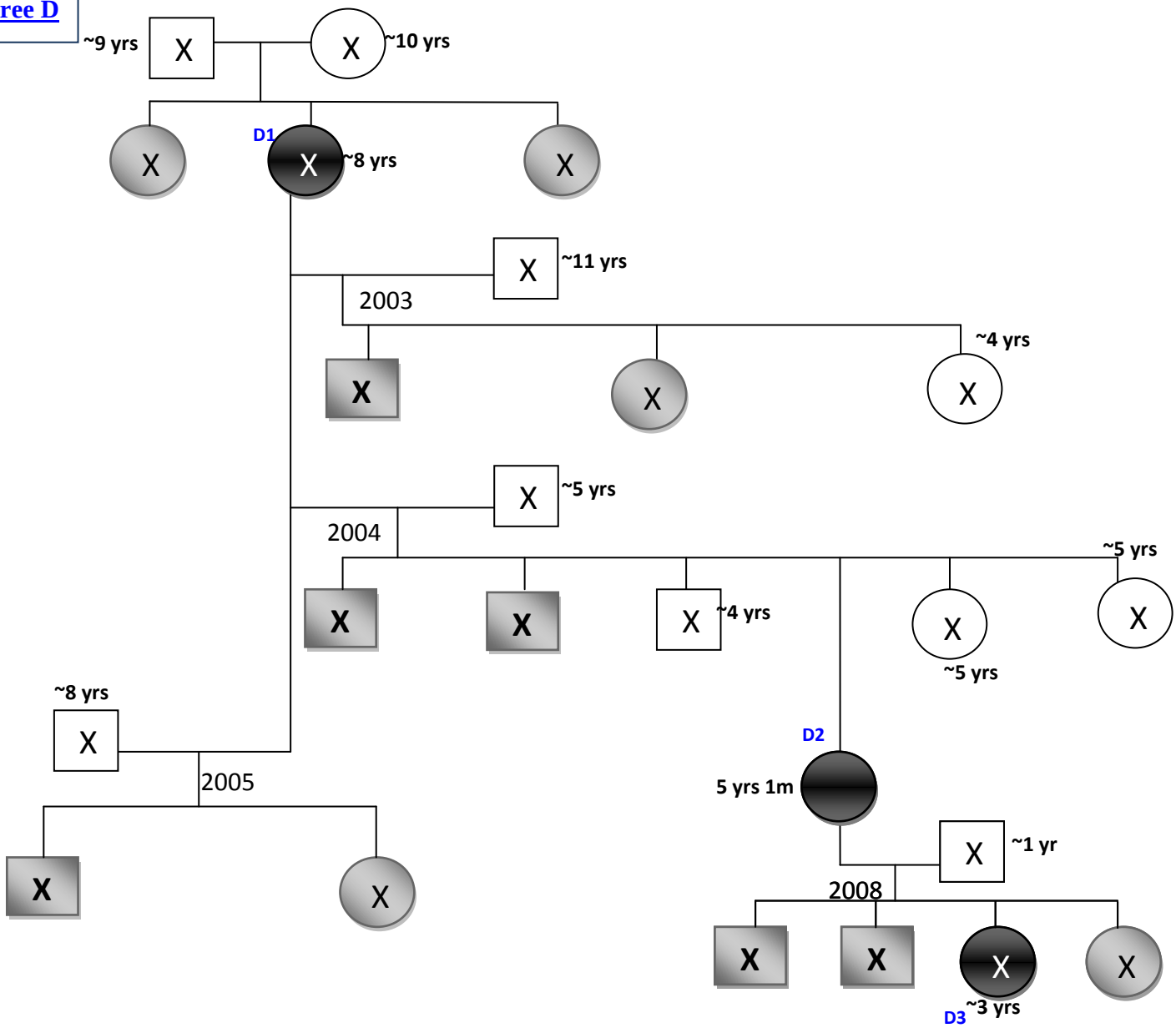

Pedigree A
A1
~7 yrs
~5 yrs
2002
A2
~7 yrs
~6 yrs
A3
~6 yrs
2003
A4
~6 yrs
~6 yrs
~6 yrs
~7 yrs
2005
~4 yrs
~6 yrs
~4yrs
2005
~3 yrs
2006
2 yrs 11m
~5 yrs
3 yrs 8m
~5 yrs
2 yrs 11m
~9yrs
2006
~1yrs
~3yrs
~4yrs
~5yrs
2008
~1yrs
~1yrs
~1yrs


Pedigree B
1996
B1
~11 yrs
~11 yrs
1999
~8 yrs
B2
2001
B3
6 yrs 9m
~12 yrs
~8 yrs
2002
B4
~6 yrs
B5
6 yrs 2m
1998
~8 yrs
2002
B6
5 yrs 2m
~5 yrs
2004
~3 yrs
~3 yrs
2003
~6 yrs
2006
~5 yrs
~5 yrs
~3 yrs
2006
~5 yrs
~5 yrs
2006
~1 yrs
~3 yrs
2001
~4 yrs
~5 yrs
~6 yrs
~6 yrs
2002
2005
B7
4 yrs 3m

Pedigree C


~9 yrs
~11 yrs
C1
~8 yrs
~4 yrs 10m
~8 yrs
~11 yrs
~5 yrs
2005
2004
~4 yrs
2001
~10 yrs
~5 yrs
~10 yrs
2004
~6 yrs 5m
C2
~6 yrs
~3 yrs
~4 yrs
2006
~3 yrs
~1 yr
~6 yrs
2006
~2yrs
2006
~5yrs
~5yrs
~8yrs
2006
~3 yrs
~2 yrs
~7 yrs
2009
~3 yrs
~2 yrs
~2 yrs
2011

Pedigree D


~9 yrs
~10 yrs
D1
~8 yrs
~11 yrs
2003
~4 yrs
~5 yrs
2004
~5 yrs
~4 yrs
~5 yrs
~8 yrs
2005
D2
5 yrs 1m
~1 yr
2008
~3 yrs
D3