Molecular Genetic Studies of Inherited Cataracts in the American Cocker Spaniel

- progress report

University of Pennsylvania, July 17, 2019

Personnel: ***University of Pennsylvania***
Gustavo Aguirre, VMD, PhD; Principal Investigator
Leonardo Murgiano, PhD; Post-doctoral fellow and Project Scientist
Dina Torjman M.Sc., Research scientist

(Doreen Becker, DVM, PhD; former Post-doctoral fellow and Project Scientist)
(Jessica Niggel, M.Sc., former Research scientist)

Objectives

The principal objective of the project is the identification of the gene(s) and genetic variants responsible for inherited cataract in American Cocker Spaniels (ACS). Our final aim consists in the development of a genetic test that can identify genetically normal, affected and carrier dogs for any variant considered deleterious or somewhat implicated with the development of cataracts. In this report we will describe the progress achieved, the last important steps implemented, and the future prospects.

Cataracts are the most common cause of vison impairment in humans and other mammals, and are very frequent ophthalmic diseases in dogs. Several breeds are affected by such condition, included the American Cocker Spaniel (ACS), with an estimated prevalence of 8-11%. Such percentages describe both acquired and inherited cataracts: the latter category contains cataract phenotypes that are clinically similar, but may have a different genetic etiology and only a superficial clinical similarity. ACS dogs with inherited cataracts are born with normal lenses, which then proceed to opacify over time, leading to blindness by 2-10 years of age.

The mechanism of inheritance in ACS has been previously proposed as being be autosomal recessive, but our subsequent observations suggested a situation more complex than the one predicted in the preliminary phase of the project.

Thanks to the outstanding contribution by the participating breeders and owners our work significantly progressed after a thorough classification of suitable and verified samples in the ACS population which included a constant re-analysis and update of the cases and controls present in our database. This allowed us to pinpoint specific areas of the genome associated in varying degree with the condition, and to refine such association with each iteration of the analysis (included the exclusion of false positives).

Our final aim remains the identification of gene(s) and vulnerability loci associated with the most common form of cataract in ACS and on validating its inheritance mechanism. We achieved such analysis of the database through tight communications with the owners and the breeding association. After obtaining a sufficient number of samples, we finally implemented an important step in the research.

Background

Cataract in ACS – nature of the samples

Cataracts are often inherited conditions. They are characterized by opacity/cloudiness of the lens, arising due to lens protein misfolding, solubility changes and aggregation and leading to vision impairment of progressive severity, occasionally demanding surgical intervention. American Cocker Spaniels are among the most commonly cataract-affected dog breeds. As previously reported, we have found a spectrum of cataract phenotypes differing in location, progression rate, laterality (unilateral or bilateral), genetic background and age of onset. We considered the latter parameter, above the rest, as the most crucial for the classification and grouping of our samples.

We stressed for a correct gathering of information about the affected and unaffected dogs and for a precise assessment of the phenotype and the selection of a good control sample group. This is essential in order to select candidate cases for cataracts predictable as having a genetic etiology. Specifically, inherited cataracts in ACS are thought to appear sometime around 2-5 years of age and progress. Nonetheless, we have found a subset of cases where cataracts, presumably inherited, begin between 5-9 years of age. Both of these age groupings have been included in the study, and analysis is done for the entire population or for each subgroup individually.

Cataracts can be caused environmental effects such as UV light exposure, mechanical trauma, poor nutrition, exposure to toxic substances. They can also occur as secondary effects of other ophthalmic diseases, such as uveitis or glaucoma. We used the maximum care in excluding any possible secondary cataract phenotype with a high likelihood of not having a genetic etiology, and thus lowering the quality the dataset.

Research on genetic diseases in companion animals

Current research in genetic diseases in domestic animals is based on three main principles: (I) Construction of a suitable dataset, obtained through the identification of cases and validated controls (II) Mapping of the variants associated with the condition studied (III) Validation through sequencing.

The importance of (I) is described and explained in the above paragraph. A number of significant steps forward have been made thanks to this approach, and below we elaborate on specific sub-phenotypes detected.

(II) is generally achieved through the use of SNP genotyping. The method uses purified DNA from blood of cases and controls that is placed on chips, specific platforms scanned for strategically selected genetic variation markers, called single nucleotide polymorphisms (SNPs). Thorough the information obtained by such experiments, the researchers can explore the presence of common (and ideally, exclusive) shared regions among the cases. Such region could be, as an example, common homozygous intervals (as it happens in recessive diseases). Analysis of markers inherited from parents and identical by descent can even pinpoint shared linked interval in heterozygous regions of the chromosome (as in dominant diseases).

Another common type of analysis is the Genome Wide Association Study (GWAS). Such study pinpoints higher frequency of certain SNPs in cases vs. controls, associating these variations with the disease. GWAS can be implemented on a wide population of dogs with reasonable computation time, and regardless of the family information about the samples. Moreover, GWAS can better predict variable degrees of association of a locus with the condition, giving away vital information in the investigation of a more complex inheritance mechanism. In fact, GWAS has been a vital part of our approach. Often, the dataset generated for GWAS analysis is also used to search shared homozygous regions among the cases.

Sequencing (III) consists, in general terms, in the determination of the exact DNA sequence of a given genomic region (of variable size, and even including a genome in its entirety). A common and fast sequencing method is the Sanger sequencing, used for the comparison of candidate mutations in cases and controls (that is, to validate whether a given mutation is associated with the condition, thus possibly being the causative one). Sanger is often used even for the development and execution of a genetic test for the disease.

Even though limited use of Sanger sequencing is relatively inexpensive, the exploration of a whole genome sequence would make it unfeasible and too expensive. To this end, Whole Genome Sequencing (WGS) methods have brought a whole new level in the exploration of genetic defects, because they allow the scientists to obtain the full information about the genome of a sequenced animal. WGS is particularly useful when the sequencing of a high amount of candidate variants in one or more cases would be unfeasible for time and money constraints.

An ideal scenario in the study of a genetic defect involves the use of SNP chip for the mapping the disease to a specific chromosomal region, and sequencing a putative candidate gene for the validation of the data once the genomic region is identified. Even in case of more than one associated/implicated region, a careful evaluation of the samples selected for WGS, a consistent dataset and a high number of controls can finally unveil the genetic etiology of the disease.

Summary of the previous work (and progress to date):

We implemented several strategies during the period of the study. As stated previously, the choice of a given approach was done depending on the quality of the dataset available at the time, and the reliability of the information. The constant influx of new samples improved the dataset on each iteration.

Candidate genes and pedigree analysis

As previously reported, while in the ongoing process of collecting sufficient samples needed for detailed genomic studies, we carried out a preliminary candidate gene analysis in order to exclude more obvious genes. As stated, the results were negative – we found no associated variant in those selected genes with the cataract phenotype (for more details about these results, see the previous Progress Reports).

In the previous reports, we described the use of the pedigree software Cyrillic. We were able to link most of our affected subjects to three common ancestors. For this reasons we hypothesized that an autosomal recessive inheritance is at play, and that such model would explain at least a significant part of our cases. Nonetheless, a deeper analysis of the data suggested that a common, shared genetic variants causing *all* the genetic cataracts in the ACS population is unlikely.

Total dogs	**793**
Total of Informative dogs	**534**
Potential cases	***93***
Bilateral	72
Unilateral or very Asymmetric	21
Controls	***441***
Too young to be properly assessed	185
Total of Excluded dogs	**259**

Samples received

Compared to the previous report, the number of dogs participating the study increased to 793 from the 769 reported last time. A short breakdown of the samples follows:

Table 1 –Total of dogs entered in the dataset. Count of dogs that are sufficiently informative, type of cases, potential controls and dogs not suitable for the study. Causes for exclusion: co-morbidity with another eye condition, doubts about diet, the dog prematurely deceased (especially if DNA/blood is missing), lack of feedback on updates (rare occurrence), lack of an official diagnosis by a certified veterinary ophthalmologist (or of monitoring post diagnosis), inconsistent records (very rare occurrence). Of the dogs shown above, only the ones with consistent records over time can be genotyped!

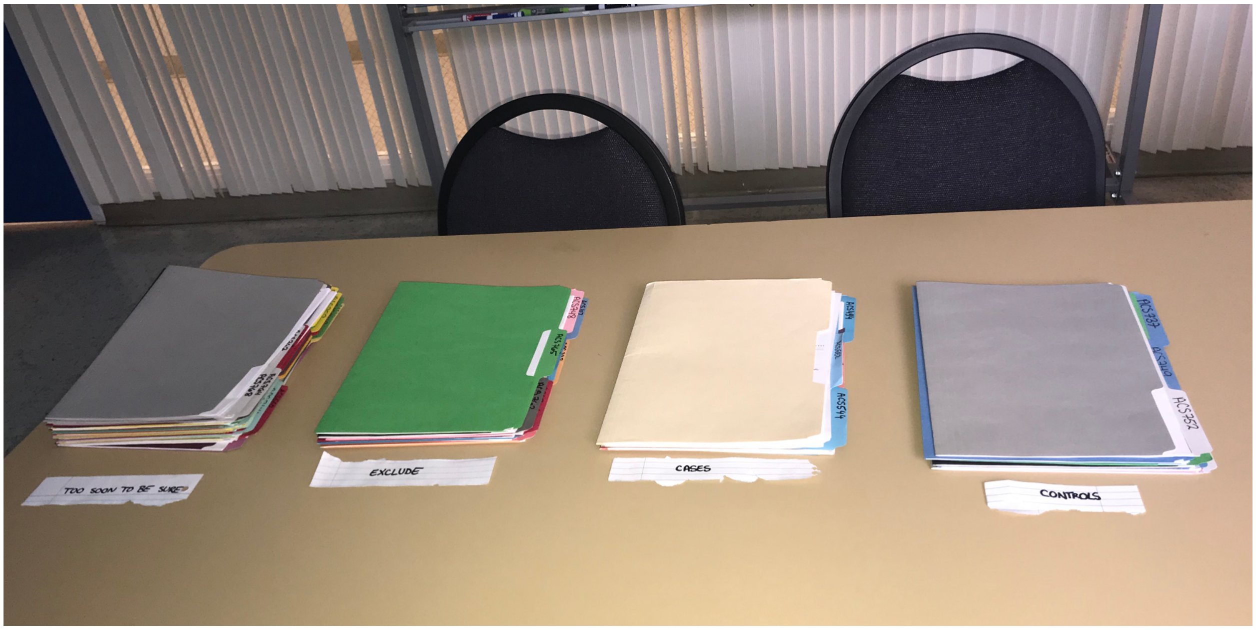
DNA samples were isolated from blood or buccal swabs by personnel at OptiGen LLC who are collaborating in the study (in this regard, we wish to thank the breeders for the fact that the overwhelming majority of samples are blood samples, easier to work with and generally bringing with better DNA yield). All of the blood samples have been sent to us in EDTA tubes to prevent clotting. We extracted the DNA from the blood samples of cases and controls considered suitable for the study.

Phenotype reassessment

We previously reported the development and use of a standardized eye exam research form. We wish to stress again that the forms are extremely useful and important to the study, we have noticed that still not every veterinary ophthalmologist will use them. This has been a problem as the forms used-OFA-CAER-are inadequate for consistent diagnosis. A proper form can be downloaded through the following link:

https://drive.google.com/open?id=1c-hbLI2sdgMyVtb1jz7gSkO9v8AAme5V
Clicking on this link will direct to a page with the document. It can be downloaded (top right) and/or printed. Please note this is an updated version of the link and the form (Jul 2019)

Each time new samples are added and a sufficient number of updates is gathered, we analyze the new information and re-classify the dogs. We make use of our carefully organized archive and classify the samples as Cases, Controls, Excluded (due to the phenotype being probably explained by a non-genetic etiology) and samples simply too young to be evaluated with certainty (therefore the assignment is withhold and are kept under observation).



Typical classification of samples during a phenotype evaluation session.

In addition to the dogs listed in Table 1, a recent round of visits allowed us to include 11 additional controls; i.e. “promoted” based on re-examination results. The contribution of the breeders is continuous, and we are soon going to examine 42 dogs so far deemed too young to tell or with incomplete records.

As previously stated, we have discovered that ACS seem to exhibit distinct sub-types of phenotypes of inherited cataract. Primarily, we registered (I) a possible diversification of the phenotypes in regard of the age of onset. We also (II) noted that there seems to be a second type of classification of the cataract phenotype, where one eye develops a cataract at an early age and several years later a second cataract appears in the other. We also (III) took into account the anteroposterior position of the cataract onset for the classification of the phenotype.

Our principal mean of classification of the phenotypes was on the age basis (I). In fact, since we started to carefully re-assess the phenotypes of the dogs, such element was our primary concern in order to include a sample in the “Cases” or “Controls” groups, and more importantly, asses the quality of the “Case” with a relevant score. Such subdivision is distinct and both groups consist of a high number of samples.

In case of (II) and (III), we considered the conditions separately (sub-phenotypes, so to say) in the initial iterations of the analysis, but we were unsure about our preliminary results because of the lower amount of samples for a given subset (e.g. “anterior unilateral cataracts samples”). After the last iteration of genotyped data, with a higher number of samples in our hand, we are elaborating strategies that can allow us to explore the possibility of association of a genomic region with a specific phenotype. As stated in the last report, we did not ignore the possibility of taking in account the phenotype sub classes (I-II-III) in light of the population structure of the dataset after our PCA analysis (see Figure 1). Nonetheless, the data gathered so far do not seem to indicate a strong effect of the sub-phenotypes indicated above compared to the stronger sub-population effect (see below).

SNP genotyping and data analysis

Since our last report, and because of the work carried out in the phenotype reassessment, we were able to increase the number of the suitable sample available for the research to 180 dogs (plus 57 dogs excluded from the case-control comparison part of the study, but still used for population analysis and statistics). A breakdown follows:

Total genotyped	**180**
Cases*	**52**
First class	27
Older age category	13
Second class	12
Controls	**71**
First class	40
Second Class	17
Third class	14
Excluded	**57**

Table 2 – Genotyped dogs. Cases (*) are subdivided in 28 bilateral, 17 asymmetrical, 5 unilateral cataracts. Excluded dogs will be soon re-evaluated for a possible re-inclusion.

As stated previously, we took advantage of the new, higher density (220k vs 170k) version of the current canine SNP chips. The new chip is ~30% more informative, with no information loss compared to the older one (that is, more SNPs were added to the new version but with full compatibility with the older one). Specific computational techniques were used to raise the information density of the old dataset at the level of the new one (“imputation”, through the popular software Beagle, extensively used by our group in other projects).

The dogs were divided in 52 cases and 71 controls. in the previous report, we stated that at the moment, we are satisfied with the cases/control ratio even taking into consideration the new samples. We wish to stress that we still need all the samples possible for the next steps of the project.

Each cases and controls subset was classified on the basis of the age of onset, laterality, anterior-posterior side of development of the cataract, and reliability of the sample (generally age-related). With the expanded dataset and the greater control cohort, proceeded to carry out updated analyses with the methods previously described and expanding the data analysis with new methods and working hypotheses.

GWAS: We carried out a whole new series of Genome Wide Association Studies (GWAS), a statistical analysis based on cases (52) and controls (71) within the population. As done previously, we used the excellent R package GenABEL (used in numerous animal genetics publications). The aim of such studies is to associate a specific genomic region and its markers to a cohort of study cases.

First and foremost, we carried out once general association of all the cases and all the controls and we compared it with the previous data. We confirmed the previously observed weak but detectable signal in the general analysis, used as a whole with no population structure adjustment. This confirmed that the higher number of cases and controls, indeed increased the power of the dataset.

Since we accumulated a greater number of controls, updated the cases, we repeated the population structure analysis as in the previous report: Principal Component Analysis (PCA) of the dataset (created by the same GenABEL software). As previously reported, roughly 80% of the total individuals would fall within one of the two sub-populations of uneven size (**Figure 1**).

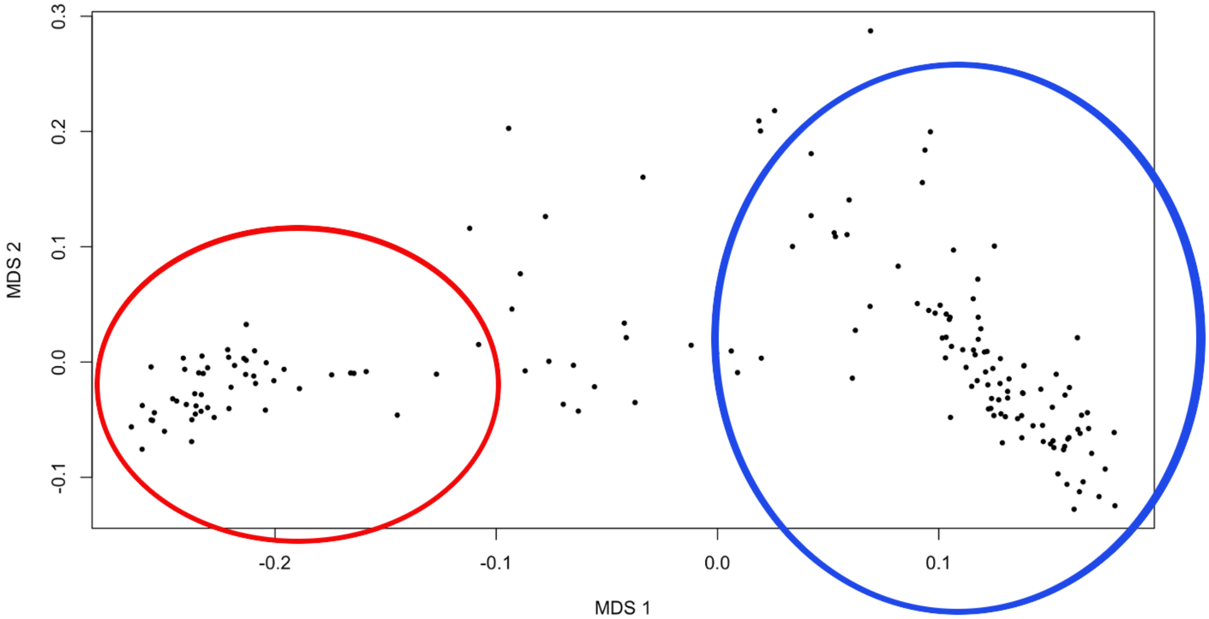


Figure 1 – Principal component analysis (PCA) of the American Cocker Spaniel cataract population. We can observe the two sub population clustering on the right (population A-blue) and left (population B-red) zones of the plot. In addition, we can observe a number of outliers not belonging to either group.

The two sub-populations were used for separate analysis, each time using as cases only the ones falling into one or another of the two sub-population.

In the case of the larger sub-population (we can call “population A”), the peaks obtained and the analysis of the quantiles confirmed the clear improvement registered in the last report. We confirmed the presence of the signal in a specific chromosomal region (as previously reported), and we confirmed the increase of signal in the secondary locus. (**Figure 2**).

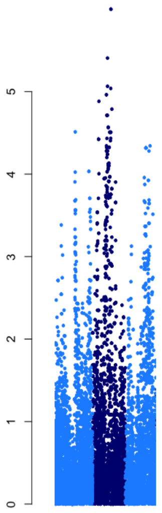


Figure 2 – Partial Manhattan plot of the case/control association for population A (right side of the PCA plot in Figure 1). Is possible to observe a suggestive peak on two specific chromosomal regions. Only the interested chromosomes, plus two other flanking it, are shown as a reference. Chromosomal number is not shown. The –$\log_{10}$(P-value) is a

function of the association – the higher, the better the association of a given genomic region with the phenotype is. The associated peak is one of the two found associated with the sub population the previous iteration.

Furthermore, we detected a suggestive peak associated with cases of the general population. (**Figure 3**).

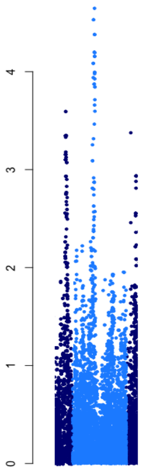


Figure 3 – Partial Manhattan plot of the case/control association for the whole population (left side of the PCA plot in Figure 1). Is possible to observe the suggestive peaks (under significance threshold. Only the interested chromosomes, plus two other flanking them are shown as reference. Chromosomal number is not shown. The –log10(P-value) is a function of the association – the higher, the better the association of a given genomic region with the phenotype is. This result could optimistically lead to a marker shared by a number of cases belonging to both populations.

Phasing: We are running the haplotype phasing with the software Beagle, that is in our experience very reliable (it has been tested in different ongoing and finished projects carried out by our group – it's also widely reported in literature). We focused primarily on the regions reported above, and on any suggestive peak identified by GWAS for populations A, B and for the total population. We counted cases and controls with the suspected haplotypes in order to identify trends (**Figure 4**).

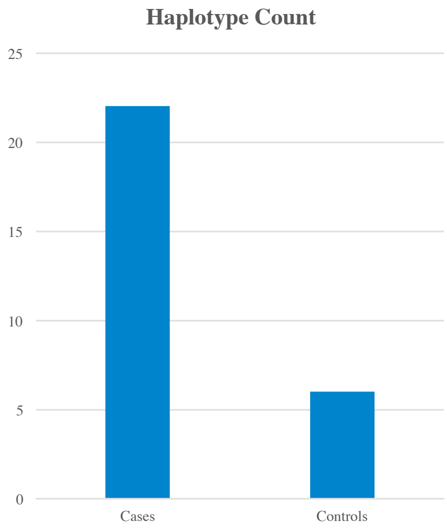


Figure 4 – Example of haplotype count between cases and controls. Note the greater frequency of the candidate haplotype in the cases. The procedure has been repeated several times for each candidate region.

With this data in our hands, and the additional controls genotyped, we moved toward the next exciting step of our project, and to our plan for a deeper and targeted data generation. In fact, we selected the cases and controls for our WGS in base of the presence and absence of the haplotypes.

Homozygosity mapping: In the previous report, we detected no homozygous region exclusive for the cases; furthermore, in the new output, we found no homozygous region present in high number in the cataract cases, that was exclusive for such cases, and thus not present in the controls. Since it's possible that the cataract condition (or at least some of these, if we are dealing with more than one within the population) is recessive, as previously suggested, but not in a single autosomal manner. Furthermore, it's also possible that two regions apparently identical between cases and controls are in fact distinct at the fine molecular level. We count on whole genome sequencing data also to elucidate this possibility.

Future prospects and plans

A complex disease: We hypothesized in the previous report that the occurrence of cataracts in American Cocker Spaniel is a probable complex of diseases. There are at least 2 genes involved that results in cataracts with similar phenotypes. As well, it is likely that a 3rd modifier gene locus is involved. As shown, a greater amount of cases and controls leads to better and more encouraging results. The selection of the appropriate sub-populations of cases and controls moved forward the analyses and the project, and we are now have been able to identify our candidate region and to implement whole genome sequencing (WGS).

Tackling the complexity: As proposed previously and confirmed in the current studies, we can not identify a simple and complete association of a single marker with the cataract in ACS. We have found that we can trace and identify trends and associations both under the assumption of a recessive disease, and under the assumption of a disease associated with loci of vulnerability not necessarily inherited on a recessive manner (we cannot, at this point, suggest a dominant inheritance – if such, the penetrance would be fairly low or dependent from the co-existence of multiple factors, not necessarily all of them genetic).

We update our immediate and future objectives as listed below, and compare them with what stated in the last report.

. A) We renew, as in the last report, our stated intention to increase the sample number in the database: a greater number of cases means to be able to enrich the specific sub-populations, and a greater number of controls allowing us to avoid false positives. The Research Scientist dedicated to the project spends a significant amount of time in the management of the database and in the interaction with the breeders and owners to obtain samples and updates, and that our database improved in numbers and diversity. We think that we reached a sufficient “critical mass” that allows us to plan for a deeper analysis, but we will quite obviously keep updating our data and gathering new samples.

. B) We previously stated that we would go through an in-depth analysis of the data output, never ignoring the slightest suggestive peak. In the current phase of the project, we are confident that we have in our hands candidate regions detected through GWAS. We previously planned an additional round of genotyping, that has been carried out. As planned, we have now a re-phased dataset, with classified genomic region with a suggestive GWAS peak under a case/control table of haplotypes and homozygous regions. This data will be compared with the WGS results.

. C) Although we have insufficient data to find a pattern correlation between specific of sub-phenotypes (age related, uni- or bi-laterality, position on the anteroposterior axis), we have reliable sub-population association. We consider such cross-reference the analyzed data still very crucial. Once the markers are identified, we could *theoretically* pinpoint a specific combination of vulnerability loci associated with the condition.

. D) As planned, we selected eight samples suitable for whole genome sequencing. As stated previously, since we could expect more than one region of the genome involved in the expression of the phenotype, the sequencing of a whole genome through WGS would be a suitable answer in order to quickly identify a small number of candidate variants. We previously stated that the price of WGS is decreasing each year, and our current collaborations allow us for a viable deal. For this reason, ad because of the different candidate regions identified, we opted so sequence four cases and four control, eight dogs in total. The WGS is presently being conducted in Berne, Switzerland and, when completed, the raw data will be file transferred to our lab. Depending on the results, additional dogs may be used for WGS.

. E) Validation, in two steps – through further sequencing, investigating the segregation of a candidate variant within the population, and/or with further experiments confirming in vitro a supposed effect of the variant on gene expression, translation, splicing. Obviously, this step will be carried out once definitive data from step D will be available. Once the WGS data will be available, our findings could lead us in two different directions: we could find evidence of one or more better defined candidate intervals and sequence more dogs in order to find additional evidence. Alternatively, we may consider using a much higher-density SNP chip technology now available for dogs that could help with the mapping of the cataract variant, with the assumption that the target regions are smaller than expected because of the age of the mutation.

We are very excited to the prospect of whole genome sequencing dog samples. The acquisition of this new, deeper and informative dataset will allow us to move the project forward. The soon-to-be-received WGS data will point us in the right direction.