Molecular Genetic Studies of Inherited Cataracts in the American Cocker Spaniel - progress report

University of Pennsylvania, July 15, 2020

Personnel:
Gustavo Aguirre, VMD, PhD; Principal Investigator
Leonardo Murgiano, PhD; Senior Research Investigator and Project Scientist
(Doreen Becker, DVM, PhD; former Post-doctoral fellow and Project Scientist).,
Jessica Niggel, M.Sc., Research scientist
(Dina Torjman B.Sc., former Research scientist)

Objectives

The principal objective is the identification of the genes and genetic variants responsible for inherited cataract in American Cocker Spaniels (ACS). Our goal consists in the development of a genetic test that can identify genetically normal, affected and carrier dogs for any variant that directly is considered deleterious or implicated with the development of cataracts. In this report we will describe the progresses achieved, the last important steps implemented, the data and feedback obtained with such implementation, and next moves.

Cataracts are the most common cause of vison impairment in humans and many 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. Data gathered to this point support such interpretation for the ACS. 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. To elaborate, we observed the presence of potential risk factors based on the sub-population observed.

As stated previously a significant element in the progress of our project was the thorough classification of suitable and verified samples in the ACS population, included a constant re-analysis and update of the cases and controls present in our database, thanks to the outstanding cooperation of the owners and breeders. This allowed us to pinpoint specific areas of the genome associated to varying degrees with the condition, and to refine such association with each iteration of the analysis (included the exclusion of false positives, which are an issue we have to keep in mind in our research).

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 breed club. After reaching a sufficient amount of samples, we planned and executed the use additional resources and techniques in order to move forward the project.

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. ACS are among the most commonly cataract-affected dog breeds.

As previously reported, we acknowledged a spectrum of cataract

phenotypes differing in location, progression rate, whether they are unilateral or bilateral, genetic background and age of onset. We considered the latter parameter, above the rest, as a most crucial factor for the classification and grouping of our samples. 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.

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. As stated previously, this is essential in order to select candidate cases for cataracts predictable as having a genetic etiology. We also calculated any possible correlation between the categories mentioned above.

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 valid 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 will recap on specific sub-phenotypes detected.

(II) is generally achieved through the use of SNP genotyping. The

method uses purified DNA, preferably obtained from blood samples 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). Research is constantly trying to improve such technology with denser chips, that equal to greater amount of information contained.

Another common type of analysis is the Genome Wide Association Study (GWAS) that uses the SNP chip platform. 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, since there is no perfect segregation of the markers between cases and controls. 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, included a genome in its entirety). A common and fast sequencing method is 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. In fact, we will soon use it extensively on specific candidate markers in order to assess their

frequency and segregation (between cases and controls) of a number of candidate variants in our population.

A limited, targeted use of Sanger sequencing is relatively cheap, but the exploration of a whole genome sequence would make it unfeasible and too expensive. On the other hand, Whole Genome Sequencing (WGS) methods have brought a whole new level in the exploration of genetic defects, because they allow 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 time and cost prohibitive if done using more conventional approaches.

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(s) 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):

The current COVID-19 pandemic situation unfortunately is still ongoing, and we are all in fact going through challenging times. During the initial lockdown phase, we were able to send samples to be run in the high density SNP chip and continued some of the WGS analysis on samples already collected and run. Although inconvenient having to work remotely, we were able to continue this work quite effectively. Now, the University of Pennsylvania carefully planned a restoration of the activities after the initial lockdown (which included lab work and access to more powerful facilities for data analysis), and now we have restarted our research activities fully.

We implemented several strategies during the period of the study. As stated previously, the choice of a given approach was done in base of

the quality of the dataset available at the moment, and the reliability of the information. The constant influx of new samples improved the dataset on each subsequent iteration.

Candidate genes and Pedigree analysis

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 previously, the results were negative, and 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 those, 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. This indicates that while some of the cataract phenotypes may appear similar, the underlying genetic cause is different, and our aim to identify the underlying genetic causes.

Samples received

Compared to the previous report, the number of dogs participating the study increased to 831 from the 793 reported last time. The number would have been higher if we had been able to receive samples and clinical records during the lockdown and university closure)March 13-June 8). A short breakdown of the samples follows:

Total of Informative dogs	**552**
Potential cases	***107***
Bilateral	79
Unilateral or very Asymmetric	28

Controls	***446***
Too young to be properly assessed for study inclusion at this time	198
Total of Excluded dogs	**279**

Table 1 –Total of dogs included 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, the dog prematurely deceased (especially if DNA/blood is missing), lack of feedback on updates (fortunately, this now is a very 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 previously collaborated in the study, or by the project Research scientist in our lab (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 resulting in better DNA yield). All of the blood samples have been sent to us in EDTA lined 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 to temporarily not use the samples awaiting future clinical updates). This in person assessment of the samples is done ~ 3x/year and includes all individuals involved in the project. It was suspended during the university closure but will be started through zoom conferencing once new samples are submitted. Because of health regulations, it is not possible for 3 individuals to sit side by side and deliberate on the samples-however, with zoom and 'page sharing', this can be done.

Typical classification of samples during a phenotype evaluation session.

In addition to the dogs listed in Table 1, a recent round of data analysis

session just prior to the pandemic closure of research laboratories allowed us to add 11 additional controls (the dogs were“promoted” based on updated re-examinations). The contribution of the breeders is continuous and important, and we previously stated that we were going to examine 42 dogs so far deemed too young to tell or with incomplete records we are planning to do it in the next period, and see what can be integrated in the sequencing and genotyping dataset.

As previously stated, we have discovered that ACS seem to exhibit distinct sub-phenotypes of inherited cataract. Primarily, we registered (I) a possible stratification 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 fellow eye. 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 in a high amount 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 plan to develop strategies that can allow us to explore the possibility of association of a genomic region with a specific phenotype.

Importantly, 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 (Figure 1). However, we previously reported that 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). It is possible that the sub-phenotypes are influenced by genetic (e.g., modifier genes) or environmental factors (e.g., diet, medications, etc). This is an area we will examine closely after the gene and disease causing mutations are identified.

SNP genotyping and data analysis

Since our last report, we improved considerably our SNP chip dataset. First and foremost, we were able to review our records and improve the amount of the second-best quality cases included in the study (re-inserted from the previously excluded dogs). Such dogs increased the amount of total cases to 62. The total of excluded dogs dropped from 57 to 48. The total of high quality controls dropped by one. A breakdown is reported in the table.

Cases*	**62**
First class	27
Older age category	25
Second class	10
Controls	**70**
First class	39
Second Class	17
Third class	14
Excluded	**48**

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 once updated examination results are obtained.

We previously took advantage of the new, higher density (220k vs 170k) 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).

In addition to that, just before the lockdown period we managed to select dogs from our best samples (60 dogs in total - 26 cases and 34 controls of the highest quality, see above) and send them to be processed for a third type of SNP chip using a new technology. Such technology allows the genotyping of the selected samples for 712k SNPs, more than three times the original information! Importantly, the older SNPs are still present and therefore can be used to impute (see above) this new information in the rest of the dataset. This is important because an additional cycle of GWAS with this new data has been carried out (both with the 60 dogs exclusively, and with the imputed dataset as a whole, see below). The reason for using the 712k chip is to identify, if possible, areas of the genome with poor coverage in the older chip versions, and these new areas analyzed might harbor candidate gene(s) that require greater scrutiny.

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). We checked whether there was some sex or age bias in the ratio of bilateral and unilateral cataract. Examples follow. Note that the phenotypes and sub-phenotypes do not seem to diverge significantly according to age and

sex.

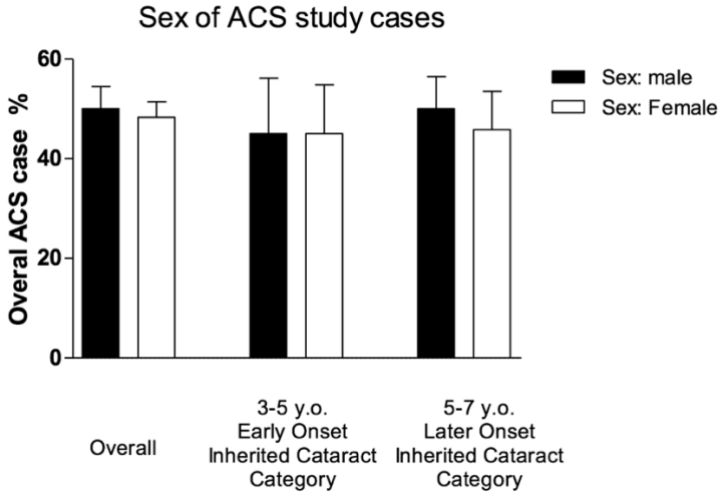

Figure 1 – distribution of the % of males and females in the genotyped cases per age of onset.

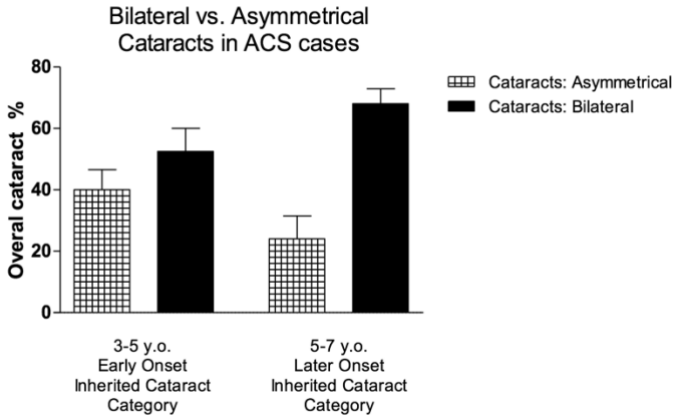

Figure 2 – distribution of bilateral and asymmetrical cases by age of onset.

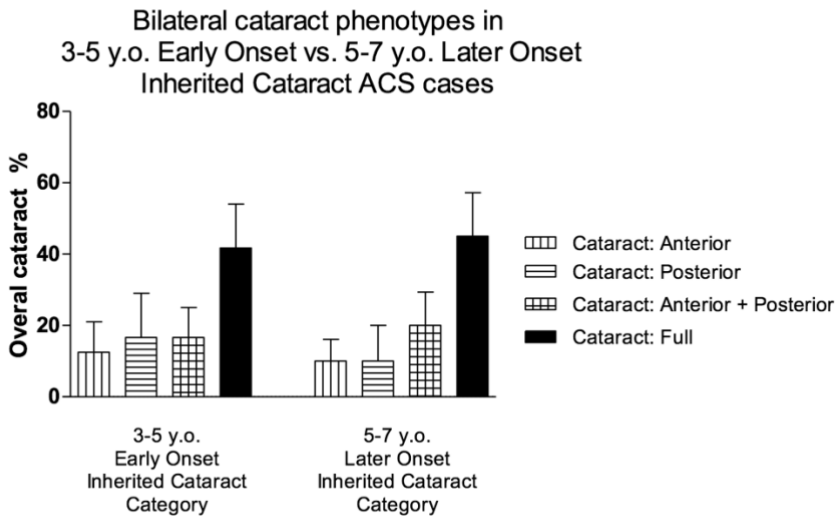

Figure 3 – distribution of cataract types in bilateral cataract cases.

GWAS: We carried out a whole new series of Genome Wide Association Studies (GWAS). Then we carried out another series of analyses using all the cases (62) and controls (70) within the whole population. As done previously, we used the excellent R package GenABEL (used in numerous animal genetics studies). The aim of such studies is to associate a specific genomic region and its markers to a cohort of study cases. In addition, we used association analysis packages from plink 2.0 in order to validate the findings and check whether the association found is consistent with one carried out with a different program.

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 before, roughly 80% of the total individuals would fall within one of the two sub-populations of uneven size (**Figure 4**).

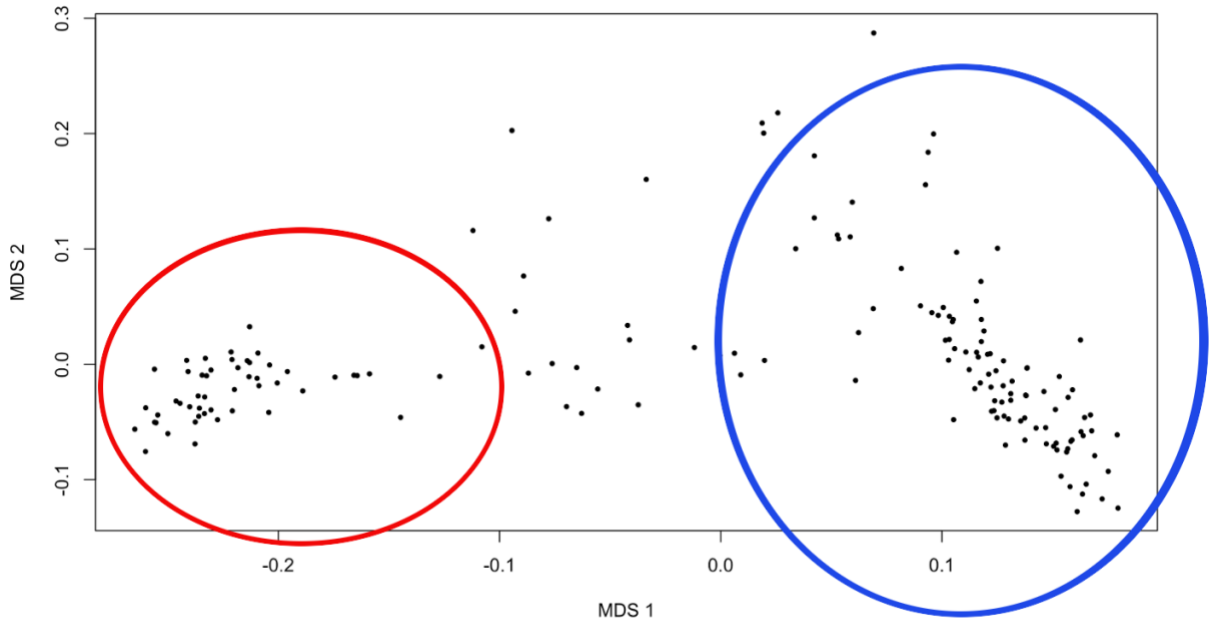

Figure 4 – PCA of the ACS cataract population. We can observe the two sub population clustering on the right (population A) and left (population B) zones of the plot. In addition, we can observe a number of outliers not belonging to any of the two. One of the candidate haplotypes is enriched in one of the two sub-populations. The overall structure remained (unsurprisingly) unchanged with the new high-density data.

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 5**).

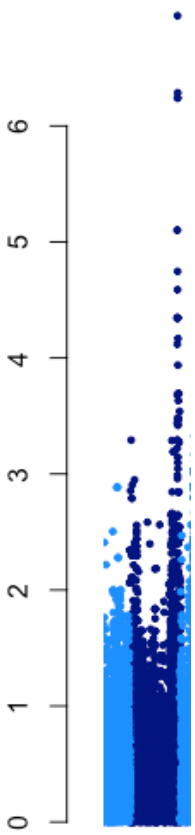

Figure 4 – Partial Manhattan plot of the case/control association for population A (right side of the PCA plot, Figure 1). It is possible to observe a suggestive peak on two specific chromosomal region, one order of magnitude bigger than the last iteration. Only the interested chromosomes, plus two other 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.

We maintained signals in regions associated with the condition across the samples and improved the allele count through the higher density SNP chip data (see below).

Phasing: As stated previously, 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 concluded 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. This is for us vital order to know “where” to search for candidate markers.

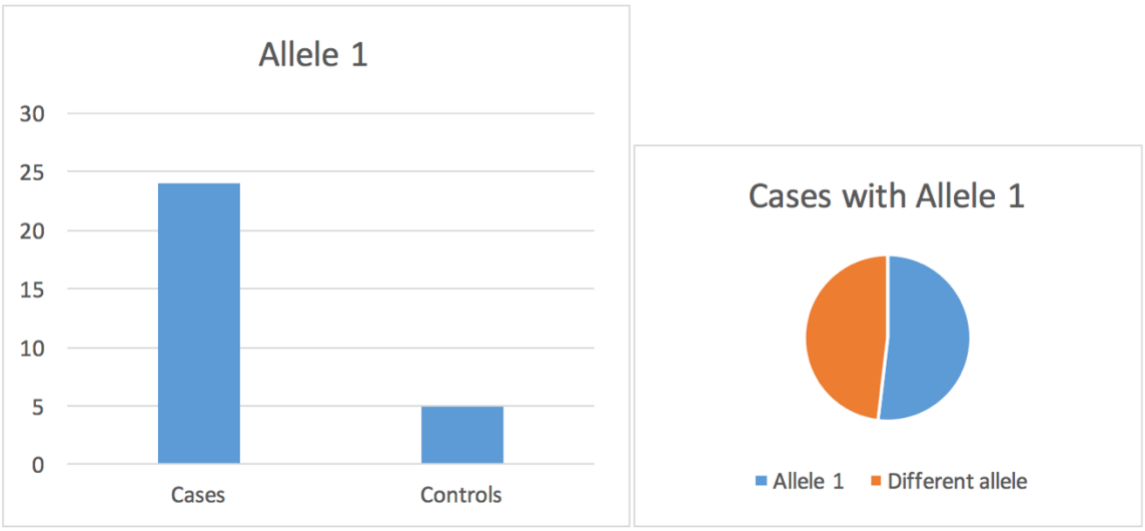

Figure 5 – Left: occurrence of Allele 1 between cases and controls. Note the greater frequency of the candidate haplotype in the cases. The procedure has been repeated several times for each candidate region. Right: pie chart with amount of cases carrying the allele.

We also considered possible low penetrance of a risk factor, selecting alleles with good signal but frequent in the controls as well. We cannot exclude anything at this stage, marker genotyping will tell!

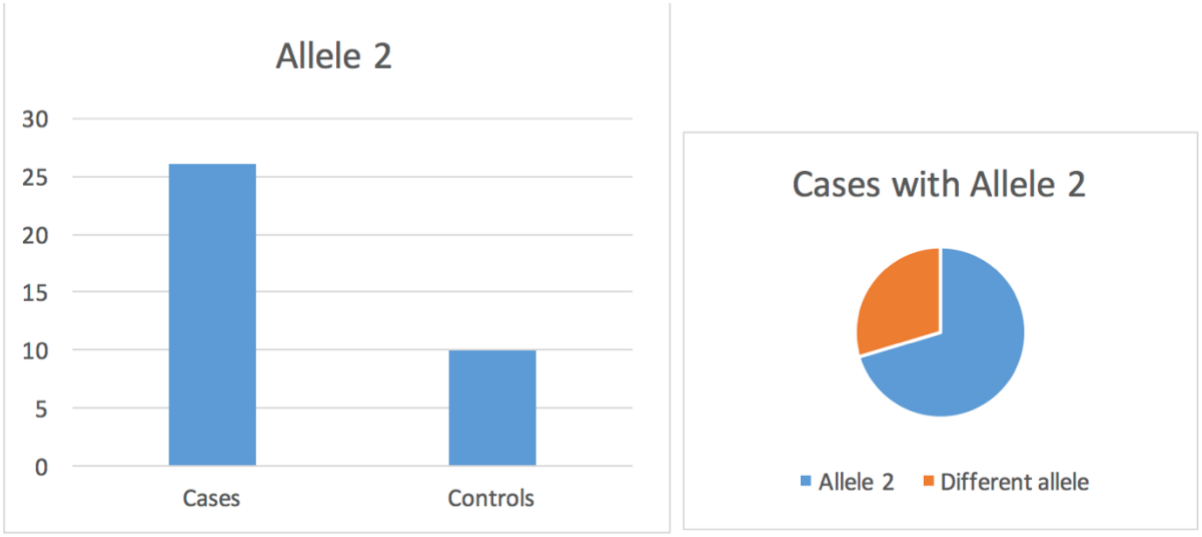

Figure 6 – Left: occurrence of Allele 2 between cases and controls. Right: pie chart with the amount of cases with this allele. Note how that ~70% of the cases have the allele.

Homozygosity mapping: In the previous report, we speculated that 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. For this reason, we are still screening regions in which most dogs (cases and controls) are homozygous searching for exclusive markers to add to the pools of the one to be tested. This is in addition to the results obtained from GWAS.

Whole genome sequencing: We implemented the planned WGS. Cases and controls of the best quality were selected, 2 cases, 2 cases selected from the older category, and 4 of the best controls. As described above, the state of the pandemics and of the lockdowns initially prevented us from

using the planned facility in Europe, but we quickly found a workaround. Data has been processed and is currently being filtered in order to proceed for the selection of markers that must be tested in the whole ACS DNA dataset we have created. We hope to have some of this data available for the August webinar.

Future prospects and plans

A complex disease: We hypothesized in the previous report that the occurrence of cataracts in American Cocker Spaniel is likely a complex of 2 or more diseases. 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: Even if we cannot show, at this moment, 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) As always, we renew 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 proceed with our plan, but we will quite obviously keep updating our data and gathering new samples, in the very least for validation.

. 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 that will focus the research using the newly acquired WGS results. The last iteration of SNP chip data generation narrowed the amount of such regions and also (potentially) improved the ratio of the cases sharing a genomic region, due to the higher resolution of the 712k chip used. If additional suitable candidates will be found, we will genotype more dogs.

. C) Our preliminary cross-reference of the data did not point out any specific correlation between laterality, i.e. unilateral or bilateral, and age of cataract onset. Nonetheless, once a smaller pool of markers will be available, we will observe their segregation with the sub-phenotypes once more. In the last report, we proposed using the 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. This has been done and it improved our results significantly, albeit less than in an ideal perfect segregation.

. D) Whole genome sequencing data analysis will be the focus of our next steps. Through the data generated, we will select suitable candidate markers to be tested within the population. Validation will then happen 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.

In the last report, we shared our excitement for the prospect of whole genome sequencing dog samples. We still think that acquisition of this informative dataset will allow us to move the project forward. At the moment, our main task is to carefully filter, test, and validate in order to refine what has been done so far. With the progress made since the last report, we are quite optimistic of the results and certain that we are using the correct approach going forward.