Molecular Genetic Studies of Inherited Cataracts in the American Cocker Spaniel

- progress report

University of Pennsylvania, December 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 remains 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. The current report will show the progress achieved so far (in relation to the last report, July 17, 2019), the resources employed and to request the support necessary in order to continue the studies and to finalize our initial goals.

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, American Cocker Spaniel (ACS) included, 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 perhaps 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 suggest a situation more complex than the one predicted in the preliminary phase of the project.

Our work significantly progressed through tight communications with the owners and the breeding association and their outstanding contribution. 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.

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 (divided in 2 age groups, see below). We considered the latter parameter, above the rest, as the most crucial for the classification and grouping of our samples, and we stress for a correct gathering of information about the affected and unaffected dogs 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 different iterations of the analysis have been carried out 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 (usually from blood) of cases and controls that is placed on SNP chips, specific platforms scanned for strategically selected genetic variants used as markers, called single nucleotide polymorphisms (SNPs). Thorough the information obtained by such experiments, the researchers can explore the presence of common (and ideally exclusive, or at least enriched) 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); a similar analysis can be obtained analyzing the studied population in phased data.

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. Conveniently, the dataset generated for GWAS analysis is also used to search shared homozygous regions among the cases, as well as phased data.

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).

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.

Samples received

In the moment we are writing this report, the current dataset is composed by 823 American Cocker Spaniels (from the 793 of the July report). The dogs genotyped in several iterations on Illumina 170k and later 220k SNP chip platforms amount to 180.

Total of Informative dogs	**539**
Potential cases	***93***
Bilateral	74
Unilateral or very Asymmetric	21
Controls	***442***
Too young to be properly assessed	210
Total of Excluded dogs	**259**

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 board 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!

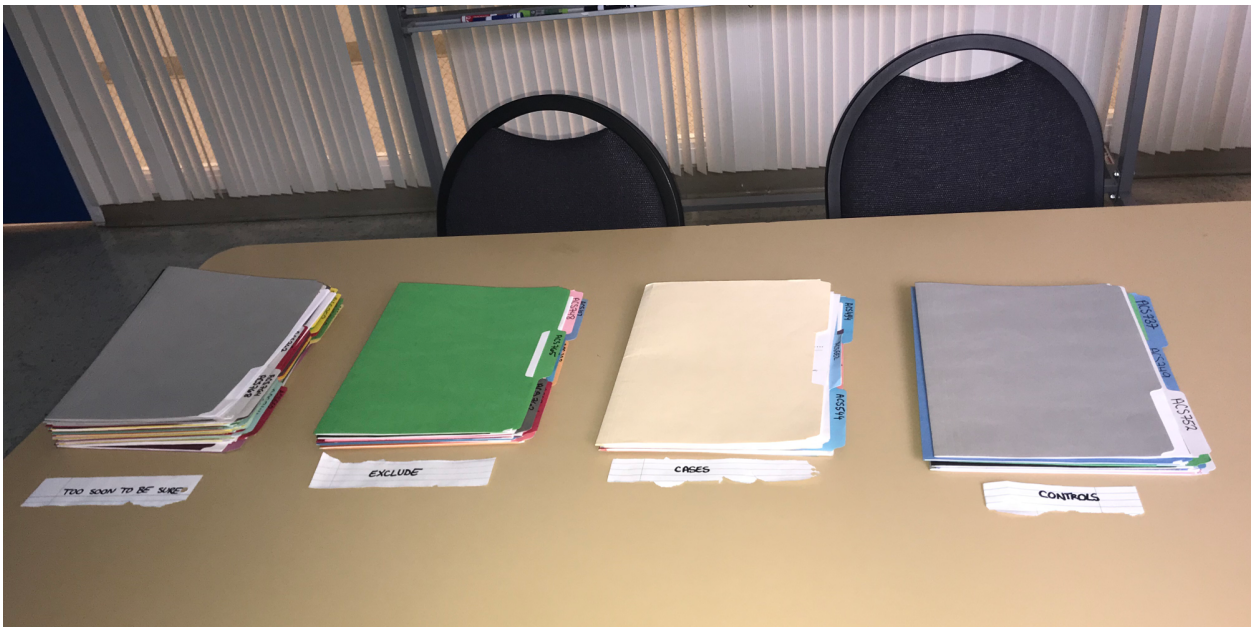
DNA samples were isolated from blood or buccal swabs by personnel at OptiGen LLC who are collaborating in the study (in this regard, we wish to thank the breeders for the fact that the overwhelming majority of samples are blood samples, easier to work with and generally bringing with better DNA yield). 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 (last update: 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 withheld 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. In the last report, we stated that we were going to examine 42 dogs so far deemed too young to tell or with incomplete records; of these, 5 dogs have been added to the records in this way as part of our quarterly data review.

As previously stated, we have discovered that ACS seem to exhibit distinct sub-types of phenotypes of cataracts that fit the inherited classification. 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") 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 developing 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. 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). Of these, 53 are case, 72 are controls (updated from the last report) and the rest are dogs previously genotyped but excluded from the analysis for various reasons.

Cases*	**53**
First class	28
Older age category	13
Second class	12
Controls	**72**
First class	41
Second Class	17
Third class	14
Excluded	**57**

Table 2 – Genotyped dogs. Cases (*) are subdivided in 28 bilateral, 18 asymmetrical, 5 unilateral cataracts. Excluded dogs are re-evaluated for a possible re-inclusion as new re-examination records are received.

As stated previously, once such resource became available, 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 for imputation and phasing, in this and in other projects).

The dogs are divided in 53 cases and 72 controls. As stated in the previous report, we are satisfied with the cases/control ratio but we added new samples if available. We reiterate 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 (53) and controls (72) within the population. As done previously, we used the excellent R package GenABEL (used in numerous animal genetics publications). We confirmed the previously observed weak but detectable signal in the general analysis, used as a whole with no population structure adjustment. We repeated the GWAS and population structure analysis Principal Component Analysis (PCA) of the dataset (created by the same GenABEL software). Expectedly, roughly 80% of the total individuals would fall within one of the two sub-populations of uneven size (see July report).

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

of the larger sub-population (we refer to this as “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. We confirmed suggestive peak associated with cases of the general population.

Phasing: We re-ran the haplotype phasing with the software Beagle, that is in our experience very reliable, and has been tested in different ongoing and finished projects carried out by our group. Albeit we focused primarily on candidate regions pointed out by GWAS, the data encompass all the markers available. We counted cases and controls with the suspected haplotypes in order to identify trends as reported previously.

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. For this reason, we have to considerthat it is possible that cataracts (or at least some of these, if we are dealing with more than one within the population) are recessive, as previously suggested, but not in a simple autosomal manner. Furthermore, also it is 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, and evaluated trends and associations as with heterozygous haplotypes.

Whole Genome Sequencing: 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 received the whole genome sequencing data for the cases and controls for our WGS based on the presence or absence of the haplotypes. As planned in the July report, once this phase of the study was concluded, we opted for WGS of eight dogs: two cases (2-5 years), two later-onset cases, and four top-tier older controls greater than 9 years. Once we processed the data, we focused on the candidate regions detected. Our first aim was to quickly identify potential variants (DNA regions that have a sequence change that could be causally disease-associated) exclusive for the cases in the suspected haplotypes. We also selected variants enriched in the cases but still present in the controls. Due to A) the

large number of potentially suitable variants and candidate regions, and B) the fact that as reported in July, the segregation of candidate haplotypes/variants between cases and controls is not perfect, this phase is not concluded yet and genotype analysis is still ongoing. Furthermore, more than a single marker could be associated at the same time with the cataract condition, forcing us to not exclude a given genetic variant outright, but, in fact, to take into account its co-presence with others.

Plans for the coming year

A complex disease: We hypothesized in the previous report that the occurrence of cataracts in American Cocker Spaniel is complex and may include more than one genetic defect each of which have a similar clinical disease phenotype. In this scenario, 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 may be 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 analysis of the WGS data.

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. In order to complete the this phase of the project, we plan to complete two critical steps during the coming year:

(I) Filtering of candidate variants obtained through WGS in cases and controls and assess their allele frequency within the population. We are giving priority to variants occurring in appropriate candidate genes (and falling within the positional candidate regions), but the completion of the genotyping analyses will still need months of work.

(II) Higher-density SNP chip genotyping for a selected amount of cases and controls with a new available technology, a canine SNP chip from Affymetrix with three times the SNP density and information compared to the Illumina one (e.g. 660,000 vs 220,000 SNPs). This new technology is costlier (+40%) than the old one but three times more informative. We plan therefore to carry out a new association study using a subset of same samples, but which, by being genotyped in the 660,000 platform will have greater power to detect association, and higher fine-mapping potential. This will reduce the number of candidate regions and even optimistically pinpoint a single, strong locus that previously was not detected because of the sparser density of the currently available data (we posit that a smaller region could be shared by the cases due the fact that the disease appeared decades ago, and there has been a high number of genetic recombinations which result in a smaller shared genetic interval).