American Cocker Spaniel Cataract – progress report

University of Pennsylvania, Feb 2, 2025

In this installment of the American Cocker Spaniel (ACS) Inherited Cataract study report, we would like to once again thank all the ACS breeders and owners for their continued support and participation in our study.
Compared to the previous report, the number of dogs participating in the study increased from 1020 to 1057, with five samples with enough over-time updates that allowed us to integrate them into our study.

Notably, the trend we observed during the last period is that the updates have diminished significantly. We understand that the progress toward the end goal has been slower than hoped. Nonetheless, we think it's essential to reiterate that without samples or updates, no progress is possible.

New technology, which can support researchers in unraveling the genetics behind simple and (in the case of the ACS) complex diseases, is developed every year. As described below, in the last year, we went through a significant effort to rebuild our whole dataset from the ground up. Any future breakthrough and subsequent mass-testing will start from the dogs currently in the study.

Our research for the cataract in ACS pertains a disease with a complex inheritance, as opposed to a simple one with a perfect segregation between causative and non-causative mutations.
Additionally, as we stated in the past, an additional complication is the presence of two major groups of cataracts – those with an early (2 to 5 years) and those of a later (more than 5 to 8 years) onset

On top of updating the reader, in this report, we will elaborate more than usual on the obstacles found in studying complex genetic diseases and the implications of these on our research.

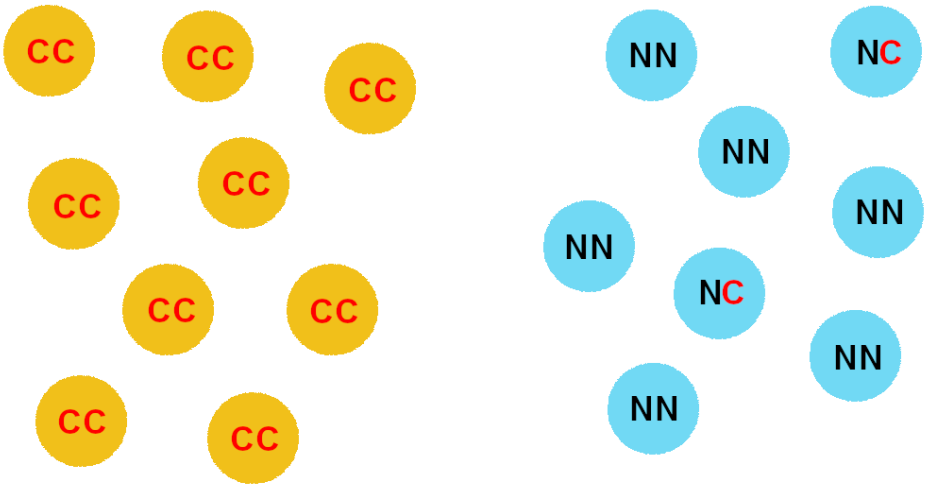

Figure 1 – Simple disease - in the example given, the disease is recessive. All the cases share two copies of the same CAUSATIVE marker (CC), and normal controls have no causative marker or just one copy (NN or NC). Since all the cases have two copies of the causative allele, controls are free from the marker, or just one copy, this analysis's statistic is clear-cut.

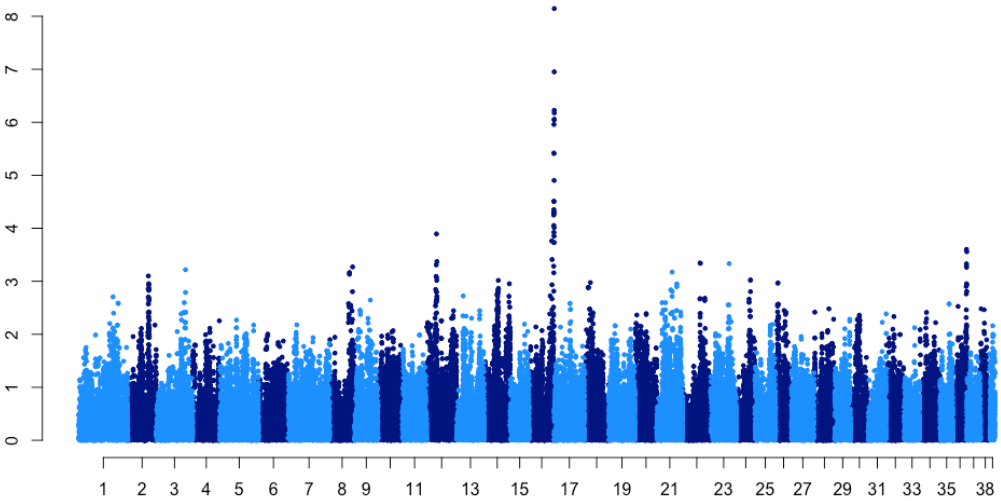

Figure 2 – Mapping of a “simple” disease – Each marker in cases and controls is compared, and a clear-cut trend is observed. In this case, the mapping used is genome-wide association (GWAS), which we use extensively (***this example is not done using ACS data***!). The peak observed here (chromosome 16) is a big hint at where we should have a look. This is how GWAS works – when a

given marker is strongly associated with the disease, the "dot" that represents it appears higher on the GWAS scale. Groups of markers with a stronger association form these peaks – that's where the researchers will look, hunting for genes.

Research on diseases with a complex inheritance requires particular attention to the dataset, especially in the quality of the phenotyping (in our case, constant updates) and in the number of samples of the dataset.

This is because the way causative markers behave in these conditions is way more complicated. The most crucial factor is that multiple markers associated with the condition exist in multiple combinations, and controls can have SOME of those markers in several copies closer to the cases!

Another contributing factor to the success of any analysis, and which helps to eliminate the above issue, is a diverse pool of samples. If only specific kennels of a breed contribute, we will never have a clear idea of the general picture.

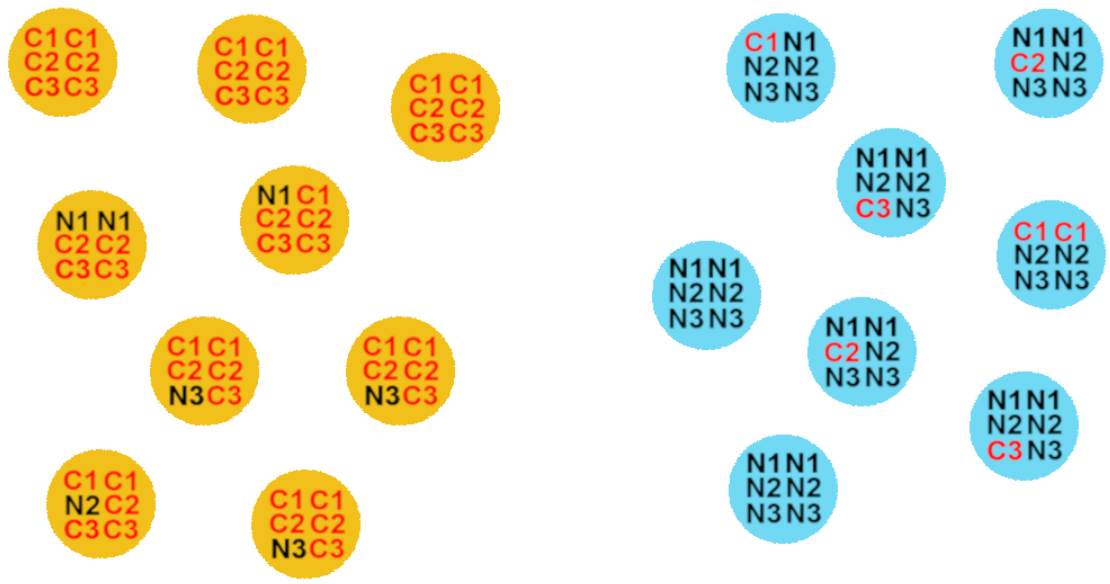

Figure 3 – Complex diseases with multiple markers. In the general population of cases and controls, various combinations of risk factors exist. For a given, SPECIFIC marker, the case can even be free of any causative copy (N N), while a control has two (C1 C1)! Why can this happen? The affected has other causative markers (C2, C3), while the control is free of any other causative marker (N2, N3). Running a comparison here is more complex, because any comparison of affected and unaffected by these markers

will have a lot of “noise”. This noise is what kept us from finding the causative markers. Only numbers -that is, a large and diverse group of dogs - help here. Plus, of course, good phenotypes obtained by constant updates.

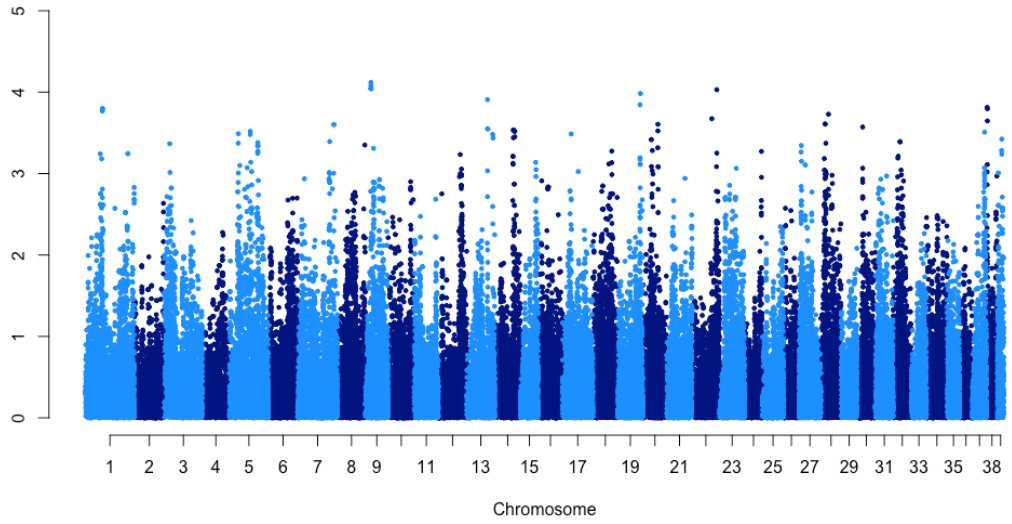

Figure 4 – A similar GWAS as in Figure 2, but with the preliminary data from dogs with a complex disease (***this example is also not done using ACS data***!). Markers have been compared, cases vs controls, in a similar way. Nonetheless, we can now see no “clear-cut” signal now. Where should we look here? It’s a forest of small, inconclusive peaks, and so many of them! There is no way to have any hint on where to look!

Thankfully, there is always room for improvement. Cases can get updates; more can be gathered. Sometimes though, even with the better low-pass tech, the problem lies in the experiment itself (as we explain below).

In our past report, we described the issues we faced due to the disease's complex inheritance and the SNP data's low informativeness. We also explained our strategy of investing heavily in whole genome sequencing and low-pass GWAS for the type of ACS population studied.

While using this strategy we obtained partial but encouraging results on a kind of cataract (late onset), we were pretty perplexed by the results for the early onset, which haven’t shown precise results. While we had many early onset cataract cases sequenced, the approach of simply focusing on sequenced cases had only partial results, probably due to what is illustrated in Figure 3 above.

To address such complexity, we decided to implement a strategy that would help us optimize the available sample pool. This strategy involves a collaborative effort with two other universities. We spent the last months carefully building a dataset that could be used for such an endeavor.

Usually, we send samples to be sequenced for low-pass to a company. This company then uses its own ”reference panel” for the imputation. What does it mean? What is imputation?

In the case of our type of research, imputation is a strategy that helps the investigators in dramatically improving the information they gather on the thousands of markers present in each dog studied.

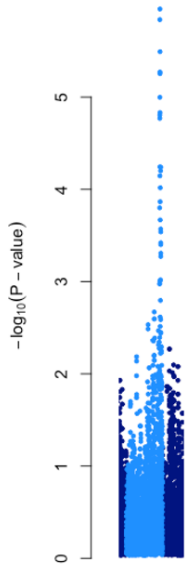

Figure 5 – A GWAS similar to the ones shown in Figure 2 and 4, but from an actual ACS report (hence the peak is only partially shown). This is, like in figure 4, a complex disease, but thanks to accurate phenotype and updates, we can observe that our research pointed out trends we can work with. The next steps are described below. Keep them coming!

When a genetic analysis is carried out, the current technology (“low pass”) generates information on a large number (hundreds of thousands or even millions, in fact) of markers for each dog. This extensive data is then used to compare cases and controls and pinpoint “zones” in the dog genome that are considered most likely to have the mutation we are looking for. In these zones, we search.

There is an obstacle, though – the technology that generates the millions of markers is not perfect (to be so, it would cost so much that an analysis with the required number of dogs would be unfeasible). What the technology does is to kind of “skim” through the dog's genetic information and “write down” only *some* of the markers. A different software will then read this partial marker information and fill in the gaps. How does it? By using a larger dataset as a

reference to fill these gaps.

Imagine reading a book, but only every three pages or so. The missing information is filled in using another larger book as a reference. The larger book, in this case is called a reference panel.

A reference panel is essentially a panel of dogs used as a “guide” for “filling up the gaps” in the data as part of the normal imputation process. While this is already reasonably accurate, the dogs usually used as references are unrelated to the ones we send, which could potentially mask or alter the result.

For this reason, we concluded that using our dogs as part of the imputation process is essential and that a collaborative effort was required to obtain an adequate number of non-ACS dogs, which is also needed for such a custom reference panel.

To do so, we started collaborating with the University of Minnesota (Dr. Steven Friedenberg). We shared the whole genome sequencing (WGS) and low-pass genotyping files with UM, and we let the UM group impute in their computer our low-pass files to generate a reliable dataset, leading to more accurate GWAS analyses. This also shows that the effort and money spent on deep (i.e., the non-lowpass, “standard” one) ACS WGS pays off because it allows us to create a reliable reference panel for our dogs, which our collaborators could use.

The imputation was carried out as follows: 157 high-quality (constant updates and clear phenotype) ACS were sequenced on "low pass". These dogs were imputed using a general panel created by the collaborators plus 23 ACS selected from our specific pool of whole genome sequenced ACS. We took a particular care in selecting a diverse population of cases and controls. Among the cases, we diversified the dogs sequenced by any difference in the phenotype, primarily age of onset.

Additionally, we began a collaboration with the University of Texas, TAMU, Dept of Small Animal Clinical Sci. The collaborators have previously sequenced tens of dogs for us. This collaboration will allow us access to thousands of control dogs we can use for the filtering and allele frequency check of any candidate variant.

We had a setback when Dr. Sigdel, the postdoc working on cataracts in ACS, left for a career change. Dr. Sigdel is a fantastic researcher, and finding a suitable substitute took time, delaying our investigation.

While this was a setback, we were determined to search for a new Postdoctoral researcher. Dr. Elizabeth Greif took over the position in January 2025 and will tackle all the generated data. Dr. Greif just obtained her PhD under Prof. Leigh Anne Clark's (UGA) supervision, with Dr. Steven Friedenberg as co-supervisor.

We look forward to the results of these improved imputed datasets that will be assigned to Dr. Greif, who is experienced in discovering markers associated with complex traits.

When this report is being written, the just-joined Dr. Greif is reviewing all the pedigree and phenotype data of each dog sequenced and used for the association, ready to use this new information in the following weeks on the latest, high-quality data.

Of note, the current ACS cataract research is incorporated into our lab, and the postdoc working on it processes the data as part of a series of projects she is dedicated to.

We expect the results obtained from the generated data will give more insight into candidate markers. Although we don't discard the possibility of an additional round of sequencing to further, the dataset (whole genome) could be required to accelerate the discovery of good candidate markers for late-onset dogs. We will know this in 3-6 months.

The scenario we are hopefully looking forward in this year time period to is that most of our effort will be focused on mass genotyping of the general available ACS population for candidate markers for early and late-onset cataract.

Whatever our new move will be, samples and updates are the only thing that can move the project forward.

Gustavo Aguirre, VMD, PhD; Principal Investigator
Leonardo Murgiano, PhD; co-Principal Investigator
Jessica Niggel, M.Sc.; Research Scientist
Anil Sigdel, DVM, PhD; Former Postdoctoral Researcher
Elizabeth Greif, PhD Postdoctoral Researcher 2025