1. Objectives/aims

Recessive lethal alleles exist benignly in breeding populations, until a sire and dam carrying them are mated. One quarter of the resulting pregnancies will be homozygous for the lethal allele and will result in an aborted pregnancy. Missed breeding opportunities are expensive. These recessive lethal alleles will increase in frequency within the population, distributed as heterozygotes, until ultimately manifesting themselves as lethal when two heterozygous "carriers" are mated[1]. If it is possible to identify these lethal alleles, then farm managers can mitigate the problem by ensuring that two carriers are never mated to one another, thus boosting the likelihood of a successful pregnancy by 25% for any carrier.

The objective of this work is to create a species agnostic service capable of integrating pathology reports as well as sequencing, analyzing, and publishing variation data for abortions in farm animals with no known etiology such as a viral or bacterial infection. This will be accomplished by executing the following specific aims.

Aim 1) Collect pathology reports and samples from abortions occurring in sheep, cattle, and horses, sequence the samples, and identify genetic variants relative to the corresponding species reference genomes.

Aim 2) Compare these variants with a database of healthy animals and identify alleles that may be responsible for the abortion phenotype.

Aim 3) Publish these findings to a queryable interface for use by breeders, and the scientific community.

Once this service is developed, an inexpensive mechanism will exist, and may be perpetuated for researchers and producers to submit aborted samples with no known etiology for the purpose of identifying lethal alleles in their sires and dams.

Preliminary work:

In projects led by Dr. Mike Heaton at the USDA, Dr. Kalbfleisch has created databases of sequence data, as well as variant data in the form of vcf, and g.vcf files for panels of 96 cattle and sheep respectively derived from popular North American breeds[2, 3]. In another project, Dr. Kalbfleisch is currently working with Dr. Jessica Petersen and Dr. Ernie Bailey to build a database of genetic variation based on the sequence data of 100 North American Thoroughbreds. In these efforts he has developed pipelines that run at the University of Kentucky High Performance Computing Center specifically for the mapping, and variant analysis of these species[2-5]. These data provide catalogs of variants that occur in these populations at at least 3% allele frequency. It can be shown that for a recessive allele to be homozygous in 1% of the population, it must be present at 18% as heterozygotes in trait "carriers", and at 10% allele frequency in the population overall. If the allele frequencies are given as p for the *A* allele, and q for the lethal *a* allele, we know that $(p+q)^2 = 1 = p^2 + 2pq + q^2$ where p^2 would be the number of *A* homozygotes, 2pq would be the number of heterozygotes *Aa*, and q^2 would be the number of animals homozygous for *a*. As such, if a recessive lethal allele were to appear and spread randomly, it could increase in frequency in the population to 10% before a homozygous event occurred at 1% $(0.9A + 0.1a)^2 = 0.81AA + 0.18Aa + 0.01aa$. Therefore, these samples of 96, 96, and 100 healthy animals for cattle, sheep and horses respectively should provide a sufficient catalog of alleles for comparison including the lethal ones that occur as heterozygotes in the larger population. Our strategy will be to sequence abortions, or animals who died shortly after birth, to identify homozygous alleles in the abortions that occur as heterozygotes in the larger populations but do not occur as homozygotes. Further, we will be able to identify all rare alleles in the animal that may be de novo lethals, and we will be able to assess chromosomal composition from read depth to identify animals with aneuploidy.

Plans for achieving the objectives: Dr. Kalbfleisch has enlisted the help of 3 groups of collaborators for sample collection

1) Dr. Jessica Petersen and Dr. David Steffen from the University of Nebraska Lincoln will provide cattle abortions, or calves who died shortly after birth with no known etiology along with pathology reports for up to 15 animals of the 60 to 70 they receive annually.
2) Dr. Brenda Murdoch and Dr. Darren Hagen from the University of Idaho, and Oklahoma State University respectively will provide sheep abortions, or lambs who died shortly after birth with no known etiology along with pathology reports for up to 15 animals.
3) Dr. Jennifer Janes from the University of Kentucky Veterinary Diagnostic Lab will provide samples from horse abortions, or foals who died shortly after birth with no known etiology along with pathology reports for up to 15 animals of the 100 they receive annually.

Up to a total of 40 animals spanning the three species will be sequenced on the Illumina short read platform to 20X coverage.

2. Furthering the aims of the AG2PI

Data with accurate, descriptive phenotypes coupled with genetic data are the foundation of the AG2PI mission. It is envisioned that this data management system will be a node in a larger network of tools created to support genotype/phenotype association studies. These data can be queried via a published application programming interface that will make all raw and derived data created by this project accessible in real time, or otherwise available for complete download. Having these data available it will be possible for breed associations to identify potentially lethal alleles in their sires and dams and manage their breeding decisions based on the probability of producing animals homozygous for lethal recessive alleles. In certain specific pairings, it can increase the probability of producing a successful pregnancy by 25%.

3. Expected outcomes & deliverables

Dr. Kalbfleisch will be collaborating with two additional scientists, Dr. Fiona McCarthy, and Dr. Elaine Norton DVM who will assist in annotating the genetic datasets with the appropriate phenotype data derived from the pathology reports. They will work with Dr. Kalbfleisch to build a web based data portal where animals may be identified by ontology terms derived from their respective pathology reports, and the variant data will be available in several formats, including annotated variant call format (VCF) files, and binary alignment (BAM) files that may be easily read via URL into commonly used browsers and tools such as the UCSC Genome Browser[6], the Broad Institutes Integrative Genomics Viewer[7], or command line tools such as Samtools[8]. The raw sequence data will be published to the NCBI Sequence Read Archive. The VCF files will contain annotations that may be used for filtering that identify novel variants, homozygous rare alleles, and aneuploidies where found. Putative lethal alleles will be run through the ENSEMBL Variant Effect Predictor[9] to annotate variants that occur in coding regions. Annotated summary VCF files of all alleles that are plausibly lethal will be identified, and available for download, or real time access.

Longer term, this will create a process, that can continue receiving, sequencing, and analyzing samples at a modest cost for breed associations to identify and report these costly alleles.

4. Qualifications of the project team

Dr. Kalbfleisch has nearly two decades of experience in analyzing, managing, and publishing high throughput genetic data. He worked for 8 years in industry with the CuraGen Corporation, and Genaissance Pharmaceuticals in building process management systems for high throughput genetic data management. While at the University of Louisville, he spun out a startup company, Intrepid Bioinformatics and built a web-based data management system supported by oracle for single nucleotide polymorphisms, and high throughput datasets. He currently manages the genetic data used for parentage and cattle introgression determination for the USYAK registry.

Dr. Fiona McCarthy and Dr. Elaine Norton DVM, have experience in the domains of creation and use of ontologies, and in genome wide association studies [10-14].

All other collaborators for this project are faculty members with either PhDs, DVMs, or both with vast cumulative experience in animal pathology, phenotyping, and in the study of the genetic basis of health and disease in farm animals.

5. Proposal timeline

In the first 3 months of the project, tissue samples and corresponding pathology reports will be shipped to the University of Kentucky where they will be collected and stored. The second three months of the project will be dedicated to the generation of sequence data, and the initiation of the data analysis. The final 6 months will be spent on the publication of the derived data, and the results they produce.

6. Engaging AG2P scientific communities & underrepresented groups

Missed breeding opportunities are expensive, especially so to smaller producers and underrepresented groups therein. We will work with the AG2P leadership and its scientific community to ensure that all stakeholders are educated about and trained on the use of this resource, and how it can improve their productivity.

Bibliography/References cited:

1. Hoff, J.L., et al., *Candidate lethal haplotypes and causal mutations in Angus cattle.* BMC Genomics, 2017. **18**(1): p. 799.
2. Heaton, M.P., et al., *Using diverse U.S. beef cattle genomes to identify missense mutations in EPAS1, a gene associated with pulmonary hypertension.* F1000Res, 2016. **5**: p. 2003.
3. Heaton, M.P., et al., *Using sheep genomes from diverse U.S. breeds to identify missense variants in genes affecting fecundity.* F1000Res, 2017. **6**: p. 1303.
4. Kalbfleisch, T.S., et al., *A SNP resource for studying North American moose.* F1000Res, 2018. **7**: p. 40.
5. Kalbfleisch, T., et al., *Using triallelic SNPs for determining parentage in North American yak (Bos grunniens) and estimating cattle (B. taurus) introgression.* F1000Res, 2020. **9**: p. 1096.
6. Karolchik, D., A.S. Hinrichs, and W.J. Kent, *The UCSC Genome Browser.* Curr Protoc Bioinformatics, 2012. **Chapter 1**: p. Unit1 4.
7. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.
8. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
9. McLaren, W., et al., *The Ensembl Variant Effect Predictor.* Genome Biol, 2016. **17**(1): p. 122.
10. McCarthy, F.M., et al., *Understanding animal viruses using the Gene Ontology.* Trends Microbiol, 2009. **17**(7): p. 328-35.
11. McCarthy, F.M., et al., *AgBase: supporting functional modeling in agricultural organisms.* Nucleic Acids Res, 2011. **39**(Database issue): p. D497-506.
12. Buza, T.J., et al., *Gene Ontology annotation quality analysis in model eukaryotes.* Nucleic Acids Res, 2008. **36**(2): p. e12.
13. Norton, E.M., et al., *Heritability of Recurrent Exertional Rhabdomyolysis in Standardbred and Thoroughbred Racehorses Derived From SNP Genotyping Data.* J Hered, 2016. **107**(6): p. 537-43.
14. Norton, E.M., et al., *Heritability of metabolic traits associated with equine metabolic syndrome in Welsh ponies and Morgan horses.* Equine Vet J, 2019. **51**(4): p. 475-480.