Title
The Pioneer 100 Horse Health Project: A Systems Biology Approach to Equine Precision Health Research

Author
Donnelly, Callum George

Publication Date
2022

Peer reviewed|Thesis/dissertation

The Pioneer 100 Horse Health Project: A Systems Biology Approach to Equine Precision Health Research

By

CALLUM GEORGE DONNELLY

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Integrative Pathobiology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Carrie Finno, Chair

Rebecca Bellone

Grace Mulcahy

Committee in Charge

2022

Table of Contents

Table of Figures and Tables

Dedication

I dedicate this thesis to all the matriarchs in my life. To my great grandmother who showed me that knowledge and kindness are life's greatest pursuits. To my grandmother who showed me the power of eagle-eyed observation. To my mother who showed me that you must be the change you want to see in the world. To my supervisor who showed me that science worth doing is worth doing well. Finally, to my wife, who showed me that life, like science, is better shared.

Abstract

Precision medicine, which is the application of high dimensional data to refine disease classification, diagnosis, and treatment, is the future of health care in the developed world. This is no less true for veterinary patients, including horses. Horses are a uniquely positioned species to apply precision medicine approaches, with a continually improving annotated genome and a number of syndromic diseases shared with humans. However, additional resources are needed to enhance the ability to move equine medicine into the precision age. This takes on three broad areas: improved phenotype data, expanded omic analysis and big data/computational approaches. This thesis addresses the areas of need to promote progress towards precision medicine through resource development and utilization to demonstrate the application in equine disease.

We begin this thesis by introducing, in Chapter 1, the Pioneer 100 Horse Health Project, a deep phenotype and multiomic resource developed specifically for use in precision medicine studies of equine health and disease. This is followed with Chapter 2 that details the application of this resource for improving the understanding of drivers of disease in Equine Metabolic Syndrome (EMS). Through deep endocrine phenotyping, married with plasma metabolomic and fecal microbiota, we establish a clear relationship between environment-host-microbiota factors that promote EMS. Next, in Chapter 3, we further investigate the role of environmental interaction and disease in the neurologic condition, equine neuroxaonal dystrophy (eNAD). Here, we evaluate the impact of vitamin E deficiency on axonal health in juvenile horses to refine the use of a previously validated biomarker, phosphorylated neurofilament heavy, in the supportive diagnosis of eNAD. We show that vitamin E

deficiency alone is enough to drive subclinical axonal damage, even if clinical eNAD does not ensue. Finally, in Chapter 4, we apply machine learning models to proteomic data sets for serum and cerebrospinal fluid (CSF) to discover disease informative biomarkers for equine neurodegenerative disease. Here, we demonstrate the use of two- and three-plex protein biomarkers that can accurately predict horses with eNAD from both neurologically normal horses, and those with the disease mimic condition, cervical vertebral compressive myelopathy.

The development of equine precision medicine resources, detailed in this thesis, are a key step forward in this discipline. To that end, availability of these data and methods through publication and repository submissions are an important step towards that goal. While validation in external cohorts will be necessary in the future, the data presented in Chapters two, three and four independently advances equine health and welfare.

Lastly, this thesis includes an addendum providing the details for biobank generation in the Functional Assembly of Animal Genomes (FAANG) consortium using two Thoroughbred stallions. Improvements to the annotation run parallel to the development of precision medicine approaches in the horses. By more precisely annotating the equine genome to include detailed information about regulatory element and tissue specific gene expression, we greatly enhance the resolution for health, disease and trait investigation in the horses.

Collectively this work demonstrates the both the approach and application of precision medicine in the horses and acts as a blueprint for future progress in the field.

Chapter 1: The Pioneer 100 Horse Health Project: A blueprint for a systems biology approach to veterinary precision medicine

Authors: Callum G. Donnelly[1], Sichong Peng[1], Veronika Rodriguez[1], Annee Nguyen[2], Kelly E. Knickelbein[3], Elizabeth Graham-Louie[4], Jessica M. Morgan[4], Rebecca R. Bellone[1,5], Myles Hammond[5], Nathan D. Price[6,7], Carrie J. Finno[1]

[1] Department of Population Health and Reproduction, School of Veterinary Medicine, University of California, Davis, CA, USA

[2] Department of Neurosurgery, School of Medicine, Duke University, Durham, NC, USA

[3] Department of Clinical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY, USA

[4] Department of Medicine and Epidemiology, School of Veterinary Medicine, University of California, Davis, CA, USA

[5] Veterinary Genetics Laboratory, University of California, Davis, CA, USA

[6] Thorne HealthTech, New York, NY, USA

[7] Institute for Systems Biology, University of Washington, Seattle, WA, USA

Abstract

Precision veterinary medicine relies on accurate phenotyping and extensive multi-omic data to make meaningful advances in the prevention, diagnosis, and treatment of disease in animals. We have established the first equine precision medicine initiative, the Pioneer 100 Horse Health Cohort, to enable enhanced understanding of genome to phenome relationships in the context of health and disease. This study is a prospective, longitudinal cohort study using a systems medicine framework for data collection. Uniquely, the population of 108 horses are

managed at a single academic institution, allowing for careful consideration of environmental conditions, access to consistent and comprehensive medical records, deep phenotype profiling and biospecimen sampling. For the first time, we demonstrate an aggregate n-of-one approach to veterinary precision medicine. This approach allows for meaningful discovery from smaller cohorts by increasing the depth of phenotype data, expanding both our phenotypic approaches and through repeated sampling. Additionally, it allows findings to straddle precision and personalized medicine. This study lays the groundwork for future veterinary precision medicine studies, where depth of data is preferential to size of the study cohort.

INTRODUCTION

The unveiling of the Precision Medicine Initiative (PMI) in 2015 forever altered the trajectory of human medical research and practice.[1] Under the PMI banner, the “All of Us Research Program” has become the cornerstone of the initiative, marrying longitudinal observational data, genomics, and medical records to better enable precision medicine goals on a broad scale. With the unique opportunity offered by this study also comes great challenges: difficulty capturing patient data, inconsistent phenotype designation and navigating disparate electronic medical databases. As an alternative to the traditional large cohort studies, the use of a systems medicine approach has emerged as a strategy to enable precision medicine studies with smaller populations.[2] By enhancing the level of data generated for a single participant, studies may enroll much smaller sample sizes, whilst maintaining meaningful and translatable discoveries. Indeed, insights have been gleaned from such multi-omic deep dives in even single individuals.[3,4] Combining such deep multi-omic data for wellness studies, the “aggregated N of 1s approach,” was validated in

the original Pioneer 100 study[5], which grew ultimately to around 5000 individuals and led to deep insights into microbiome effects on health[6,7,8,9], the interplay of genetics and lifestyle interventions[10,11], biological aging[12], and potential early warning signs for disease.[13,14,15,16] This Pioneer 100 study established the paradigm for integrating deep phenotype collection with multi-omic analysis within a single individual, and comparing these individual trajectories across a population. Thus, systems medicine studies straddle the border between personalized and precision medicine, enabling discoveries on an individual and population basis.

Within the field of veterinary medicine, there are examples of large-scale observational studies that have pivoted toward precision medicine, including the Golden Retriever Lifetime Study, 99 Lives Sequencing Consortium and the Dog Aging project.[17,18,19] These projects have focused on uncovering the environmental, genetic and lifestyle factors associated with the development of cancer and other age-related diseases. Similarly, these projects have taken a traditional large cohort framework, with decentralized phenotype and sample collections, coupled with a centralized repository and database. Veterinary precision medicine studies have the potential to both directly affect animal health and, importantly, they also have enormous translational potential for human health. While veterinary precision medicine is in early stages, strides have been made in small animals toward precision diagnosis and treatment of feline storage diseases, bleeding diatheses, and numerous forms of canine neoplasia.[20.21,22-23] To date, however, there have been no formalized precision medicine studies in equine patient populations.

The horse represents an ideal companion animal species for the establishment of a dedicated cohort-based precision medicine study. Horses are long lived, can be housed in groups, and share several relevant health traits with humans.[24] To this end, the Pioneer 100 Horse Health Project (P100HHP) was initiated in 2019 to characterize the genomic, environmental and physiomic drivers of health and disease in adult horses. The goal of the P100HHP is to intentionally emulate prior and ongoing human precision medicine studies and create the first dedicated cohort of research companion animals for use in precision medicine. This cohort study has some significant advantages as participating animals are institution owned, allowing for an unprecedented level of prospective phenotype characterization, multi-omic analysis, biorepository creation and medical record access. Additionally, this cohort of animals is not subject to the same level of participant dropout or inherent variation of phenotype designation that can be found when large numbers of observers and owners are involved in a single cohort study.

The objective of this study is to describe the population characteristics for the P100HHP cohort. Further, we lay the road map for use of deep phenotyping and biorepository creation for future multi-omic analyses. Lastly, we report genomic variation within the P100HHP cohort.

RESULTS

Cohort Description

Since October 2019, 108 horses have been enrolled into the P100HHP cohort, with enrollment ongoing to maintain approximately 100 horses in the study. Age was used as the primary selection criteria for entering the study (*Methods;* **Fig. 1.1**), with

a median age of 12 years (yrs) for females (range 2-18 yrs), and a median age of 9 yrs for males (range 2-18 yrs). Horses included in the cohort have been resident at UC Davis Center for Equine Health (UCD-CEH) for a median of four years (range 0-17 yrs) at the time of inclusion in the study. Age and time in residence at UCD-CEH were positively correlated (R=0.60, P=6.08 x 10^{-12}).

Included in this cohort are 60 females (entire N=52, ovariectomized (OVX) N=8) and 48 males (entire N=3, orchiectomized= 45). Breed was a secondary selection criterion, with seven breed groups represented in the P100HHP cohort and encompassing 20 recognized breeds (**Fig. 1.2***)*. Compared to the most recent census of horse breeds present in the United States, the proportions of breeds represented did not differ significantly to the general horse population (z= -1.47 P=0.14).[25] The cohort does not capture all highly abundant breeds in the United States, however, it does account for all the major clades in domestic horses except for the most primitive (i.e. Nordic breeds and ponies).[26]

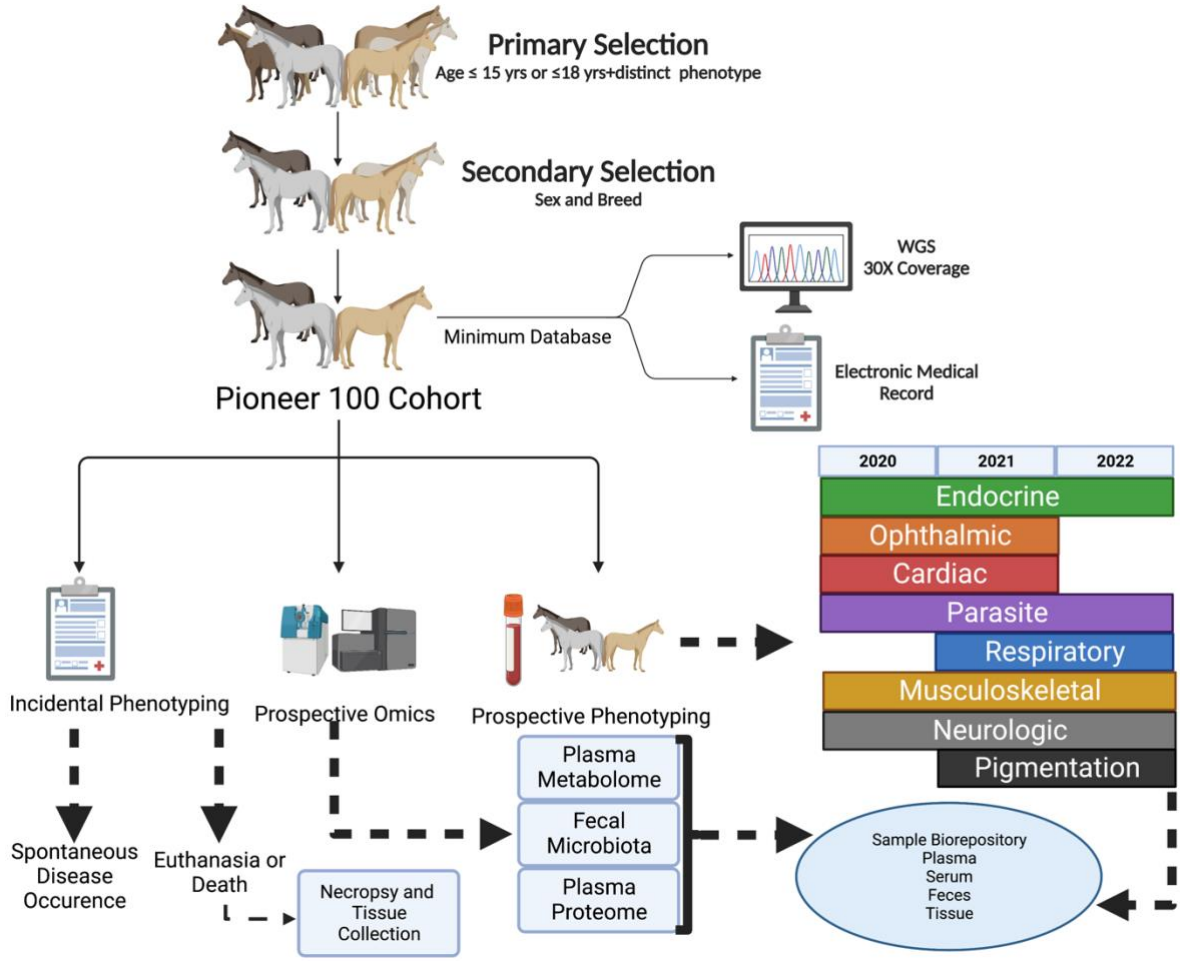

Figure 1.1 Schematic overview of the Pioneer 100 Horse Health Cohort. Animals were actively selected into the cohort from the resident population of horses at the UC Davis Center for Equine Health. Animals were required to meet age, breed, and sex ratio selection prior to entry into the study. The minimum amount of data collection for animals included in the study requires high coverage whole genome sequence (WGS) and an accessible medical record. Animals undergo extensive prospective phenotyping by body system and sample collection for biorepository creation. Disease phenotypes that occur spontaneously or were know at entry into the study are also documented. Figure created with Biorender.

Horses are managed in dry lots, with limited access to growing forage. This is typical of horse management in California, and many semi-arid areas of the United States. Horse diets primarily consist of orchard grass (*Dactylis glomerata*) and/or alfalfa/lucerne (*Medicago sativa)* hay sourced from Northern California and Oregon, with sporadic and infrequent use of supplementary processed grain-based feed stuffs. Except for intact males, horses are maintained in herds, ranging from two to

twenty horses per enclosure. All horses are vaccinated annually against *Clostridium tetani*, eastern equine encephalitis virus, western equine encephalitis virus, equine influenza virus, equine herpes virus 1&4, and rabies virus. Horses receive annual dental occlusal surface adjustment by licensed veterinarians. As research and teaching horses, all horses in this study are also used as needed for projects not related to the P100HHP concurrently with the P100HHP.

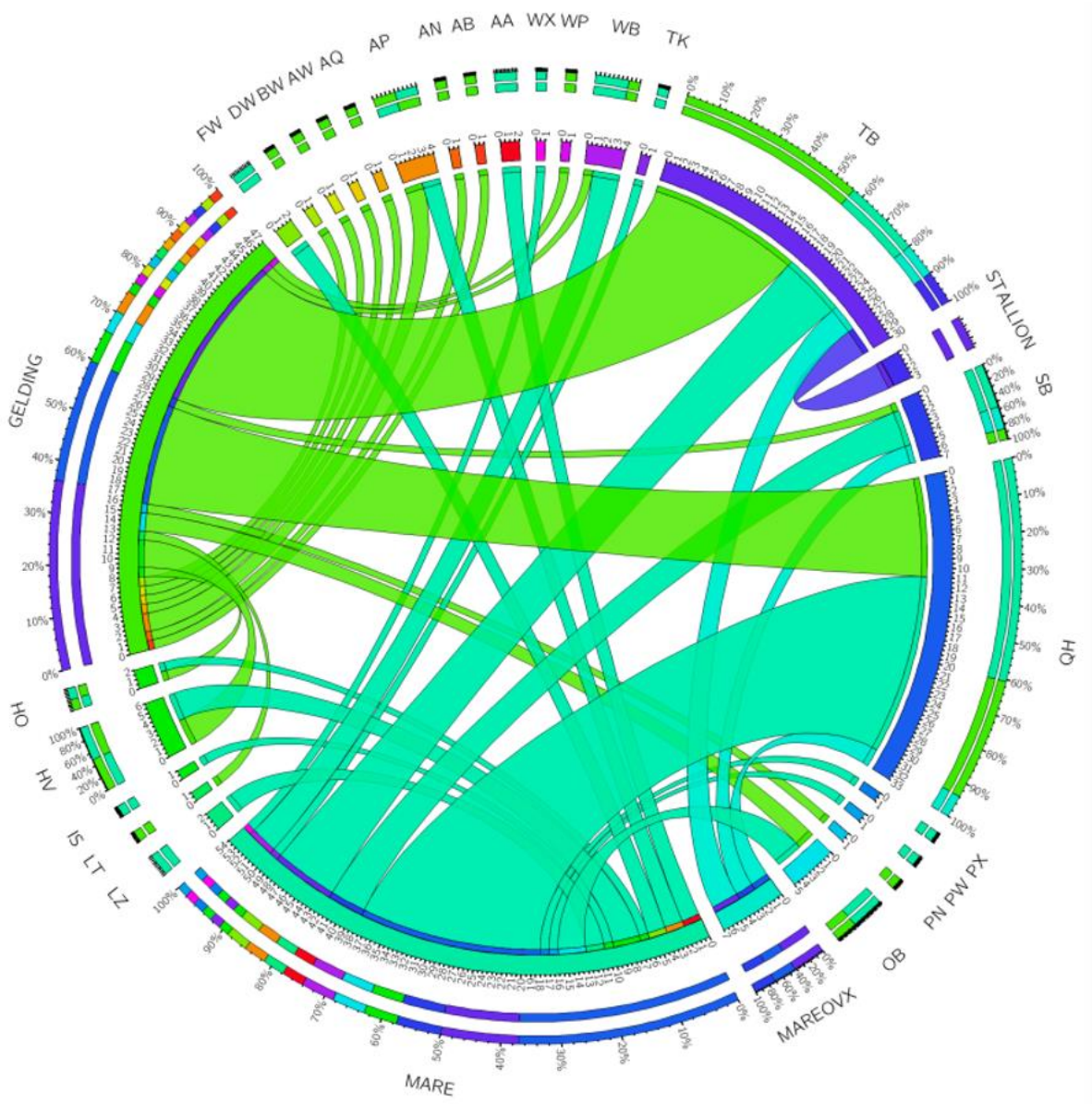

Figure 1.2 Circos plot detailing the breed and sex composition of the Pioneer 100 Horse Health Cohort. The outside tracks show the proportion (%) of either breed or sex (Mare, Gelding, MareOVX and Stallion). The inside track is the count of each breed or sex. Ribbons link related sets of data (i.e. number of each breed per sex), with the size of ribbon proportionate to the size of the related data (i.e. more animals result in a bigger ribbon). Abbreviations: AA-Appaloosa , AB-Arabian, AN-Andalusian, AP- American Paint, AQ- Appendix Quarter Horse, AW-American Warmblood, BW-Belgian Warmblood, Dutch Warmblood- Dutch Warmblood, FW- , HO-Holsteiner, HV-Hanoverian, IS-Irish Sport Horse, LT-Lusitano, LZ-Lippizaner, OB-Oldenberg, PN- Percheron, PW-, PX-Percheron Cross, QH- Quarter Horse, SB-Standardbred, TB- Thoroughbred, TK- Trakehner, WB-Warmblood, WP-Wesphalian, WX-Warmblood cross. Figure created using Circos V0.63-9.

Genome Data

Genomic Variants

Individual breeds were condensed to 7 breed groups: Quarter Horse type, Thoroughbred, Standardbred, Warmblood, Iberian, Draft and Arabian. Genotypic variants (single nucleotide variations; SNVs) were inspected by principal component analysis using these breed groups (**Fig. 1.3).** Breed groupings aligned well with genomic variants. As expected, Thoroughbred horses clustered most closely, with Quarters Horses having the most diversity.[26] This was confirmed by analyzing runs of homozygosity (ROH) for both breeds, with Thoroughbreds having a mean across genome ROH of 148,464 $\pm$ 24,497 base pairs and Quarter Horses having a mean ROH of 81,141 $\pm$ 11,509 base pairs (**Fig. 1.4**). Data visualization allowed for the confirmation of breed for one horse (526; Thoroughbred), previously of undocumented breed. It also demonstrated that this approach is able to capture individuals that may be comprised of two breeds (e.g., 1493 and 1494).

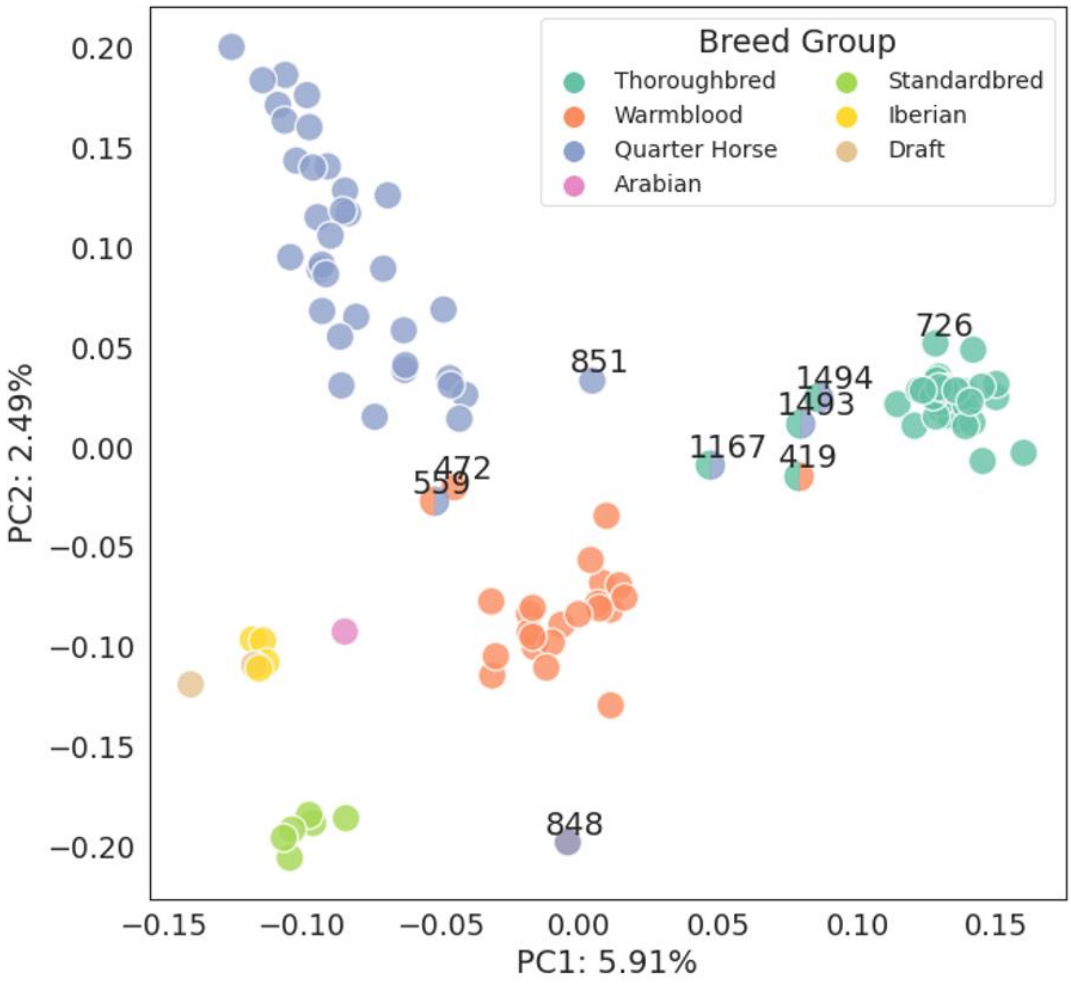

Figure 1.3 Principal Component Analysis of genomic variants by breed groups. Clustering of individuals agreed strongly with breed group assignment. Visual assessment of PCA confirmed animals that were comprised of two breed groups (419, 559, 1167, 1493, 1494). Additionally, one individual that had previously been of unknown breed was confirmed as a Thoroughbred (726). This animal was assigned as a suspected Warmblood based on available records; however, based on the genomic profile and phenotype, horse 726 fit best into the Thoroughbred group.

Whole genome data were screened for known disease-causing variants.[27] Seven disease-causing variants were detected in this population *(***Supplementary Table 1 S1***)*. Except for one animal that was homozygous for the *MYH1* variant and one for the *TPRM1* variant, all other animals with variants detected were heterozygous. Only one animal was heterozygous for multiple variants (Quarter Horse; *PPIB* and *MYH1*). Variants were detected at expected allele frequencies and in breeds previously

reported to harbor the variants.[28] An additional variant for gait, *DMRT3,* was also screened for in this cohort. As expected, all horses in the Standardbred breed group were homozygous for the variant allele.[29] Additionally, two animals from the Quarter Horse group were heterozygotes for this trait.

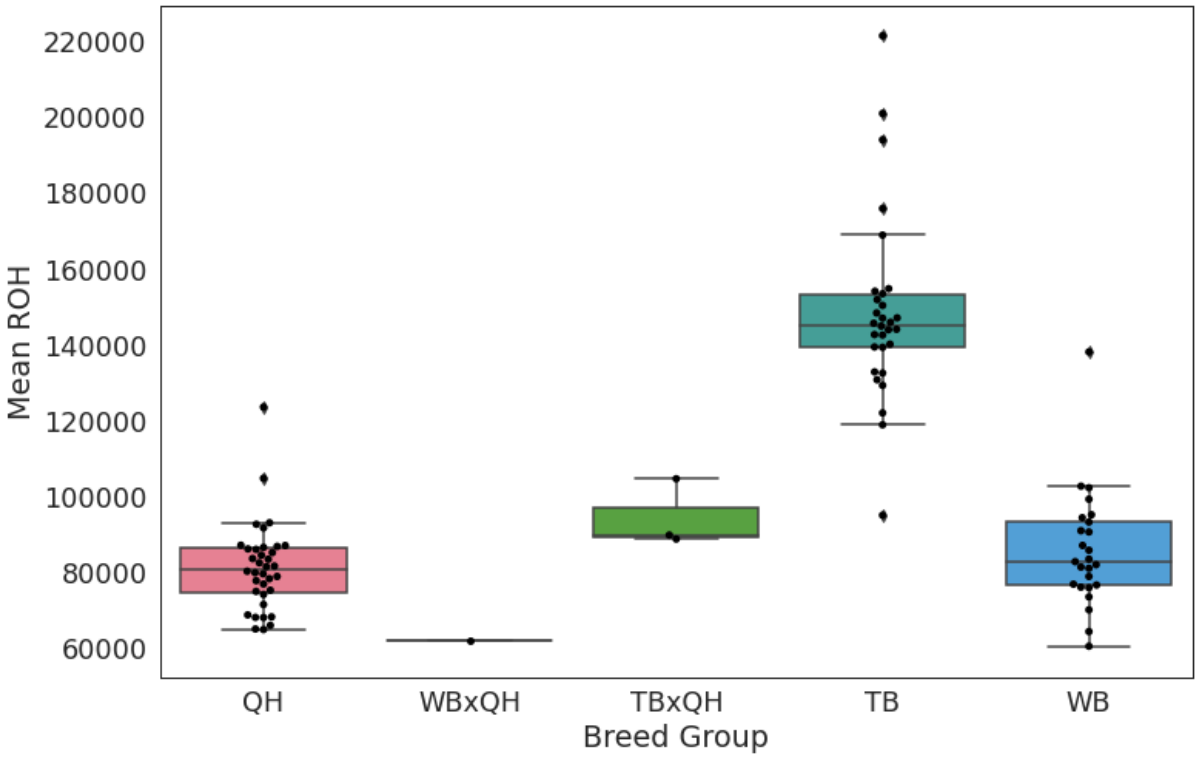

Figure 1.4 Box and Whisker plot of runs of homozygosity (ROH). Horses from the Thoroughbred breed group had the highest across genome ROH (mean ± SD; 148,464 ± 24,497 bp). The other two major groups, Quarter Horse (81,141 ± 11,509 bp*)* and Warmblood (85,629 ± 15,418 bp), had approximately half the ROH length as Thoroughbreds. The small number of Thoroughbred x Quarter Horse crossed horses demonstrated an intermediary ROH. Abbreviations; QH-Quarter Horse, WBxQH-Warmblood cross Quarter horse, TB-Thoroughbred, TBxQH- Thoroughbred cross Quarter Horse, WB-Warmblood, ROH-Runs of Homozygosity; bp-base pairs

Quarter horses were screened for the missense variant in glycogen branching enzyme-one (*GBE1),* responsible for glycogen branching enzyme deficiency (GBED). Based on whole genome sequencing data analysis, no individuals were identified to have this variant. Homozygotes for this trait would not be expected as the trait is homozygous lethal in neonatal and juvenile horses.[30] However, one horse

in the cohort was a known heterozygote, having been previously genotyped, as well as producing a clinically affected homozygote foal. Yet, due to the high GC content in this region, there was a sequencing bias that did not allow for the detection of heterozygotes from whole-genome sequences. Thus, the limitations of genotyping via whole-genome sequencing were apparent for this particular disease variant. Genotypes of the known carrier and its affected foal were independently verified by the UC Davis Veterinary Genetics Laboratory.

Table 1 S1: Disease associated variants detected by GATK genotyping of whole genome sequence data from the Pioneer 100 Cohort.

Phenotype	**Gene**	**# Heterozygotes *(Breeds)***	**# Homozygotes *(Breed)***	**Allele Frequency**
Cerebellar abiotrophy	Target of EGR1 (*TOE1*)	N=1 *Arabian*	NA	0.0045
Congenital Stationary Night Blindness	Transient receptor potential cation channel, subfamily M, member 1 (*TRPM1)*	N=1 Appaloosa	N=1 Appaloosa	0.014
Fragile Foal Syndrome Type I	*Procollagen-lysine, 2-oxoglutarate 5-dioxygenase 1 (PLOD1)*	N=2 *Thoroughbred, Hanoverian*	NA	0.0099
Gaitedness	*Doublesex and mab-3 related transcription factor 3 (DMRT3)*	N=2 *Quarter Horse*	N=6 *Standardbred*	0.064
Hereditary Equine Regional Dermal Asthenia	Peptidyl-prolyl cis-trans isomerase B (*PPIB)*	*N=4 Quarter Horse*	NA	0.018
Ileocolonic aganglionosis	Endothelin receptor B (*EDNRB)*	N=2 *American Paint Horse*	NA	0.0099

Myosin Heavy Chain Myopathy	Myosin Heavy Chain-1 (*MYH1)*	N=5 *Quarter Horse*	N=1 *Quarter Horse*	0.032
Polysaccharide Storage Myopathy 1	Glycogen synthase 1 (*GYS1)*	N=5 *American Paint Horse, Quarter Horse, Draft*	NA	0.023

Phenotype Data

Retrospective Disease Phenotypes

Electronic medical data and historical data associated with each horse donation was used to generate retrospective health related phenomes for each horse (see *Methods*). The mean number of discrete disease phenotypes by body system was 1.34 ± 1.18 and ranged from 0 to 5 per animal. The most encountered body system affected by a discrete process was musculoskeletal, with 37 horses having a total of 49 disease processes. Retrospective phenotypes by body system are summarized in **Supplementary Fig. 1 S1 and Supplementary Table1 S2.**

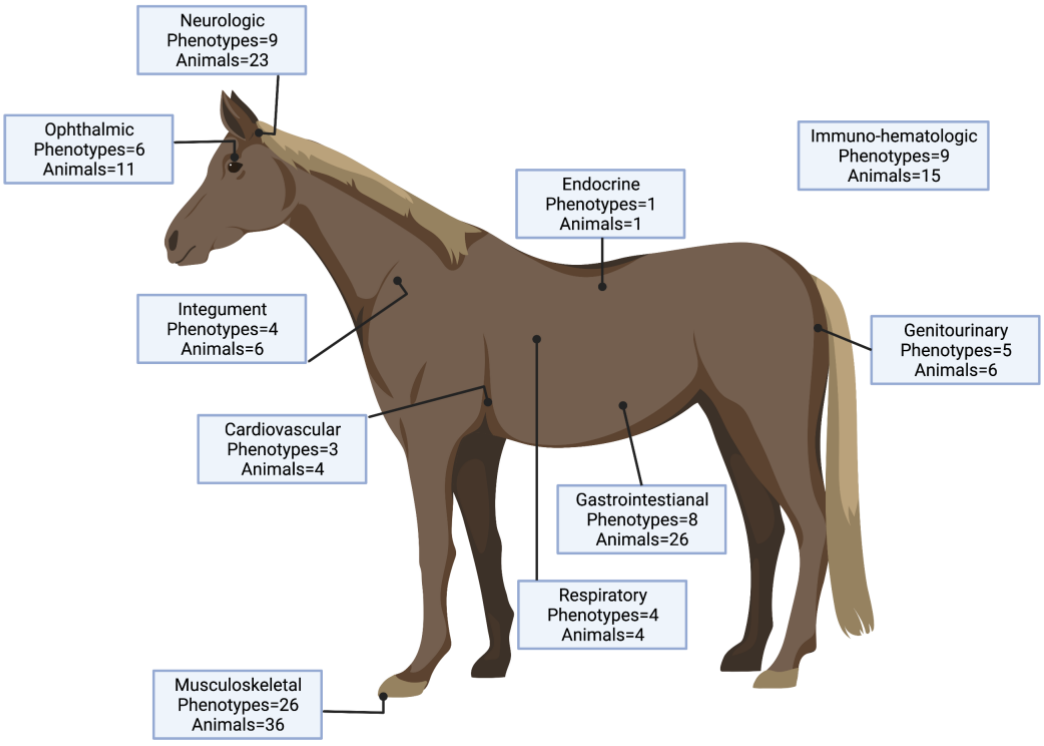

Figure 1 S1 Schematic representation of retrospective disease phenotypes by body system. Animals may have more than one phenotype per body system and accounts for more phenotypes than animals in four body systems. These four body systems (musculoskeletal, gastrointestinal, neurologic and immune-hematologic) account for the majority of retrospective phenotypes. Animals included in this cohort may have undiagnosed phenotypes or have suspected phenotypes that could not be corroborated by the medical record. Figure created with Biorender.

Supplementary Table 1 S2: Retrospective phenotypes derived from medical record mining of the Pioneer 100 cohort. Phenotypes were only assigned where clinical signs for a disease are unambiguous and/or where corroborating clinical data was present including (but not limited to) diagnostic imaging, complete blood count, biochemistry, bacterial culture, and immunodiagnostics. Where a phenotype represented a process across two systems in one animal, it is reflected in summaries for both systems (e.g. uveitis in ophthalmic and immunologic).

Musculoskeletal	**Number of Animals**
Laminitis	4
Patellar fracture	1
Lateral trochlear ridge of femur fracture	1
3rd Phalangeal fracture	1
Degenerative Suspensory ligament desmopathy	2
Distal interphalangeal osteoarthritis	1
Tenosynovitis	1
Navicular syndrome	7
Deep digital flexor tendinopathy/itis	2
Carpal osteoarthritis	2
Osteochondrosis mid-sagittal 3rd metacarpal	1
Sequestrum	1
Superficial digital flexor tendinopathy	1
Deep digital flexor tendon sheath tenosynovitis	1
Osteochondrosis lateral trochlear ridge of talus	1
Intertarsal osteoarthritis	1
Pedal osteitis	1
Overriding dorsal spinous process	2
Septic arthritis	1
Superficial digital flexor tendon subluxation	1
Osteoarthritis C5-6/C6-7	3
Distal inter-phalangeal flexural deformity	1
Osteochondrosis distal intermediate ridge of tibia	2
Sub solar abscess	1
Osteoarthritis proximal inter-phalangeal joint	1
Vitamin E Dependent Myopathy	3
Gastrointestinal	
Large colon impaction	2

Esophageal obstruction	1
Equine gastric ulcer syndrome	1
Colic (unspecified, medically resolved)	20
Right dorsal displacement of colon	1
Eosinophilic colitis	1
Chronic diarrhea	1
Nephro-splenic entrapment	1
Neurologic	
Ataxia (unspecified)	4
Ataxia (equine neuroaxonal dystrophy likely)	8
Ataxia (compressive myelopathy likely)	3
Equine neuroaxonal dystrophy (confirmed)	1
Cervical vertebral compressive myelopathy	3
Shivers	1
Tinnitus	1
Narcolepsy	1
Trigeminal neuralgia	1
Ophthalmic	
Ulcerative keratitis	1
Non-ulcerative keratitis	3
Corpora nigra cyst	1
Uveitis	3
Cataract	2
Phthisis bulbi	1
Genitourinary	
Metritis	1
Endometritis	2
Cystitis	1
Recurrent pregnancy loss	1
Hemospermia	1
Integument	
Urticaria	2
Squamous cell carcinoma	1
Coronary bad desmitis	1
Chronic hypertrophy of frog	1
Immune-hematologic	
Thrombophlebitis	2
Habronemiasis	1
Corynebacterium pseudotuberculosis infection	4
Urticaria	3
Atypical Equine Thrombasthenia	1
Uveitis	3
Vaccine adverse event	4
Ulcerative lymphangitis	1
Systemic inflammatory response syndrome	2
Respiratory	
Guttural pouch empyema	1
Sinusitis	1

Fourth brachial arch defect	1
Bronchopneumonia	1
Cardiovascular	
Thrombophlebitis	2
Aortic insufficiency	1
Aorto-cardiac fistula	1
Endocrine	
Pars pituitary intermedia dysfunction	1

Incidental Terminal Phenotypes

To date, nine horses have left the study due to terminal disease. Two horses were euthanized as a result of worsening neurologic disease, with a diagnosis of equine neuroaxonal dystrophy (eNAD) in one horse and compressive cervical myelopathy in the other. One horse was euthanized due to fulminant heart failure, secondary to a space occupying thymic carcinoma. Two horses were euthanized due to worsening laminitis. One horse was euthanized for persistent detrusor dysfunction of the urinary bladder. One horse was euthanized due to fulminant bilateral pelvic limb degenerative suspensory ligament desmitis. One horse received a fatal head injury while at pasture. The final horse had severe incarceration of the jejunum and ileum secondary to a mesenteric rent and was euthanized due to the inoperability of the damaged small intestine. All diagnoses were confirmed by complete necropsy performed by board certified veterinary anatomic pathologists. All animals had extensive tissue and biofluid collection shortly after euthanasia or death (**Supplementary Table 1 S3***)*.

Table 1 S3: Standard list of tissues collected at time of necropsy for horses enrolled in the Pioneer 100 cohort.

<u>Musculoskeletal</u>
Cartilage (Tarso-crural)
Coronary periople (thoracic)
Deep digital flexor tendon (thoracic)
Superficial digital flexor tendon (thoracic)
Suspensory ligament (thoracic)
Gluteal muscle

<u>Integument</u>
Coronary periople (thoracic)
Lamina (thoracic)
Neck skin
Abdominal adipose tissue
Nuchal adipose tissue

<u>Cardiovascular</u>
Left atrium free wall
Left ventricle free wall
Right atrium free wall
Right ventricle free wall

<u>Respiratory</u>
Left lung

<u>Endocrine</u>
Thyroid
Adrenal cortex (left ot right)
Adrenal medulla (left ot right)
Pancreas

<u>Urogenital</u>
Kidney cortex (left)
Kidney medulla (left)
Urinary bladder
Testis (if applicable)
Epididymis (if applicable)
Ovary (if applicable)
Uterus (if applicable)

<u>Gastrointestinal</u>
Liver (left lobe)
Spleen
Esophagus
Stomach
Duodenum
Jejunum
Ileum
Cecum
Right dorsal colon
Right ventral colon
Left dorsal colon
Left ventral colon
Small colon

<u>Nervous System</u>
Dura mater
Parietal cortex
Occipital cortex
Frontal cortex
Temporal cortex
Corpus callosum
Thalamus
Hypothalamus
Pituitary
Pons
Cerebellar vermis
Cerebellum lateral hemisphere
Spinal cord
- C1
- C6
- T8
- L1

Sciatic nerve
Dorsal root ganglia

<u>Eye</u>
Cornea
Retina

<u>Fluids</u>
Plasma (EDTA & heparinized)
Serum
Cerebrospinal fluid
Urine

Prospective Phenotyping

All animals have been enrolled in an extensive prospective phenotyping program (**Fig, 1.1**)**.** Phenotypes are intentionally collected by body system. Incidental prospective phenotypes are collected from the medical record in a similar way to retrospective phenotypes. Prospective phenotype data is collected on a rolling basis over the observation period. In tandem with phenotyping, a large biorepository of serum, plasma, cerebrospinal fluid, peripheral blood mononuclear cells, urine, and feces have been collected from participating horses. Primary collections occur twice per year, with additional samples collected as directed by phenotyping efforts.

Repeated sampling is a cornerstone the n-of-one approach allowing for improved resolution of phenotype designation as well as to capture individual variation over time. To demonstrate the utility of this approach, we report plasma creatine kinase (CK) concentrations in horses with polysaccharide storage myopathy type-1 (PSSM1). Four individuals from the Quarter Horse breed group were heterozygous for the *GYS1* variant (c.926G>A, p.R309H) that is causative for PSSM1. Using four age, sex, and breed group matched controls from the P100HHP cohort without the *GYS1* variant, prospective and retrospective CK values were compared. CK values were above the reference range (>278 μkat/L) in at least one timepoint in five out of eight animals. However, when all timepoints were considered, 37.5% were above the reference range in animals with the *GYS1* variant as compared to 15.4% of samples from horses without the variant (χ^2= 1.7 P=0.1). Repeated sampling improved the specificity from 50% to 87.5% for this biomarker in association with the causative variant (**Supplementary Fig. 1 S2***)*. This highlights the importance of longitudinal

depth to improve phenotype resolution when a smaller cohort is employed for precision medicine.

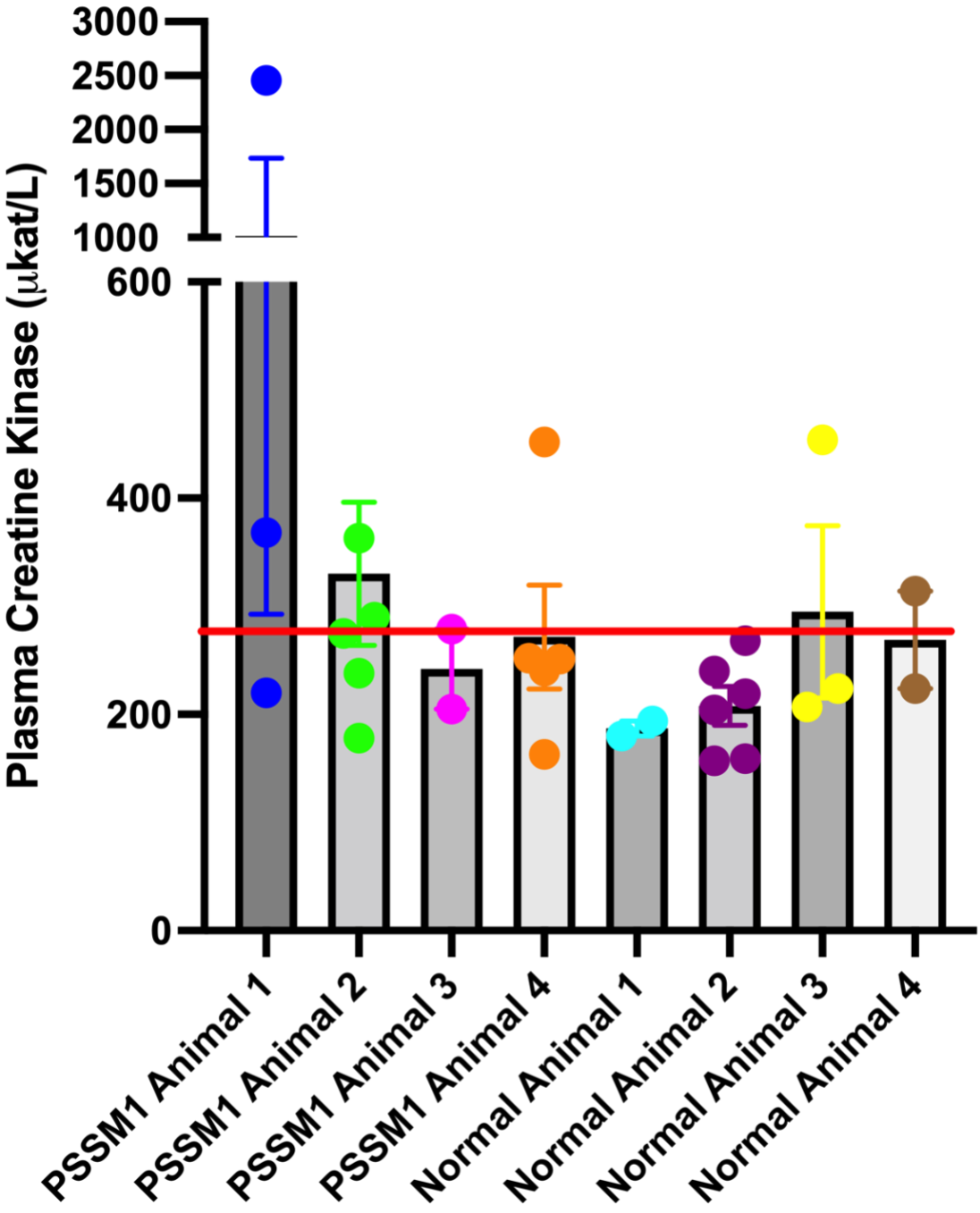

Figure 1 S2 Individual values ± SEM for Creatine Kinase (CK) a muscle leakage enzyme that may be elevated in animals with Polysaccharide Storage myopathy Type 1 (PSSM1). All available prospective and retrospective CK were compared in four Quarter Horses with (*GYS1*:c.926G>A, p.R309H) and compared to four age,

breed, sexed matched horses without the variant. The red line denotes the upper laboratory reference limit (278 μkat/L).

Discussion

This is the first publication to describe the population structure and genomic data for the P100HHP cohort. This project has successfully demonstrated the implementation of a dedicated herd of research horses for use in precision medicine studies. Precision medicine requires high-level integration of genomic, omic data (physiome), environmental exposures (exposome) and exhaustive phenotyping (phenome). It is only with this multi-part integration that veterinary research will move to enable the implementation of precision medicine for horses.

Making whole genome data available to the equine research community is of great importance for the collective advancement towards precision medicine in horses. Use of whole-genome sequencing approaches to veterinary genomics has become increasingly accepted as the price of sequencing has become more affordable.[32] Approximately 500 equine genomes have been made publicly available to date.[33,33] This contrasts with the human equivalent, where hundreds of thousands of genomes (or whole exome) datasets are available for use by the research community.[34] Our data enhances the already publicly available dataset by adding an additional 108 genomes from 20 breeds. Additionally, ours is the first study of this magnitude to be able to appropriately control for environmental influences. Further strengths of this study are that each genome will be associated with a high degree of phenome data and the majority of animals in this study are available for repeated and continued phenotyping. This contrasts with the currently available genomes, in which the associated metadata is primarily restricted to breed and sex and a single disease or

trait phenotype. Dedicated companion public phenome data is very rarely available to veterinary genomic researchers. This study lays the groundwork for leveraging smaller sample sizes with deep phenotype data that will be made available in subsequent publications.

This study has demonstrated the utility of medical records as a source of phenomic data. The use of electronic medical records has previously been proposed as a source of data for precision veterinary medicine.[35] However, implementation of medical record data is hampered by the array of medical record software, number of users inputting data and variable health and disease terminology used to annotate records. As such, hand sorting remains a necessary step when using this data, as it is in the current study. This approach ensures that the data interpreted as a disease phenotype is stringent. In previous canine prospective cohort studies, data capture for disease phenotypes relies on the interpretation of various attending clinicians, which engenders an inherent level of variability and possibility for misclassification.[36] All retrospective phenotypes in the P100HHP study were assigned by a single expert observer interpreting the medical record and only assigning a phenotype when clinical interpretations were corroborated with additional clinical data or imaging. While taking this approach ensures a high level of stringency, it likely also results in a degree of type II error, due to poor record annotation or lack of corroborating evidence for assigning a phenotype. To address this potential error in phenotyping, an intensive prospective phenotype effort is ongoing to catalogue phenotypes on a body systems basis in this cohort.

A strength of the P100HHP study is that all animals are maintained at a single facility with largely the same management for all horses. Thus, very similar exposure factors are experienced by all enrolled horses. The amount of time that an animal has been in the herd is correlated with the age of the animal. This is important, as age and exposure time are relatively proportionate, and as such effects of age, exposure and age/exposure interaction may be assessed in future studies using this data set. While the consistency of environmental exposure factors is valuable, it is also a caveat to the interpretation of this data when used in integrated analyses across cohorts. The dry lot management, typical of the Western United States, is distinct from that of other geographic areas. Particularly, the availability of pasture grazing opportunity is limited in this population and therefore typical grazing associated risk factors (e.g., parasites) should be viewed in light of this. Similarly, other environmental determinants of disease phenotypes, such as air quality during wildfire events, are also tied to the geographic area, time of year and/or season. This study addresses issues related to geotemporal exposure through repeated sampling across time and seasons to allow for future analysis of these interrelated factors.

The cohort composition was constructed primarily to include middle-aged animals that would remain in the study for its foreseeable duration. Coupled with this is a diverse set of breeds, that was intentionally selected to be representative of the broader United States horse population. While incorporating various breeds may reduce the power for discovery of binary health and disease associated variants within individual breeds, it does present opportunities. Namely, for complex traits including parasite susceptibility, metabolic regulation, endocrinopathies and others that occur on a continuum [37,38], we have the unique opportunity to look deeply within

and across breeds and time in a highly controlled environment. This will be particularly important when this project reaches maturity and integrates additional omic data with the current genomic and phenomic data. Additionally, to advance precision medicine findings that can be adopted widely in clinical equine practice, investigations will need to adopt breed-specific and breed non-specific interpretations of integrated data to further classify health and disease phenotypes. Therefore, while using an across breed cohort has inherent challenges, it also has unique benefits that are unmet by evaluating a single breed.

Genomic population structure aligned well with assigned breed group in this study. Analysis of genomic traits must take population structure into account when using whole genome data.[24] Within the P100HHP cohort, there is inherent, expected, and demonstrable population structure. In past equine genomic investigations, population structure introduced by multiple breeds would have been viewed only as a limitation. However, with the advent of more sophisticated analysis tools that account for both structure and linkage disequilibrium, this may be overcome.[39] Use of across breed analysis has the potential to enhance the ability to investigate complex disease states by leveraging larger data sets and previous genome wide associations (GWA).[40] Analysis across GWA studies is a key step to the development of precision medicine tools, particularly the refinement of disease phenotype designation based on the attributable genetic architecture.[39]

Genotyping pipelines performed well for single nucleotide and small indel variants that are associated with diseases in horses. However, this study also highlights the importance of considering strand bias in variant calling for short read whole genome

sequence data. The causal nonsense variant for GBED[29] lies within a GC rich locus, and no heterozygotes were called in this study. This is despite there being a known heterozygote within the herd. Close inspection of the region demonstrated that good coverage was maintained, with an average of 20x. However, the strand bias for this sample was 19:1 for reference and alternate, respectively. Allele bias balance correction has not yet been evaluated for equine genomes. With improved annotation of the equine genome, techniques that allow for allele bias correction may abrogate false negative genotyping calls.[41] Additionally, this challenge indicates that high coverage sequencing is likely still necessary in precision medicine investigations to resolve genotyping errors when using model species, including the horse. Canine precision medicine cohorts have used low pass sequencing and imputation to generate whole genome data sets.[17,19] The advantage that these studies have over the current is a larger sample size. However, it has been suggested that, when sequencing depth falls below 2X, sample size alone will not improve the accuracy for GWA.[42] Thus, a highly phenotyped cohort, with high coverage whole-genome sequencing, such as the P100HHP, may provide a necessary resource to enable larger studies with lower coverage genome sequencing.

In conclusion, we report here on the development of the first dedicated single population of horses maintained for the study of precision medicine. The addition of 108 genomes to the already publicly available equine genomes will enhance the ability of researchers to investigate genomic traits on a broader population basis in horses. Additionally, we lay the road map for the integration of phenomic and physiomic data in a systems biology framework to enable precision medicine

discoveries. Subsequent integrated multidimensional data sets will augment the available genomic data, enhancing our ability to investigate the complex interaction between genes, individuals and the environment for health and disease traits in horses.

Methods

Animal Selection

All horses are resident teaching and research animals housed at the Center for Equine Health at the University of California, Davis. All procedures for this study were approved by the Institutional Animal Care and Use Committee (protocol #21700). Animals were selected into the cohort if they were less than 15 years old at the commencement of the study. An exception was made for animals that had unique or informative phenotypes if these animals were also less than or equal to 18 years of age. Time in residence at the Center was not used as a selection criterion, and therefore recently acquired horses were also included. Enrollment into the study is currently open, such that the total number of horses remains approximately 100 living individual horses. Following age, the second criteria for selection was sex, with the intent to enroll females: males at approximately 1:1. The final horses included were then hand selected from the available horses that met age and sex requirements to reflect the horse population of the United States.

Whole-Genome Sequencing

Venous blood was acquired by jugular venipuncture into EDTA vacutainers. Whole blood was used to isolate DNA using the Promega Wizard Genomic DNA Kit. Concentration and ds260/280 absorbance were measured using a Qiagen Qiexpert.

A minimum concentration of 50 ng/μL of DNA with a 260/280 ratio of 1.8-1.9 was submitted for whole genome sequencing. DNA fragmentation was evaluated by gel electrophoresis using 1.5% agarose, with an adjusted concentration of 50 ng/μL for each animal. Libraries were prepared by Novogene. Libraries were sequenced on an Illumina NovaSeq with a 150bp pair-end reads with a target depth of 20X.

Whole-Genome Data Processing

Raw sequences were trimmed to remove adapters and low-quality bases using TrimGalore[43] and Cutadapt[44] and subsequently aligned to reference genome, EquCab3.0[45], using BWA MEM[46] with default parameters. Aligned reads were filtered using SAMTools[47] to remove duplicates and multimapping reads.

Whole-Genome Population Structure

Variants were called using GATK (v4.1.9.0) HaplotypeCaller and GenotypeGVCFs following best practice for germline variant calling.[48] Called variants were filtered using VCFtools and GATK VariantFiltration (v4.1.9.0) to retain only single nucleotide variants with depth-normalized quality scores over 2 (QD > 2) and minor allele frequencies over 0.1 (MAF > 0.1).[49,50] Final variant set was then used to calculate eigenvectors and eigenvalues for all samples using PLINK(v1.90).[51] The same variant set was also used to estimate runs of homozygosity (ROH) using BCFtools.[52] Results of principal component analysis (PCA) and ROH were visualized using matplotlib and seaborn.[53,54]

Pathogenic Variant Frequencies

A list of known pathogenic variants were collected from the Online Mendelian Inheritance in Animals (OMIA) database.[27] EquCab2.0 coordinates were converted to EquCab3.0 coordinates using NCBI Genome Remapping Service (https://www.ncbi.nlm.nih.gov/genome/tools/remap). All samples were then genotyped at these loci using GATK HaplotypeCaller (v4.1.9.0) and results manually confirmed in IGV.[35] Large InDels were confirmed by comparing average coverage of these loci to background coverage (randomly sampled loci across whole genome).

Medical Record Phenotype

Medical records are housed in a proprietary electronic medical record system; Veterinary Medical & Administrative Computer System (VMACS). The electronic record software is built, maintained, and housed within the University of California, Davis, School of Veterinary Medicine. Herd management related data is recorded using the commercial software package Wise Option (Wise Option Inc, Frisco, TX). Clinical record data was hand annotated for disease phenotypes that have occurred prior to the enrollment of the animal in the study. Disease phenotypes were only included if a definitive diagnosis was reached and could be corroborated through laboratory measure or imaging. If a disease phenotype had more than one recorded event, it was regarded as a single phenotype. All clinical disease phenotypes were assigned by one experienced investigator (CGD).

Prospective Phenotyping

Prospective phenotyping is summarized in **Supplementary Fig. 1 S3**. Phenotyping pipelines were based on those developed by the International Mouse Phenotyping

Consortium.[55] Detailed phenotype data continues to be collected and is not presented here.

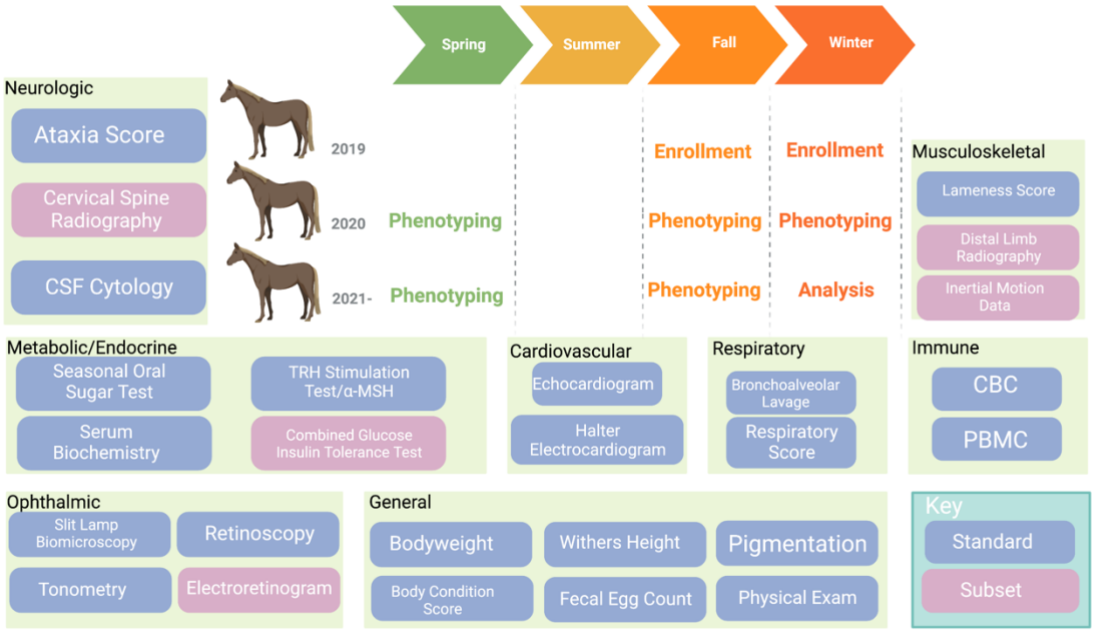

Figure 1 S3 Schematic representation of prospective deep phenotype data collection. Phenotyping by body system occurs on a rolling basis. Selection of phenotypes was based on protocols developed by the International Mouse Phenotyping Consortium.[54] Additionally prospective phenotype was informed by the availability of medical record derived phenotypes to improve body system phenotype data that were underrepresented (i.e., endocrine). Ongoing phenotype efforts are a source of future collaborative opportunities for the Pioneer 100 Horse Health Project. Abbreviations: CSF-Cerebrospinal Fluid; α-MSH- α-Melanocyte Stimulating Hormone; TRH-Thyrotropin Releasing Hormone; PBMC-Peripheral Blood Mononuclear Cell. Figure created with Biorender.

Incidental Terminal Phenotypes

Decision to euthanize animals was made only following veterinary advice. Immediately prior to euthanasia, blood samples were obtained to harvest serum and plasma. Animals were euthanized by barbiturate overdose. Immediately following euthanasia, cerebrospinal fluid was obtained from the atlanto-occipital space and

urine collected by sterile catheterization. All animals were subject to complete gross and histologic necropsy. Extensive tissue collection was carried out with flash frozen and formalin-fixed paraffin embedded tissue retained from all animals.

Data Analysis

Descriptive statistics were generated for breed, sex, and age distributions. Z-scores were calculated using the proportion of represented breeds in the P100HHP cohort compared to the most recent horse population survey generated by the United States Department of Agriculture.[25] Age and time in residence correlation were analyzed by Spearman's rank correlation test. Data were analyzed using Prism V 9.2.0 (GraphPad Software LLC, San Diego).

Data Availability

Whole genome data can be accessed using the National Center for Biotechnology Information Sequence Read Archive (NCBO SRA) with the accessions PRJNA841639, PRJNA553581 and PRJNA601992. All remaining data are within the manuscript and its Supplementary Data files.

Funding

This study was supported by the University of California, Davis Center for Equine Health with funds provided by the State of California parimutuel fund, Platinum Performance Inc. and private donors. Graduate support was provided by private donation. Support for CJF was provided by the National Institutes of Health (NIH) (L40 TR001136).

References

1. Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *New England journal of medicine*, *372*(9), 793-795.
2. Yurkovich, J. T., Tian, Q., Price, N. D., & Hood, L. (2020). A systems approach to clinical oncology uses deep phenotyping to deliver personalized care. *Nature reviews. Clinical oncology*, *17*(3), 183–194. https://doi.org/10.1038/s41571-019-0273-6
3. Chen, R., Mias, G. I., Li-Pook-Than, J., Jiang, L., Lam, H. Y., Chen, R., Miriami, E., Karczewski, K. J., Hariharan, M., Dewey, F. E., Cheng, Y., Clark, M. J., Im, H., Habegger, L., Balasubramanian, S., O'Huallachain, M., Dudley, J. T., Hillenmeyer, S., Haraksingh, R., Sharon, D., … Snyder, M. (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, *148*(6), 1293–1307. https://doi.org/10.1016/j.cell.2012.02.009
4. Smarr L. (2012). Quantifying your body: a how-to guide from a systems biology perspective. *Biotechnology journal*, *7*(8), 980–991. https://doi.org/10.1002/biot.201100495
5. Price ND, Magis AT, Earls JC, et al. A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nat Biotechnol*. 2017;35(8):747-756. doi:10.1038/nbt.3870
6. Wilmanski, T., Diener, C., Rappaport, N., Patwardhan, S., Wiedrick, J., Lapidus, J., Earls, J. C., Zimmer, A., Glusman, G., Robinson, M., Yurkovich, J. T., Kado, D. M., Cauley, J. A., Zmuda, J., Lane, N. E., Magis, A. T., Lovejoy, J. C., Hood, L., Gibbons, S. M., Orwoll, E. S., … Price, N. D. (2021). Gut microbiome pattern reflects healthy ageing and predicts survival in

humans. *Nature metabolism*, *3*(2), 274–286. https://doi.org/10.1038/s42255-021-00348-0

7. Wilmanski, T., Rappaport, N., Earls, J. C., Magis, A. T., Manor, O., Lovejoy, J., Omenn, G. S., Hood, L., Gibbons, S. M., & Price, N. D. (2019). Blood metabolome predicts gut microbiome α-diversity in humans. *Nature biotechnology*, *37*(10), 1217–1228. https://doi.org/10.1038/s41587-019-0233-9
8. Manor, O., Dai, C. L., Kornilov, S. A., Smith, B., Price, N. D., Lovejoy, J. C., Gibbons, S. M., & Magis, A. T. (2020). Health and disease markers correlate with gut microbiome composition across thousands of people. *Nature communications*, *11*(1), 5206. https://doi.org/10.1038/s41467-020-18871-1
9. Levy, R., Magis, A. T., Earls, J. C., Manor, O., Wilmanski, T., Lovejoy, J., Gibbons, S. M., Omenn, G. S., Hood, L., & Price, N. D. (2020). Longitudinal analysis reveals transition barriers between dominant ecological states in the gut microbiome. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(24), 13839–13845. https://doi.org/10.1073/pnas.1922498117
10. Zubair, N., Conomos, M. P., Hood, L., Omenn, G. S., Price, N. D., Spring, B. J., Magis, A. T., & Lovejoy, J. C. (2019). Genetic Predisposition Impacts Clinical Changes in a Lifestyle Coaching Program. *Scientific reports*, *9*(1), 6805. https://doi.org/10.1038/s41598-019-43058-0
11. Diener, C., Qin, S., Zhou, Y., Patwardhan, S., Tang, L., Lovejoy, J. C., Magis, A. T., Price, N. D., Hood, L., & Gibbons, S. M. (2021). Baseline Gut Metagenomic Functional Gene Signature Associated with Variable Weight Loss Responses following a Healthy Lifestyle Intervention in

Humans. *mSystems*, *6*(5), e0096421. https://doi.org/10.1128/mSystems.00964-21

12. Earls, J. C., Rappaport, N., Heath, L., Wilmanski, T., Magis, A. T., Schork, N. J., Omenn, G. S., Lovejoy, J., Hood, L., & Price, N. D. (2019). Multi-Omic Biological Age Estimation and Its Correlation With Wellness and Disease Phenotypes: A Longitudinal Study of 3,558 Individuals. *The journals of gerontology. Series A, Biological sciences and medical sciences*, *74*(Suppl_1), S52–S60. https://doi.org/10.1093/gerona/glz220
13. Wainberg, M., Magis, A. T., Earls, J. C., Lovejoy, J. C., Sinnott-Armstrong, N., Omenn, G. S., Hood, L., & Price, N. D. (2020). Multiomic blood correlates of genetic risk identify presymptomatic disease alterations. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(35), 21813–21820. https://doi.org/10.1073/pnas.2001429117
14. Magis, A. T., Rappaport, N., Conomos, M. P., Omenn, G. S., Lovejoy, J. C., Hood, L., & Price, N. D. (2020). Untargeted longitudinal analysis of a wellness cohort identifies markers of metastatic cancer years prior to diagnosis. *Scientific reports*, *10*(1), 16275. https://doi.org/10.1038/s41598-020-73451-z
15. Heath, L., Earls, J. C., Magis, A. T., Kornilov, S. A., Lovejoy, J. C., Funk, C. C., Rappaport, N., Logsdon, B. A., Mangravite, L. M., Kunkle, B. W., Martin, E. R., Naj, A. C., Ertekin-Taner, N., Golde, T. E., Hood, L., Price, N. D., & Alzheimer's Disease Genetics Consortium (2022). Manifestations of Alzheimer's disease genetic risk in the blood are evident in a multiomic analysis in healthy adults aged 18 to 90. *Scientific reports*, *12*(1), 6117. https://doi.org/10.1038/s41598-022-09825-2

16. Su, Y., Yuan, D., Chen, D. G., Ng, R. H., Wang, K., Choi, J., Li, S., Hong, S., Zhang, R., Xie, J., Kornilov, S. A., Scherler, K., Pavlovitch-Bedzyk, A. J., Dong, S., Lausted, C., Lee, I., Fallen, S., Dai, C. L., Baloni, P., Smith, B., … Heath, J. R. (2022). Multiple early factors anticipate post-acute COVID-19 sequelae. *Cell*, *185*(5), 881–895.e20. https://doi.org/10.1016/j.cell.2022.01.014
17. Guy, M. K., Page, R. L., Jensen, W. A., Olson, P. N., Haworth, J. D., Searfoss, E. E., & Brown, D. E. (2015). The Golden Retriever Lifetime Study: establishing an observational cohort study with translational relevance for human health. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1673), 20140230. https://doi.org/10.1098/rstb.2014.0230
18. Buckley, R. M., & Lyons, L. A. (2020). Precision/Genomic Medicine for Domestic Cats. *The Veterinary clinics of North America. Small animal practice*, *50*(5), 983–990. https://doi.org/10.1016/j.cvsm.2020.05.005
19. Creevy, K.E., Akey, J.M., Kaeberlein, M. *et al.* An open science study of ageing in companion dogs. *Nature* **602,** 51–57 (2022). https://doi.org/10.1038/s41586-021-04282-9
20. Li, R. H., Ontiveros, E., Nguyen, N., Stern, J. A., Lee, E., Hardy, B. T., & 99 Lives Cat Genome Consortium. (2020). Precision medicine identifies a pathogenic variant of the ITGA2B gene responsible for Glanzmann's thrombasthenia in a cat. *Journal of veterinary internal medicine*, *34*(6), 2438-2446. https://doi.org/10.1111/jvim.15886
21. Mauler, D. A., Gandolfi, B., Reinero, C. R., O'Brien, D. P., Spooner, J. L., Lyons, L. A., ... & Wildschutte, J. H. (2017). Precision medicine in cats: Novel Niemann-Pick type C1 diagnosed by whole-genome sequencing. *Journal of*

veterinary internal medicine, *31*(2), 539-544. https://doi.org/10.1111/jvim.14599

22. Pang, L. Y., & Argyle, D. J. (2016). Veterinary oncology: biology, big data and precision medicine. *The Veterinary Journal*, *213*, 38-45. https://doi.org/10.1016/j.tvjl.2016.03.009
23. Wang, G., Wu, M., Durham, A. C., Mason, N. J., & Roth, D. B. (2021). Canine Oncopanel: A capture-based, NGS platform for evaluating the mutational landscape and detecting putative driver mutations in canine cancers. *Veterinary and Comparative Oncology*. https://doi.org/10.1111/vco.12746
24. James A. Roth, Christopher K. Tuggle, Livestock Models in Translational Medicine, *ILAR Journal*, Volume 56, Issue 1, 2015, Pages 1–6, https://doi.org/10.1093/ilar/ilv011
25. United States Department of Agriculture. Demographics of the U.S. Equine Population, 2015. Animal and Plant Health Inspection Service; 2017 Feb
26. Petersen, J. L., Mickelson, J. R., Cothran, E. G., Andersson, L. S., Axelsson, J., Bailey, E., Bannasch, D., Binns, M. M., Borges, A. S., Brama, P., da Câmara Machado, A., Distl, O., Felicetti, M., Fox-Clipsham, L., Graves, K. T., Guérin, G., Haase, B., Hasegawa, T., Hemmann, K., Hill, E. W., … McCue, M. E. (2013). Genetic diversity in the modern horse illustrated from genome-wide SNP data. *PloS one*, *8*(1), e54997. https://doi.org/10.1371/journal.pone.0054997
27. Online Mendelian Inheritance in Animals, OMIA. Sydney School of Veterinary Science, {date of download}. World Wide Web URL: https://omia.org/

28. Tryon, R. C., Penedo, M. C., McCue, M. E., Valberg, S. J., Mickelson, J. R., Famula, T. R., Wagner, M. L., Jackson, M., Hamilton, M. J., Nooteboom, S., & Bannasch, D. L. (2009). Evaluation of allele frequencies of inherited disease genes in subgroups of American Quarter Horses. *Journal of the American Veterinary Medical Association*, *234*(1), 120–125. https://doi.org/10.2460/javma.234.1.120

29. Promerová, M., Andersson, L. S., Juras, R., Penedo, M. C., Reissmann, M., Tozaki, T., Bellone, R., Dunner, S., Hořín, P., Imsland, F., Imsland, P., Mikko, S., Modrý, D., Roed, K. H., Schwochow, D., Vega-Pla, J. L., Mehrabani-Yeganeh, H., Yousefi-Mashouf, N., G Cothran, E., Lindgren, G., … Andersson, L. (2014). Worldwide frequency distribution of the 'Gait keeper' mutation in the DMRT3 gene. *Animal genetics*, *45*(2), 274–282. https://doi.org/10.1111/age.12120

30. Ward, T. L., Valberg, S. J., Adelson, D. L., Abbey, C. A., Binns, M. M., & Mickelson, J. R. (2004). Glycogen branching enzyme (GBE1) mutation causing equine glycogen storage disease IV. *Mammalian genome : official journal of the International Mammalian Genome Society*, *15*(7), 570–577. https://doi.org/10.1007/s00335-004-2369-1

31. Marshall, D. A., MacDonald, K. V., Robinson, J. O., Barcellos, L. F., Gianfrancesco, M., Helm, M., ... & Phillips, K. A. (2017). The price of whole-genome sequencing may be decreasing, but who will be sequenced?. *Personalized medicine*, *14*(3), 203-211. https://doi.org/10.2217/pme-2016-0075

32. Durward-Akhurst, S. A., Schaefer, R. J., Grantham, B., Carey, W. K., Mickelson, J. R., & McCue, M. E. (2021). Genetic Variation and the

Distribution of Variant Types in the Horse. *Frontiers in genetics*, *12*, 758366. https://doi.org/10.3389/fgene.2021.758366

33. Jagannathan, V., Gerber, V., Rieder, S., Tetens, J., Thaller, G., Drögemüller, C., & Leeb, T. (2019). Comprehensive characterization of horse genome variation by whole-genome sequencing of 88 horses. *Animal genetics*, *50*(1), 74-77. https://doi.org/10.1111/age.12753
34. Siva, N. (2008). 1000 Genomes project. *Nature biotechnology*, *26*(3), 256-257. https://doi.org/10.1038/nbt0308-256b
35. McCue, M. E., & McCoy, A. M. (2019). Harnessing big data for equine health. *Equine veterinary journal*, *51*(4), 429–432. https://doi.org/10.1111/evj.13080
36. Abul-Husn, N. S., & Kenny, E. E. (2019). Personalized Medicine and the Power of Electronic Health Records. *Cell*, *177*(1), 58–69. https://doi.org/10.1016/j.cell.2019.02.039
37. Gold, S., Regan, C. E., McLoughlin, P. D., Gilleard, J. S., Wilson, A. J., & Poissant, J. (2019). Quantitative genetics of gastrointestinal strongyle burden and associated body condition in feral horses. *International journal for parasitology. Parasites and wildlife*, *9*, 104–111. https://doi.org/10.1016/j.ijppaw.2019.03.010
38. Cox, R. M., McGlothlin, J. W., & Bonier, F. (2016). Hormones as Mediators of Phenotypic and Genetic Integration: an Evolutionary Genetics Approach. *Integrative and comparative biology*, *56*(2), 126–137. https://doi.org/10.1093/icb/icw033

39. Tian, C., Gregersen, P. K., & Seldin, M. F. (2008). Accounting for ancestry: population substructure and genome-wide association studies. *Human molecular genetics*, *17*(R2), R143–R150. https://doi.org/10.1093/hmg/ddn268
40. Sesia, M., Bates, S., Candès, E., Marchini, J., & Sabatti, C. (2021). False discovery rate control in genome-wide association studies with population structure. *Proceedings of the National Academy of Sciences*, *118*(40).
41. Buniello, A., MacArthur, J., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., Flicek, P., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, *47*(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120
42. Muyas, F., Bosio, M., Puig, A., Susak, H., Domènech, L., Escaramis, G., Zapata, L., Demidov, G., Estivill, X., Rabionet, R., & Ossowski, S. (2019). Allele balance bias identifies systematic genotyping errors and false disease associations. *Human mutation*, *40*(1), 115–126. https://doi.org/10.1002/humu.23674
43. Krueger, F. (2015). Trim galore. *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files*, *516*(517).
44. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, *17*(1), 10-12.
45. Kalbfleisch, T. S., Rice, E. S., DePriest, M. S., Jr, Walenz, B. P., Hestand, M. S., Vermeesch, J. R., O Connell, B. L., Fiddes, I. T., Vershinina, A. O., Saremi, N. F., Petersen, J. L., Finno, C. J., Bellone, R. R., McCue, M. E.,

Brooks, S. A., Bailey, E., Orlando, L., Green, R. E., Miller, D. C., Antczak, D. F., … MacLeod, J. N. (2018). Improved reference genome for the domestic horse increases assembly contiguity and composition. *Communications biology*, *1*, 197. https://doi.org/10.1038/s42003-018-0199-z

46. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*, *25*(14), 1754-1760.
47. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079
48. Van der Auwera GA & O'Connor BD. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition).* O'Reilly Media
49. Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin, 1000 Genomes Project Analysis Group, The variant call format and VCFtools, *Bioinformatics*, Volume 27, Issue 15, 1 August 2011, Pages 2156–2158, https://doi.org/10.1093/bioinformatics/btr330
50. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81. https://doi.org/10.1086/519795
51. Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, and Durbin R. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* (2016) 32(11) 1749-51

52. J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.
53. Waskom, M. L., (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021, https://doi.org/10.21105/joss.03021
54. Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754
55. Brown, S. D., & Moore, M. W. (2012). The International Mouse Phenotyping Consortium: past and future perspectives on mouse phenotyping. *Mammalian genome : official journal of the International Mammalian Genome Society*, *23*(9-10), 632–640. https://doi.org/10.1007/s00335-012-9427-x

Chapter 2: Integrated Metabolome-Microbiota Analysis Identifies Molecular Drivers of Equine Metabolic Syndrome

Authors: Callum G. Donnelly[1], Erin Oberhaus[2] Michele Coleman[3], Jane M. Manfredi[4], Stephanie Valberg[5], Carrie J. Finno[1]

[1] Department of Population Health and Reproduction, School of Veterinary Medicine, University of California, Davis

[2]School of Animal Science, Louisiana State University

[3] Department of Large Animal Medicine, College of Veterinary Medicine, University of Georgia

[4]Department of Pathobiology and Diagnostic Investigation, College of Veterinary Medicine, University of Michigan

[5] Department of Large Animal Clinical Sciences, College of Veterinary Medicine, University of Michigan

Abstract

Regulation of metabolic function relies not only host mechanisms, but also the complex interaction with the resident microbiota. The metabolome lies at the crossroads of this interaction, bidirectionally signaling systemic, nutritional, and environmental inputs to maintain homeostasis. Metabolic syndrome typifies dysregulation of this communication, resulting in obesity, insulin resistance, dyslipidemia and hypertension. Using a systems biology approach, we established the relationship between host-metabolome-microbiota for metabolic syndrome in a population of deeply phenotyped horses. Our analyses indicate that a permissive relationship between the imbalance of primary and secondary bile acids drives

susceptibility to insulin dysregulation, independently of obesity and dyslipidemia. Individuals with higher circulating levels of primary (i.e. host-derived) bile acids segregate strongly with metabolic syndrome phenotypes, including insulin dysregulation.

Introduction

The collection of phenotypes defining metabolic syndrome (MetS) are common to both humans and horses; obesity, insulin resistance, dyslipidemia and hypertension.[1,2] Despite differences in diet types, horses encounter MetS at a similar rate to that of the human population, with up to 40% of horses considered obese or overweight .[3,4] This shared health crisis makes horses a powerful naturally-occurring disease analogue to determine intrinsically shared mechanisms driving MetS.

The gastrointestinal (GI) microbiome is increasingly recognized as an active participant in the development of obesity and MetS.[5] Much of this effect in humans has been attributed to the modernization of food production systems in the post-industrialized world. Principally, people of western societies have a distinct, yet less complex, GI microbial population regarded as the 'industrial' microbiome. This stands in contrast to non-western and indigenous GI microbiomes, where richness and diversity are maintained, and the risk of metabolic related diseases is reduced.[7,8] Rodents have predominated in experimental studies of the 'industrial' microbiome. The premise being that, with successively more restrictive microbial colonization of rodents housed in modern vivaria, this situation approximates the same shift seen with industrialization.[9,10] While valuable, rodent studies cannot fully recapitulate the occurrence of both the industrial/domestic GI microbiome and spontaneous MetS.

Domestication of horses delineates one of the initial steps towards an industrialized society.[11] Similar to industrialization of human society, domestication of horses had a comparable effect on the microbiota. Non-domesticated horses maintain a higher microbial diversity than their domestic counterparts.[12] In addition to differing dietary habits of domestic horses, they now also share modifying exposures similar to those of humans, namely the use of antimicrobial medications.[12] To this end, we have developed a cohort of horses that have the necessary density of phenotype, omic and environmental data to evaluate the complex relationships that lead to MetS. This population of horses, known as the Pioneer 100 Horse Health Project (P100HHP) is the only dedicated group of research animals maintained for systems biology-based investigation of health and disease traits.

Defining the microbial contribution to homeostasis requires a medium for host-microbiome communication. The most readily studied of these is the metabolome.[13] The bidirectional synthesis and metabolism of small molecules provides direct and indirect signals to the host to maintain a number of metabolic functions.[14] Numerous studies have indicated that integration of metabolome and microbiome data enables a more complete picture for drivers of health and disease, including MetS.[15-17] Indeed, the integration of multiple high dimensional data sets is a necessary feature of systems-based approaches to disease investigation.[18] In this study, we address the limitations of smaller sample size with high density multidimensional omic and phenotype data. The effectiveness of this approach in resolving microbiome-metabolome signatures in health and disease has been well validated with the human Arivale cohort.[18-20] In the original study, 108 individuals were sufficient to

define important correlational families in cross-omic and phenotypic datasets. In a similar manner, our study has generated high-density datasets for over 100 individual horses, with the intent of defining microbiome-metabolome signatures for MetS, in addition to improving the molecular classification of MetS sub/endotypes.

Using data generated by the P100HHP, we aim to explore host and microbiota directed dysregulation of metabolic homeostasis to improve clinical designation and management of this syndrome. In this study, we hypothesized that MetS will segregate into sub/endotypes, termed metabotypes, that are defined by specific metabolomic signatures. We further hypothesize that metabolome-microbiome integration will allow for the detection of host, microbiome, and environmental directionality underlying MetS in horses.

Results

Study Design and Cohort Description

This study used data generated from the P100HHP cohort; a longitudinal and deeply phenotyped population of horses (n=108) living at a single research facility.[21] Population characteristics of the horses have previously been reported.[21] Individuals in this arm of the study were repeatedly evaluated for the cluster of phenotypic traits associated with MetS; insulin dysregulation, pars pituitary intermedia dysfunction and obesity. Briefly, individuals were repeatedly evaluated for insulin regulation by oral sugar test (OST), baseline adrenocorticotropin hormone (ACTH) and thyrotropin-releasing hormone stimulation test (TRH-stim), body condition scoring and body weight measurement (see *Methods*). Individuals were considered MetS if they were insulin dysregulated (ID) in ≥ 2/3 OST matched metabolomic datasets (n=8). If an

individual was ID periodically (Spring or Fall) in <2 OST they were designated as seasonally MetS (SMetS; n=29). Individuals that remained insulin regulated at all time points were considered metabolically healthy (REG; n=35). Individuals were considered affected by equine pars pituitary intermedia deficiency (PPID) based on results of TRH-stimulation testing and suggestive physical phenotype (hypertrichosis, retrobulbar fat deposition, muscle atrophy), in accordance with current diagnostic guidelines.[22] Only two individuals (604 and 858) met diagnostic criterion for PPID and so this phenotype was not considered further in the analysis. An individual was regarded as obese if the average body condition score (BCS) was >7/9 (n=10) and overweight if >6/9 (n=23) over the entire observation period. .[23]

Individuals were also phenotyped for other, non-MetS health-related phenotypes as described elsewhere.[21] In concert with phenotype assessment, contemporaneous samples were collected for evaluation of plasma metabolome and fecal microbiota (see *Methods*). As part of the entry requirement for individuals into the study, whole genome sequence data and electronic medical record data had already been collected.[21]

Outlier Metabolites Segregate with MetS

To characterize the metabolite communities of individuals, the mean metabolite concentrations across all time points were evaluated by outlier analysis (see *Methods*).[24] Metabolites with ≥ 5 individuals/metabolite outlier were selected for further analysis. The mean number of outliers per individual was 22 ± 47. One individual had considerably more metabolite outliers (354 metabolites) as compared to the rest of the cohort. After initially segregating alone, this individual was not

included in further analysis. Unsupervised hierarchical clustering of outliers by Euclidean distance identified two major clades in the data (**Fig. 2.1**). These major clades matched insulin regulation states, with a more insulin regulated clade (Clade 1) and a more insulin dysregulated clade (Clade 2). In the MetS group, 85% of individuals segregated to Clade 1 and 82% of regulated individuals clustered in Clade 2. SMetS individuals were intermediary between the extreme ends, with half occurring in each of the major clades. Segregating major clades were not driven by breed (*P*=0.3*)*, age (*P*=0.2*)* or sex (*P*=0.4*)*. Mean body condition scores (relative adiposity) were significantly higher for the more insulin dysregulated clade (mean ± S.D.; 6.7 ± 0.9) as compared to the more insulin regulated clade (5.9 ± 0.8; *P*=0.0003). Similarly, MetS individuals had higher mean BCS (7.1 ± 0.9) as compared to REG individuals (5.8 ± 0.7; *P*=0.0007), regardless of segregating clade. SMetS individuals did not separate from MetS (*P=0.4)* nor from REG (*P*=0.05) individuals based on BCS. Metabolite outliers were closely associated with the overlapping phenotypes of obesity and/or insulin dysregulation for MetS and REG individuals.

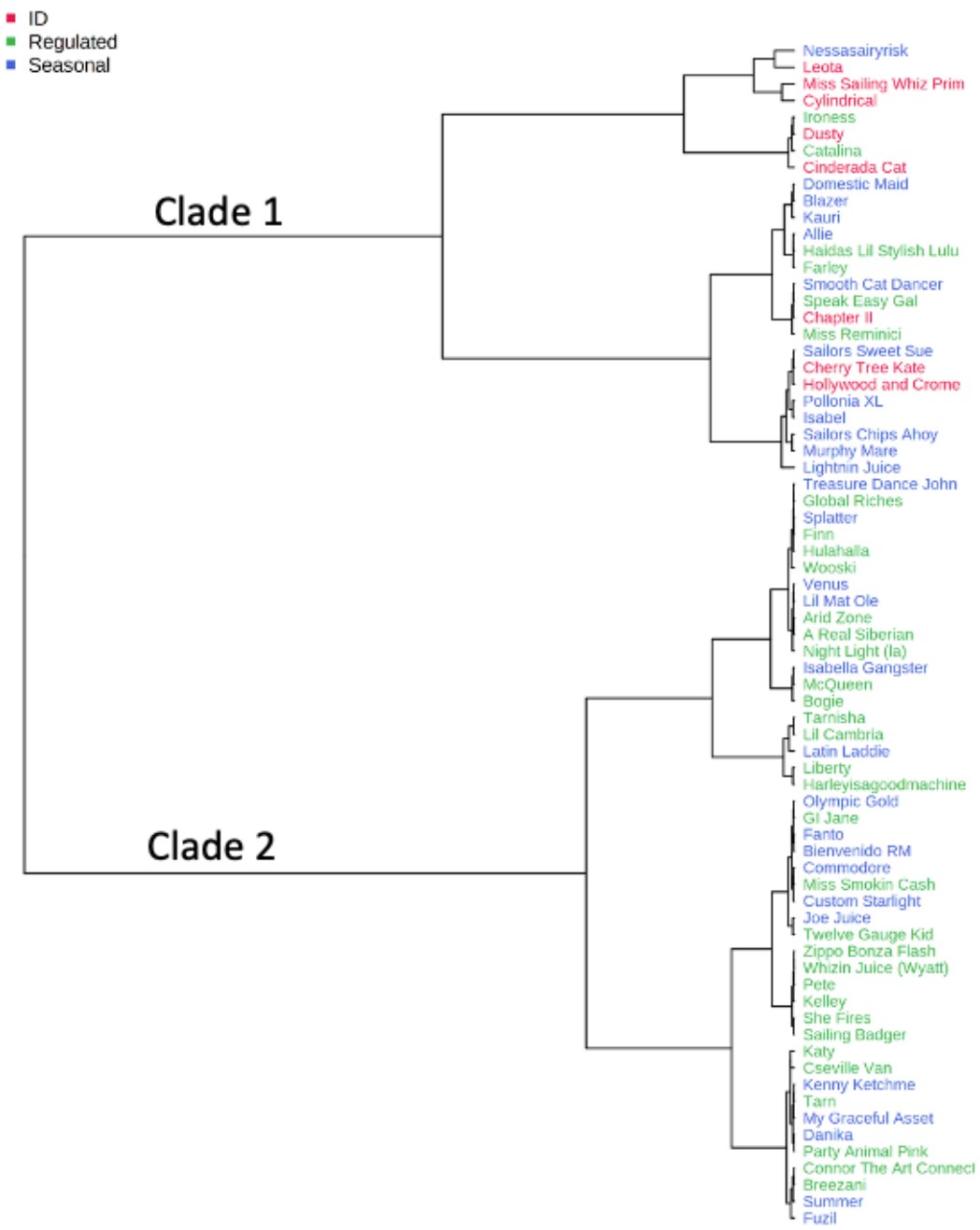
ID
Regulated
Seasonal
Clade 1
Clade 2
Nessasairyrisk
Leota
Miss Sailing Whiz Prim
Cylindrical
Ironess
Dusty
Catalina
Cinderada Cat
Domestic Maid
Blazer
Kauri
Allie
Haidas Lil Stylish Lulu
Farley
Smooth Cat Dancer
Speak Easy Gal
Chapter II
Miss Reminici
Sailors Sweet Sue
Cherry Tree Kate
Hollywood and Crome
Pollonia XL
Isabel
Sailors Chips Ahoy
Murphy Mare
Lightnin Juice
Treasure Dance John
Global Riches
Splatter
Finn
Hulahalla
Wooski
Venus
Lil Mat Ole
Arid Zone
A Real Siberian
Night Light (Ia)
Isabella Gangster
McQueen
Bogie
Tarnisha
Lil Cambria
Latin Laddie
Liberty
Harleyisagoodmachine
Olympic Gold
GI Jane
Fanto
Bienvenido RM
Commodore
Miss Smokin Cash
Custom Starlight
Joe Juice
Twelve Gauge Kid
Zippo Bonza Flash
Whizin Juice (Wyatt)
Pete
Kelley
She Fires
Sailing Badger
Katy
Cseville Van
Kenny Ketchme
Tarn
My Graceful Asset
Danika
Party Animal Pink
Connor The Art Connect
Breezani
Summer
Fuzil

Figure 2.1 Dendrogram of hierarchical clustering based on outlier analysis of metabolites. More insulin dysregulated individuals segregated in clade 1 and more insulin regulated individuals segregated in clade 2. Animals were classed as MetS (ID; red), SMetS (Seasonal; blue) or Regulated (green).

MetS subtypes are defined by metabolite classes

We next evaluated the data for metabolite classes that were driving subclusters in Clade 1 to define metabotypes. Two major subclusters emerged, one driven primarily by MetS individuals and the other by SMetS individuals. The MetS predominant subcluster was dominated by monoarachidonic triglyceride (TG) species, ranging in length from 14 to 18 carbons, and could be summarized as an overabundance of unsaturated TGs. The second major sub-cluster was defined by the increased abundance of primary bile acid and bile acid conjugates (**Fig. 2.2**).

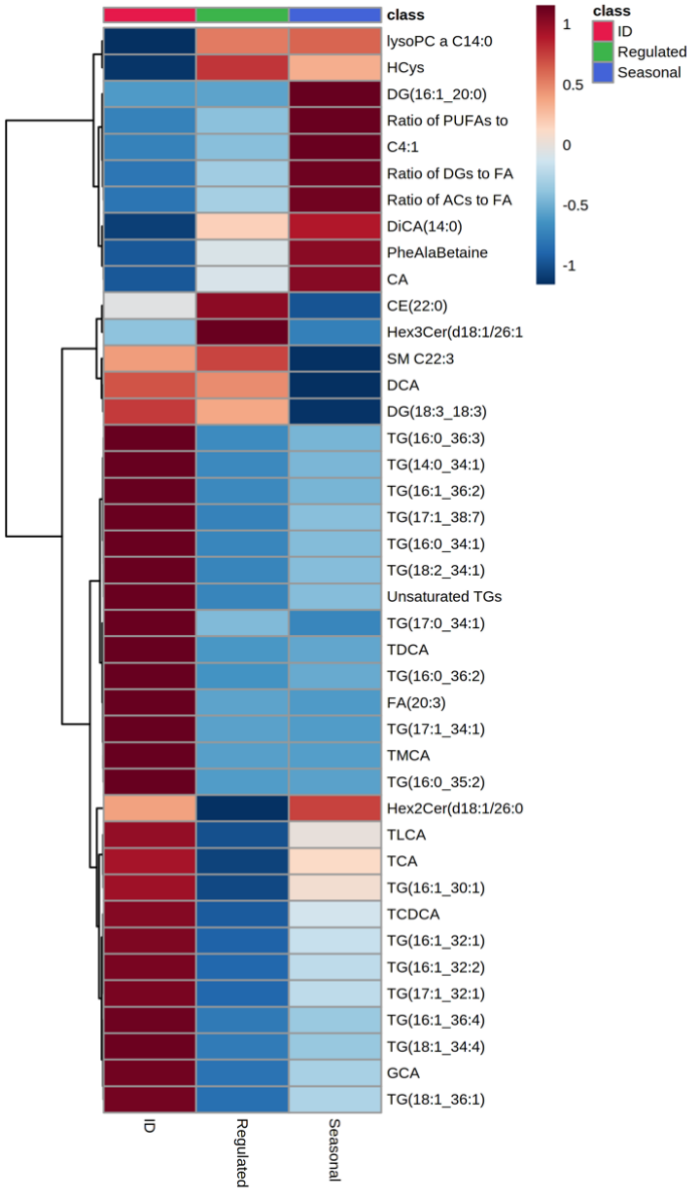

Figure 2.2 Heatmap of top 25 metabolites or calculated indices used in hierarchical analysis of outlier data. Abbreviations: lysoPC- lysophosphotidylcholine; HCys- homocysteine; DG-diglyceride; PUFA- polyunsaturated fatty acid; C4:1- butenylcarnitine; FA- fatty acid; AC- acetylcarnitines; PheAlaBetaine- phenylalanine betaine; CA- cholic acid; CE- ceramide; Hex3Cer- trihexosylceramide; SM- sphingomyelin; DCA- deoxycholic acid; TG- triglyceride; TDCA- taurodeoxycholic acid; TMCA- tauro-b-muricholic acid; Hex2Cer- dihexosyloceramide; TLCA-

taurolithocholic acid; TCA- taurocholic acid; TCDCA- taurochenodeoxycholic acid; GCA- glycocholic acid

Sparse partial least squares discriminant analysis (sPLSDA) of individuals by insulin group recapitulated these results, with the first component defined by TGs in MetS individuals (**Fig. 2.3A).** The subsequent two components were defined by both primary and secondary bile acids that provided distinction between SMetS and REG individuals (**Fig. 2.3B**). In addition, sphingomyelins, ceramides, budenylcarnitine and calculated indices of acetylcarnitines and diglycerides provide further nuance between SMetS and REG individuals.

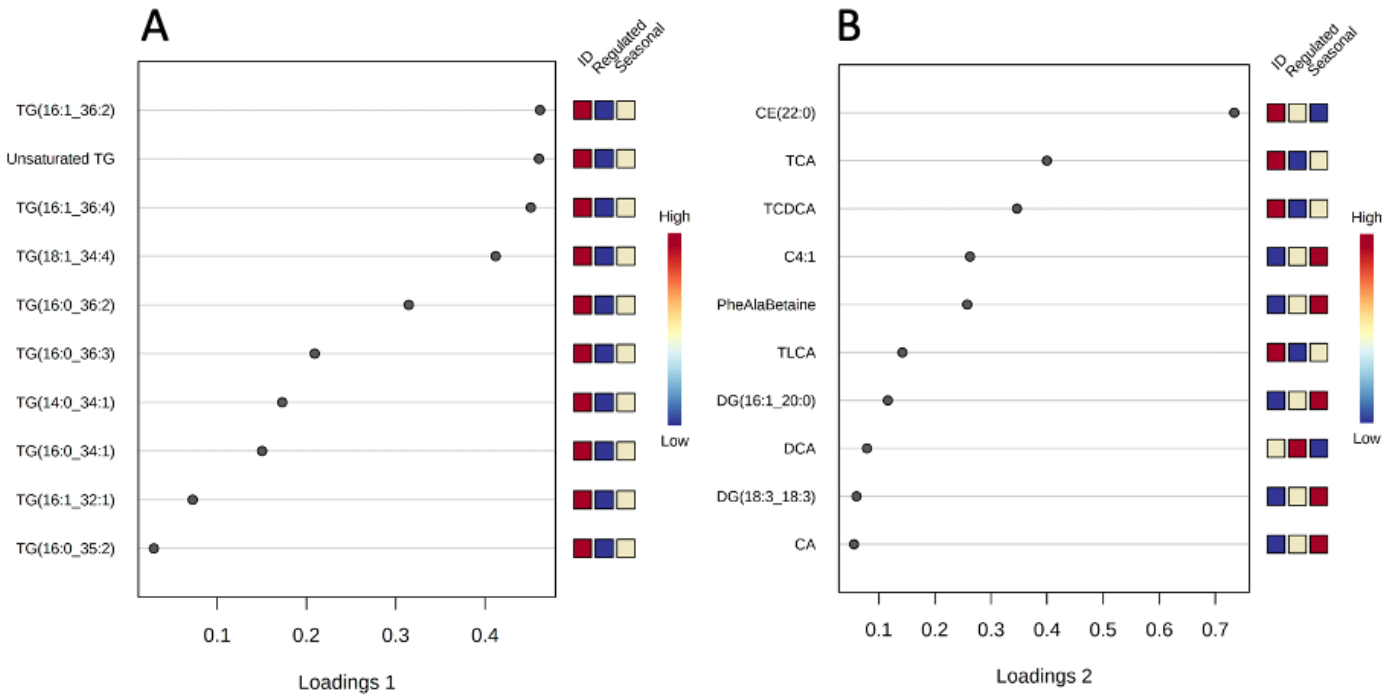

Figure 2.3 Loading plots 1 (**A**) and 2 (**B**) for sPLSDA analysis defining sub-clades in Clade 1. Abreviations: TG- triglyceride; CE- ceramide; TCA- taurocholic acid; TCDCA- taurochenodeoxycholic acid; C4:1- butenylcarnitine; PheAlaBetaine-

phenylalanine betaine; TLCA- taurolithocholic acid; DG-diglyceride; DCA-deoxycholic acid; CA- cholic acid

Bile acid enrichment indicates microbiota-metabolome interaction in MetS

The primary bile acid (BA) conjugate taurocehnodeoxycholic acid (TCDCA) was elevated in individuals with MetS compared to REG individuals (median (95%CI); 2.19 (1.55-4.21) μmol/L vs 1.41 (1.28-1.86) μmol/L, *P*=0.03). Taurocholic acid (TCA) was also elevated in MetS and SMetS individuals (2.19 (1.34 (1.02-2.12) μmol/L vs 0.95 (0.9-1.27) μmol/L, *P*=0.05). Although it did not reach FDR-corrected pairwise significance, the secondary BA deoxycholic acid (DCA) was a discriminant feature in sPLSDA analysis, with higher enrichment in REG individuals (**Fig. 2.3**). The balance of primary and secondary BAs is reflective of the microbial environment, with secondary BAs produced by the gastrointestinal microbiota.[25] Given this relationship, the contribution of the microbiota to the prevailing metabolic phenotype was evaluated. To date, only a select group of species of gram-positive Firmicutes, mostly in the genus *Clostridium,* have shown the ability to metabolize primary BAs.[26] In this population of individuals, post-prandial insulin concentrations were inversely correlated with the proportion of bacterium in the Clostridiales family (**Fig. 2.4A**). Similarly, TCDCA also had an inverse association with Clostridiales, although this failed to reach corrected significance (*P*=0.08). However, TDCA did have a significant positive relationship with the phyla Proteobacter (*P*=0.03, **Fig. 2.4B**). Proteobacter species are not known to metabolize BA and are associated with dysbiosis.[27] We conclude that the prevailing metabolic environment reflects microbial population shifts that may promote a MetS phenotype mediated by BA dysregulation.

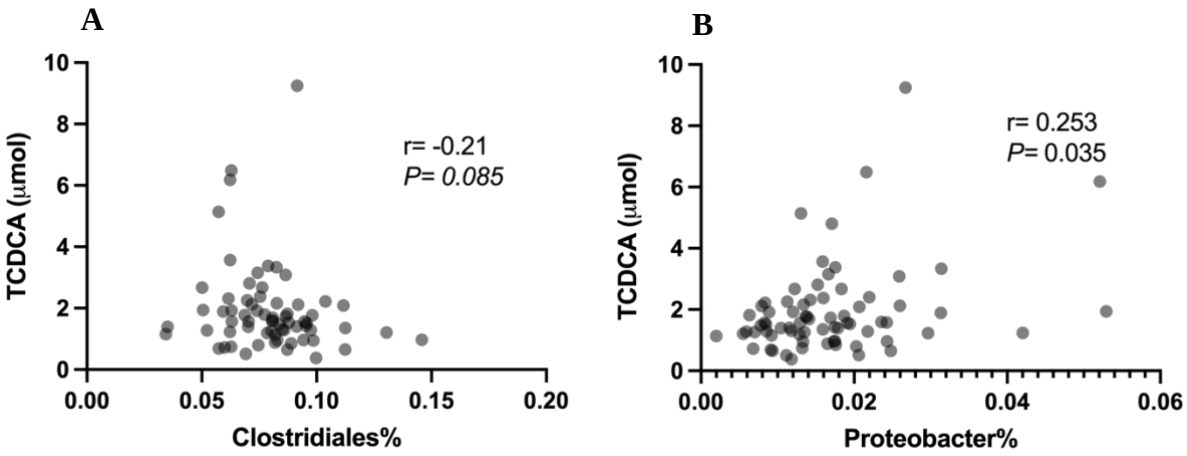

Figure 2.4 Correlation plots for post prandial insulin and Clostridales (**A**) and the bile acid taurochenodeoxycholic acid (TCDCA) and Proteobacter (**B**).

Discussion

For the first time, we have demonstrated the use of a systems approach to defining metabotypes for MetS in horses. Particularly, we have demonstrated that MetS and SMetS horses segregate into two predominant groups; a TG dominant and a BA dominant metabotype. Further, we have detected a relationship between the metabolome and microbiota that associates with phenotypes of MetS and may be mediated by BA dysregulation. Collectively, we demonstrate the modifying effect that the prevailing microbial-metabolome interaction has on the development of MetS in horses. This cements the contention that horses are highly suitable model of spontaneous MetS.

The association of metabolomic signatures with MetS in horses that allows for distinct subtypes to be defined is an important step toward improving the diagnosis and treatment of the syndrome. This has largely not been a focus of previous

metabolomic studies, where the predominant aim has been metabolomic biomarker or metabogenomic marker discovery.[28-31] Additionally, previous studies have been hampered by one or more limitations including small sample size, single breed investigation and single time point sampling.[28-31] The current study was specifically designed to address these limitations and to detect metabolomic signatures that are biologically reflective of MetS and MetS subtypes.

Dyslipidemia, particularly hypertriglyceridemia, is strongly associated with insulin resistance and type 2 diabetes in humans.[32] We have demonstrated that enrichment for unsaturated TGs defines a sub-cluster that is dominated by MetS individuals and, to a lesser degree, SMetS or REG individuals. Various unsaturated TGs evaluated in this study distinguished MetS from SMetS and REG horses. Outlier analysis indicated that specific dysregulated TGs represents MetS better in some individuals than total TGs. Additionally, this relationship is not linear, with the severity of hypertriglyceridemia not predictive of severity of ID, merely that the two are likely to exist together. In support of this, the individual removed from the group analysis (#1182) had a total unsaturated TGs 20 times higher than that of the next highest horse, yet had a normal fasted insulin and mildly inappropriate post-prandial response. In obese horses not phenotyped for ID, untargeted lipidomic profiles identified highly complex obesity associated signatures.[29] However, due to individuals coming from multiple farms (environments), clustering in this study was highly biased. In individuals phenotyped for MetS from a single environment, untargeted lipidomics identified many dysregulated lipid species, with a predominance of TGs.[33] Collectively, the current study confirms that TGs are associated non-linearly with the MetS phenotype. We add to the existing known

metabolites and lipids associated with MetS and suggest that TG identifies a metabotype within the MetS spectrum. Further measurement of total TGs may not be sufficient to capture the nature of dysregulation in MetS individuals. Therefore, development of assays for specific TGs may improve diagnosis and targeted therapy for equine MetS.

Bile acid (BA) signaling has become a fundamental pathway for the development of MetS in humans.[32] To date, the effect of BAs, as they relate to equine MetS, has not been demonstrated. Through their interaction with the farsenoid X receptor (FXR) and Takeda G protein receptor 5 (TGR5), BAs have direct roles on incretin secretion, gluconeogenesis, glycogen synthesis, energetic regulation and inflammation in multiple organs.[32] Although total bile acids are routinely measured in horses, compositional analysis has rarely been reported. From a limited sample number, it was determined that TCDCA and TCA are the major components (>80%) of bile acid in horses.[34] In that study, TLCA and DCA were not identified. The discrepancy likely lies in the analytic sensitivity to detect the lower abundance BAs as we have been able to in the current study. Importantly, the relative and absolute abundance of primary BAs segregated with MetS and SMetS. We directly show that reduced postprandial tolerance of glucose is associated with elevated primary BAs in both MetS and SMetS animals. Mirroring that of human MetS, higher relative abundance of primary versus secondary BAs promotes insulin resistance.[35] Our data suggest that BA signaling may be a permissive factor in the development of MetS in horses and that alterations in the BA profile may, in part, explain horse's periodicity of insulin dysregulation in SMetS. Although further study is required, compositional BA analysis may indicate animals at risk of developing ID and subsequently MetS. In the

current study, TCDCA concentrations were more closely correlated with insulin than that of TGs. This relationship is expected, as BAs have a direct effect on glucose homeostasis, whereas TGs are both a consequence and indirect reason for insulin dysregulation.[36] Importantly, we present a new avenue for MetS diagnostic and therapeutic research in horses with specific reference to BA dysregulation. Additionally, this feature elevates the horses as a naturally-occurring model of MetS for translational studies as it recapitulates the human syndrome with an additional layer of fidelity.[36]

Bile acid regulation is emblematic of the interaction between the host and microbiota. Microbial derived secondary bile acids are well characterized in humans[38] but have yet to be evaluated in horses. While lower in overall abundance, the secondary BA, DCA, tended to segregate with more metabolically normal individuals. DCA is the product of cholic acid metabolism, primarily from bacteria of the *Clostridium* genus.[26] DCA directly promotes insulin sensitivity and is under investigation as a supplement for MetS individuals.[39] Importantly, the relationship between BAs and the microbiota is bidirectional, with one affecting the other. To date, defining the microbiota population in obese and MetS horses has been challenging. Again, efforts have been hampered by small sample size and single sampling opportunities. On a microbiota population level, the current study does not differ greatly from previous studies, with no differences in the overall richness or evenness between MetS and REG individuals.[29] The current study does have the advantage of being able to specifically detect individual shifts in microbiota over time. In conjunction with BA dysregulation, we show that, at least at the family level, there is a relationship with Clostridiales and insulin regulation. This interaction appears to also show periodicity,

being detectable in spring and not fall. Additionally, we demonstrate that, with increased levels of primary BA, the microbiota becomes richer in Proteobacter. Proteobacter enterotypes have been associated with antibiotic use and diarrhea in horses and are thought to represent a dysbiosis shift.[27] Indeed, with antibiotic use, there is widespread reduction of Firmicutes, which include the family Clostridiales.[40] Conversely, not all Firmicutes promote glucose homeostasis. *Rumminococcaceae* have been highly and consistently associated as a pathobiont in the development of type 2 diabetes in humans.[41] *Rumminococcaceae* are a major constituent of the equine microbiota due to the high cellulose content of their diets.[12] Typically, this species is associated with a healthy microbiota.[41] Within a certain range, this likely still the case, but, in the current study, we suggest that when *Rumminococcaceae* predominates and falls out of regulation with other Firmicuts, such as *Lachnospiracciae,* it is associated with reduced glucose tolerance.

Collectively, we demonstrate a relationship between the microbiota, BA regulation and the development of phenotypes associated with equine MetS. Although this finding is promising, it must be tempered with the limitations of the techniques used. Without metagenomic and/or metatranscriptomic approaches, species identity and the necessary array of genes responsible for BA metabolism cannot be fully determined. Follow-up evaluation using techniques that allow more precise designation of microbiota constituents at the genus or species level is required to validate the current findings. Similarly, while it is a strength of this study that all individuals reside at the one institution, it also means that validation using additional individuals from different environments will be required to validate these findings in other environments.

The complex relationship between the host, microbiota and environment have been difficult to capture in the context of equine MetS. These data are the first to show that there is a potential microbiota-metabolome crosstalk that promotes a tendency to MetS. With our systems biology approach, we have been able to leverage a smaller sample size through deep phenotype and multiomic evaluation. Importantly, we demonstrate that bile acids are fundamentally associated with equine MetS and that this has both a host and microbial component. Future studies establishing the direct mechanisms of bile acid signaling are needed to develop new therapies around this novel pathophysiological insight.

Methods

Animals

Animals used in this experiment are a part of the Pioneer 100 cohort that has previously been described.[21] All procedures in this experiment were performed in accordance with the institutional care and use committee guidelines for the University of California, Davis (Protocol #21700). Body condition scoring (BCS) was performed as part of the mandatory deep phenotype protocol according to Henneke during Spring of 2020 and 2021.[23] All animals were scored by two experts with the consensus score recorded as the adiposity phenotype. Animals additionally had contemporaneous still photographs and videos at walking and trotting gaits recorded for verification of BCS and lameness phenotypes.

Endocrine Phenotypes

Animals were phenotype in Spring (May) and Fall (Oct-Nov) of 2020 and 2021. During spring phenotyping, insulin regulation and pars pituitary intermedia dysfunction (PPID) phenotypes were determined at the same sampling event according to Hodge et al (2019).[43] Insulin regulation was evaluated by oral sugar test (OST) as previously described.[43] Horses were acclimated to stalls and fasted overnight. Baseline serum and plasma samples were obtained between 6am and 7am for all horses in both Spring and Fall. In the Spring sampling, horses were administered 1 mg of thyrotropin releasing hormone (TRH) intravenously, with plasma sampled ten minutes following administration. Immediately following the TRH stimulation test, 0.15 mL/kg Lite Karo syrup was orally administered. Plasma and serum samples were obtained 60 and 90 minutes following oral sugar administration. In the Fall sampling, only OST was performed. All blood samples were immediately placed on ice and processed into plasma or serum (by centrifugation) within 30-60 minutes of sampling. Samples were stored at -80°C until analysis and analysed within six months of collection.

Plasma samples at baseline and at 10 minute post-TRH stimulation were analyzed for adrenocorticotropin releasing hormone (ACTH) using a chemiluminescent immunometric assay validated for use in horses (Diagnostic Products Cor, Los Angeles, CA, USA).[44] Plasma insulin concentration was measured at baseline, 60- and 90-minutes post oral sugar test. Insulin concentrations were determined by an equine validated immunoradiometric assay (Immuno-Biological Laboratories, Minneapolis, MN, USA).[45]

Targeted Metabolomics

Metabolomic analysis was performed on heparinized plasma samples obtained at the same time as endocrine phenotype samples for a subset of individuals (n=72). Three time points were evaluated for each individual spanning one year, with samples taken in November 2020, May 2021, and October-November 2021. Samples were analyzed using the Biocrates Quant500 (Biocrates Life Sciences AG, Innsbruk, Austria) quantitative targeted metabolomic panel. This panel includes 630 directly measured metabolites/lipids as well as 234 calculated values and indices. All samples were analyzed at the Center for Metabolomics, Institute of Metabolic Diseases, Baylor Scott & White Research Institute. Analysis of metabolomic data in this report used mean values derived from the three time points.

Outlier analysis was performed according to the method developed by Walatch (2020).[24] A false discovery rate (FDR) of 0.1 was used to capture outlying metabolites. Metabolite outliers with ≥ 5 individuals/metabolite were used in subsequent analysis. Outlier data analysis was performed using MetaboAnalyst 5.0 using the one factor package. Hierarchical analysis was performed using Euclidean distance with clustering by Ward's method. Sparse partial least squares discriminant analysis (sPLSDA) was used to reduce dimensionality of data and define prevailing metabolite groups.[46] Pairwise comparison between insulin regulation groups was performed by ANOVA, with Tukey's post-hoc testing and adjusting multiple comparison by FDR of 0.1.

Fecal Microbiota

Fecal samples were collected at five-time points; May 2020, November 2020, March 2021, May 2021 and October/November 2021. Free catch samples were used. Only samples where passage was observed were retained. Samples were immediately placed on ice and stored for transport and were transferred to -80°C within 3 hours of collection. Sample genomic DNA extraction and 16S rRNA sequencing was performed by Laragen Inc (Culver City, CA, USA). Paired-end sequencing of the V3-V4 region of the 16S domain was performed, with a target read length of 250-base pairs and a targeted depth of 100,000 reads per sample. All samples were sequenced on an Illumina MiSeq platform. 16S data was analyzed using the QIIME2 environment (QIIME2; https://qiime2.org). Demultiplexed reads were filtered for quality using the environment default parameters. Sequence variants were obtained and assigned sub-operational-taxonomic-units (sOTUs) with Deblur (V2022.2.0).[47] Sequences were classified against the Greengenes reference database (v13_8) and filtered at 97% sequence identity. Count tables with assigned taxonomy were downloaded for subsequent analysis.

Data Availability

Data available upon reasonable request.

References

1. Durham, A. E., Frank, N., McGowan, C. M., Menzies-Gow, N. J., Roelfsema, E., Vervuert, I., Feige, K., & Fey, K. (2019). ECEIM consensus statement on equine metabolic syndrome. *Journal of veterinary internal medicine*, *33*(2), 335–349. https://doi.org/10.1111/jvim.15423
2. Ahmed, M., Kumari, N., Mirgani, Z., Saeed, A., Ramadan, A., Ahmed, M. H., & Almobarak, A. O. (2022). Metabolic syndrome; Definition, Pathogenesis, Elements, and the Effects of medicinal plants on it's elements. *Journal of diabetes and metabolic disorders*, *21*(1), 1011–1022. https://doi.org/10.1007/s40200-021-00965-2
3. Thatcher, C. D., Pleasant, R. S., Geor, R. J., & Elvinger, F. (2012). Prevalence of overconditioning in mature horses in southwest Virginia during the summer. *Journal of veterinary internal medicine*, *26*(6), 1413–1418. https://doi.org/10.1111/j.1939-1676.2012.00995.x
4. Flegal, K. M., Kruszon-Moran, D., Carroll, M. D., Fryar, C. D., & Ogden, C. L. (2016). Trends in Obesity Among Adults in the United States, 2005 to 2014. *JAMA*, *315*(21), 2284–2291. https://doi.org/10.1001/jama.2016.6458
5. Hild, B., Dreier, M. S., Oh, J. H., McCulloch, J. A., Badger, J. H., Guo, J., Thefaine, C. E., Umarova, R., Hall, K. D., Gavrilova, O., Rosshart, S. P., Trinchieri, G., & Rehermann, B. (2021). Neonatal exposure to a wild-derived microbiome protects mice against diet-induced obesity. *Nature metabolism*, *3*(8), 1042–1057. https://doi.org/10.1038/s42255-021-00439-y
6. Sonnenburg, J. L., & Sonnenburg, E. D. (2019). Vulnerability of the industrialized microbiota. *Science (New York, N.Y.)*, *366*(6464), eaaw9255. https://doi.org/10.1126/science.aaw9255

7. Clemente, J. C., Pehrsson, E. C., Blaser, M. J., Sandhu, K., Gao, Z., Wang, B., Magris, M., Hidalgo, G., Contreras, M., Noya-Alarcón, Ó., Lander, O., McDonald, J., Cox, M., Walter, J., Oh, P. L., Ruiz, J. F., Rodriguez, S., Shen, N., Song, S. J., Metcalf, J., … Dominguez-Bello, M. G. (2015). The microbiome of uncontacted Amerindians. *Science advances*, *1*(3), e1500183. https://doi.org/10.1126/sciadv.1500183
8. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, Segata N. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell. 2019 Jan 24;176(3):649-662.e20. doi: 10.1016/j.cell.2019.01.001. Epub 2019 Jan 17. PMID: 30661755; PMCID: PMC6349461.
9. Rosshart SP, Vassallo BG, Angeletti D, Hutchinson DS, Morgan AP, Takeda K, Hickman HD, McCulloch JA, Badger JH, Ajami NJ, Trinchieri G, Pardo-Manuel de Villena F, Yewdell JW, Rehermann B. Wild Mouse Gut Microbiota Promotes Host Fitness and Improves Disease Resistance. Cell. 2017 Nov 16;171(5):1015-1028.e13. doi: 10.1016/j.cell.2017.09.016. Epub 2017 Oct 19. PMID: 29056339; PMCID: PMC6887100.
10. Rosshart, S. P., Herz, J., Vassallo, B. G., Hunter, A., Wall, M. K., Badger, J. H., McCulloch, J. A., Anastasakis, D. G., Sarshad, A. A., Leonardi, I., Collins, N., Blatter, J. A., Han, S. J., Tamoutounour, S., Potapova, S., Foster St Claire, M. B., Yuan, W., Sen, S. K., Dreier, M. S., Hild, B., … Rehermann, B. (2019). Laboratory mice born to wild mice have natural microbiota and model human

immune responses. *Science (New York, N.Y.)*, *365*(6452), eaaw4361. https://doi.org/10.1126/science.aaw4361

11. Orlando, L. (2020). Ancient genomes reveal unexpected horse domestication and management dynamics. *BioEssays*, *42*(1), 1900164.

12. Kauter, A., Epping, L., Semmler, T., Antao, E. M., Kannapin, D., Stoeckle, S. D., Gehlen, H., Lübke-Becker, A., Günther, S., Wieler, L. H., & Walther, B. (2019). The gut microbiome of horses: current research on equine enteral microbiota and future perspectives. *Animal microbiome*, *1*(1), 14. https://doi.org/10.1186/s42523-019-0013-3

13. Payab, M., Tayanloo-Beik, A., Falahzadeh, K., Mousavi, M., Salehi, S., Djalalinia, S., Ebrahimpur, M., Rezaei, N., Rezaei-Tavirani, M., Larijani, B., Arjmand, B., & Gilany, K. (2021). Metabolomics prospect of obesity and metabolic syndrome; a systematic review. *Journal of diabetes and metabolic disorders*, *21*(1), 889–917. https://doi.org/10.1007/s40200-021-00917-w

14. Agus, A., Clément, K., & Sokol, H. (2021). Gut microbiota-derived metabolites as central regulators in metabolic disorders. *Gut*, *70*(6), 1174–1182. https://doi.org/10.1136/gutjnl-2020-323071

15. Fromentin, S., Forslund, S. K., Chechi, K., Aron-Wisnewsky, J., Chakaroun, R., Nielsen, T., Tremaroli, V., Ji, B., Prifti, E., Myridakis, A., Chilloux, J., Andrikopoulos, P., Fan, Y., Olanipekun, M. T., Alves, R., Adiouch, S., Bar, N., Talmor-Barkan, Y., Belda, E., Caesar, R., … Pedersen, O. (2022). Microbiome and metabolome features of the cardiometabolic disease spectrum. *Nature medicine*, *28*(2), 303–314. https://doi.org/10.1038/s41591-022-01688-4

16. Hu, X., Zhou, R., Li, H., Zhao, X., Sun, Y., Fan, Y., & Zhang, S. (2022). Alterations of Gut Microbiome and Serum Metabolome in Coronary Artery

Disease Patients Complicated With Non-alcoholic Fatty Liver Disease Are Associated With Adverse Cardiovascular Outcomes. *Frontiers in cardiovascular medicine*, *8*, 805812. https://doi.org/10.3389/fcvm.2021.805812

17. Vallianou, N., Christodoulatos, G. S., Karampela, I., Tsilingiris, D., Magkos, F., Stratigou, T., Kounatidis, D., & Dalamaga, M. (2021). Understanding the Role of the Gut Microbiome and Microbial Metabolites in Non-Alcoholic Fatty Liver Disease: Current Evidence and Perspectives. *Biomolecules*, *12*(1), 56. https://doi.org/10.3390/biom12010056

18. Price ND, Magis AT, Earls JC, et al. A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nat Biotechnol*. 2017;35(8):747-756. doi:10.1038/nbt.3870

19. Wilmanski, T., Diener, C., Rappaport, N., Patwardhan, S., Wiedrick, J., Lapidus, J., Earls, J. C., Zimmer, A., Glusman, G., Robinson, M., Yurkovich, J. T., Kado, D. M., Cauley, J. A., Zmuda, J., Lane, N. E., Magis, A. T., Lovejoy, J. C., Hood, L., Gibbons, S. M., Orwoll, E. S., … Price, N. D. (2021). Gut microbiome pattern reflects healthy ageing and predicts survival in humans. *Nature metabolism*, *3*(2), 274–286. https://doi.org/10.1038/s42255-021-00348-0

20. Wilmanski, T., Rappaport, N., Earls, J. C., Magis, A. T., Manor, O., Lovejoy, J., Omenn, G. S., Hood, L., Gibbons, S. M., & Price, N. D. (2019). Blood metabolome predicts gut microbiome α-diversity in humans. *Nature biotechnology*, *37*(10), 1217–1228. https://doi.org/10.1038/s41587-019-0233-9

21. Donnelly, C.G., S. Peng, V. Rodriguez, A. Nguyen, K.E. Knickelbein, E Williams Louie, J.M. Morgan, M Hammond, R.R. Bellone[1,5,] N.D. Price, Carrie J Finno (2022). The Pioneer 100 Horse Health Project: A blueprint for a systems biology approach to veterinary precision medicine. *In Submission*

22. Equine Endocrinology Group : https://sites.tufts.edu/equineendogroup/. Accessed May 6, 2022

23. Henneke, D. R., Potter, G. D., Kreider, J. L., & Yeates, B. F. (1983). Relationship between condition score, physical measurements and body fat percentage in mares. *Equine veterinary journal*, *15*(4), 371-372.

24. Walach J, Filzmoser P, Kouřil Š, Friedecký D, Adam T. Cellwise outlier detection and biomarker identification in metabolomics based on pairwise log ratios. J Chemom. 2020 Jan;34(1):e3182. doi: 10.1002/cem.3182. Epub 2019 Dec 2. PMID: 32189829; PMCID: PMC7063692.

25. Ridlon, J. M., Kang, D. J., Hylemon, P. B., & Bajaj, J. S. (2014). Bile acids and the gut microbiome. *Current opinion in gastroenterology*, *30*(3), 332–338. https://doi.org/10.1097/MOG.0000000000000057

26. Doden, H. L., Wolf, P. G., Gaskins, H. R., Anantharaman, K., Alves, J., & Ridlon, J. M. (2021). Completion of the gut microbial epi-bile acid pathway. *Gut microbes*, *13*(1), 1–20. https://doi.org/10.1080/19490976.2021.1907271

27. Shin, N. R., Whon, T. W., & Bae, J. W. (2015). Proteobacteria: microbial signature of dysbiosis in gut microbiota. *Trends in biotechnology*, *33*(9), 496–503. https://doi.org/10.1016/j.tibtech.2015.06.011

28. Delarocque, J., Frers, F., Feige, K., Huber, K., Jung, K., & Warnken, T. (2021). Metabolic changes induced by oral glucose tests in horses and their diagnostic use. *Journal of veterinary internal medicine*, *35*(1), 597–605. https://doi.org/10.1111/jvim.15992

29. Coleman, M. C., Whitfield-Cargile, C. M., Madrigal, R. G., & Cohen, N. D. (2019). Comparison of the microbiome, metabolome, and lipidome of obese and

non-obese horses. *PloS one*, *14*(4), e0215918. https://doi.org/10.1371/journal.pone.0215918

30. Delarocque, J., Reiche, D. B., Meier, A. D., Warnken, T., Feige, K., & Sillence, M. N. (2021). Metabolic profile distinguishes laminitis-susceptible and -resistant ponies before and after feeding a high sugar diet. *BMC veterinary research*, *17*(1), 56. https://doi.org/10.1186/s12917-021-02763-7

31. Patterson Rosa, L., Mallicote, M. F., Long, M. T., & Brooks, S. A. (2020). Metabogenomics reveals four candidate regions involved in the pathophysiology of Equine Metabolic Syndrome. *Molecular and cellular probes*, *53*, 101620. https://doi.org/10.1016/j.mcp.2020.101620

32. Shapiro, H., Kolodziejczyk, A. A., Halstuch, D., & Elinav, E. (2018). Bile acids in glucose metabolism in health and disease. *The Journal of experimental medicine*, *215*(2), 383–396. https://doi.org/10.1084/jem.20171965

33. Elzinga, S., Wood, P., & Adams, A. A. (2016). Plasma lipidomic and inflammatory cytokine profiles of horses with equine metabolic syndrome. *Journal of Equine Veterinary Science*, *40*, 49-55.

34. Washizu, T., Tomoda, I., & Kaneko, J. J. (1991). Serum bile acid composition of the dog, cow, horse and human. *The Journal of veterinary medical science*, *53*(1), 81–86. https://doi.org/10.1292/jvms.53.81

35. Wu, T., Bound, M. J., Standfield, S. D., Jones, K. L., Horowitz, M., & Rayner, C. K. (2013). Effects of taurocholic acid on glycemic, glucagon-like peptide-1, and insulin responses to small intestinal glucose infusion in healthy humans. *The Journal of clinical endocrinology and metabolism*, *98*(4), E718–E722. https://doi.org/10.1210/jc.2012-3961

36. Grundy S. M. (1999). Hypertriglyceridemia, insulin resistance, and the metabolic syndrome. *The American journal of cardiology*, *83*(9B), 25F–29F. https://doi.org/10.1016/s0002-9149(99)00211-8

37. Ramírez-Pérez, O., Cruz-Ramón, V., Chinchilla-López, P., & Méndez-Sánchez, N. (2017). The Role of the Gut Microbiota in Bile Acid Metabolism. *Annals of hepatology*, *16*(Suppl. 1: s3-105.), s15–s20. https://doi.org/10.5604/01.3001.0010.5494

38. Figueiredo, G., Gomes, M., Covas, C., Mendo, S., & Caetano, T. (2022). The Unexplored Wealth of Microbial Secondary Metabolites: the Sphingobacteriaceae Case Study. *Microbial ecology*, *83*(2), 470–481. https://doi.org/10.1007/s00248-021-01762-3

39. Stacpoole P. W. (1989). The pharmacology of dichloroacetate. *Metabolism: clinical and experimental*, *38*(11), 1124–1144. https://doi.org/10.1016/0026-0495(89)90051-6

40. Costa, M. C., Arroyo, L. G., Allen-Vercoe, E., Stämpfli, H. R., Kim, P. T., Sturgeon, A., & Weese, J. S. (2012). Comparison of the fecal microbiota of healthy horses and horses with colitis by high throughput sequencing of the V3-V5 region of the 16S rRNA gene. *PloS one*, *7*(7), e41484. https://doi.org/10.1371/journal.pone.0041484

41. Gurung, M., Li, Z., You, H., Rodrigues, R., Jump, D. B., Morgun, A., & Shulzhenko, N. (2020). Role of gut microbiota in type 2 diabetes pathophysiology. *EBioMedicine*, *51*, 102590. https://doi.org/10.1016/j.ebiom.2019.11.051

42. Julliand, V., & Grimm, P. (2016). HORSE SPECIES SYMPOSIUM: The microbiome of the horse hindgut: History and current knowledge. *Journal of animal science*, *94*(6), 2262–2274. https://doi.org/10.2527/jas.2015-0198

43. Hodge, E., Kowalski, A., Torcivia, C., Lindborg, S., Stefanovski, D., Hart, K., Frank, N., & van Eps, A. (2019). Effect of thyrotropin-releasing hormone stimulation testing on the oral sugar test in horses when performed as a combined protocol. *Journal of veterinary internal medicine*, *33*(5), 2272–2279. https://doi.org/10.1111/jvim.15601

44. Beech, J., McFarlane, D., Lindborg, S., Sojka, J. E., & Boston, R. C. (2011). α-Melanocyte--stimulating hormone and adrenocorticotropin concentrations in response to thyrotropin-releasing hormone and comparison with adrenocorticotropin concentration after domperidone administration in healthy horses and horses with pituitary pars intermedia dysfunction. *Journal of the American Veterinary Medical Association*, *238*(10), 1305–1315. https://doi.org/10.2460/javma.238.10.1305

45. Kerrigan, L. E., Thompson, D. L., Jr, Chapman, A. M., & Oberhaus, E. L. (2020). Effects of Epinephrine, Detomidine, and Butorphanol on Assessments of Insulin Sensitivity in Mares. *Journal of equine veterinary science*, *85*, 102842. https://doi.org/10.1016/j.jevs.2019.102842

46. Chong, J., Wishart, D. S., & Xia, J. (2019). Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. *Current protocols in bioinformatics*, *68*(1), e86. https://doi.org/10.1002/cpbi.86

47. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. mSystems. 2017

Mar 7;2(2):e00191-16. doi: 10.1128/mSystems.00191-16. PMID: 28289731; PMCID: PMC5340863.

Chapter 3: Vitamin E depletion is associated with subclinical axonal degeneration in juvenile horses

Authors: Callum G. Donnelly[1] and Carrie J. Finno[1]

[1] Department of Population Health and Reproduction, School of Veterinary Medicine, University of California, Davis

Keywords: Neuroaxonal degeneration, juvenile, biomarker, alpha-tocopherol, neurofilament

Under Review, Equine Veterinary Journal

Abstract

Background

Phosphorylated neurofilament heavy (pNfH), a marker of neuroaxonal damage, is elevated in horses with equine neuroaxonal dystrophy (eNAD). However, the temporal dynamics of this biomarker during the post-natal risk period are not understood.

Objective

To measure serum and cerebrospinal fluid (CSF) pNfH concentrations in juvenile foals across the post-natal window of susceptibility for eNAD.

Study Design

Case-control experimental study.

Methods

pNfH concentrations were measured using frozen serum and CSF collected from thirteen foals raised in a vitamin E deficient environment from one to six months of age. Four of these foals were produced by eNAD-affected dams, developed clinical signs consistent with eNAD and had a diagnosis confirmed by histopathology. The remaining nine foals, produced by healthy mares, were vitamin E depleted and remained clinically healthy. An additional cohort of foals, produced by healthy mares, were supplemented with vitamin E (α-tocopherol; α-TOH) from birth and sampled similarly.

Results

Serum α-TOH concentrations were significantly higher in vitamin E supplemented healthy foals. Serum pNfH concentrations did not differ significantly between groups at any time point. CSF pNfH concentrations increased with age in healthy vitamin E depleted foals ($p=0.0009$); an effect that was not observed in healthy vitamin E supplemented foals.

Main limitations

A genetically susceptible cohort supplemented with vitamin E was not available for comparison to clinically affected vitamin E deplete foals.

Conclusion

We demonstrate that vitamin E depletion may elevate CSF pNfH in otherwise healthy juvenile foals by six months of age. We highlight an important cofactor to consider when interpreting CSF pNfH concentrations in juvenile horses.

Introduction

Equine neuroaxonal dystrophy (eNAD) is a neurodegenerative disease of juvenile horses, resulting in damage to spinal sensory tracts.[1] It manifests clinically as spinal ataxia with profound proprioceptive deficits and is currently the second most common diagnosis for equine spinal ataxia.[2] It was first identified in foals from New York state in the late 1970's and early 1980's and has long been associated with vitamin E deficiency.[3,4] The disease bears many similar features to that of Ataxia with Vitamin E deficiency (AVED) of humans (OMIM #277460;(2)). Whilst the genetic background differs, it appears that the molecular drivers of neuroaxonal degeneration are largely conserved.[5-7] The role of vitamin E in protection from clinical disease in genetically susceptible juvenile humans and horses has been well demonstrated.[4,8] Early studies in foals found that supplementation with vitamin E, in the form of α-tocopherol, significantly reduced the prevalence of eNAD in animals with a high suspicion of genetic susceptibility.[4] The impact of vitamin E depletion on the neurologic health of horses not genetically susceptible to eNAD has been less well characterized. While it has presumed that both susceptibility and deficiency are required for neuroaxonal degeneration, this has yet to be examined in the context of eNAD *in vivo.*

Currently, the genetic etiology of eNAD is unknown and a definitive diagnosis requires histologic examination of the spinal cord and brainstem.[1] Recent advances have been made towards the antemortem diagnosis of the condition using biomarkers of axonal degeneration and metabolites of α-tocopherol.[9,10] Phosphorylated neurofilament heavy (pNfH) has shown promise as a diagnostic aid for animals suspected of eNAD.[9] Phosphorylated NfH is a cytoskeletal protein of the axon and, with axonal damage, it is released into the cerebrospinal fluid (CSF) and serum. Use of pNfH as a diagnostic tool has previously been demonstrated in children with encephalitis/encephalopathy with reversible splenial lesions[11], bacterial meningitis[12], hypoxic ischemic encephalopathy[13], children with cardiac lesions[14,15] and pre-term infants[16]. The use of this marker has not been evaluated in individuals with AVED, nor in the context of vitamin E deficiency in children. Additionally, given the invasive nature of CSF collection in children, a relevant animal model of juvenile neurodegeneration that can be repeatedly sampled has tremendous translational benefits.

Age is an important interacting factor to consider when assessing the pNfH concentration, especially in CSF. In normal individuals, neurofilament concentrations will increase with age, either from reduced CSF turnover or low-level age-related axon degeneration.[17] Similar results were identified in healthy horses by Edwards et al (2021), with age weakly correlated to serum pNfH and moderately correlated to CSF pNfH.[9] However, in neurologically abnormal horses, age was inversely correlated to CSF pNfH concentrations.[9] Since many of these neurologically abnormal horses included horses affected with eNAD, the question arose if vitamin E

depletion alone, without concurrent apparent neurologic disease, could lead to elevations in CSF pNfH concentrations early in life.

Vitamin E inadequacy is a pervasive health issue amongst people and animals.[18] There is evidence that vitamin E inadequacy, as measured by circulating α-tocopherol concentrations, may be increasing in prevalence amongst humans.[19] The ability to study both naturally occurring neurologic disease and vitamin E deficiency in a relevant animal model has strong translational benefits. Therefore, the aim of this study was two-fold. Firstly, to evaluate serum and CSF pNfH concentrations over time for the first six months of life in a cohort of eNAD susceptible and non-susceptible foals. Secondly, we aimed to evaluate the impact of vitamin E depletion on serum and CSF pNfH concentrations in otherwise healthy foals throughout post-natal development. We hypothesized that foals with a genetic propensity for eNAD will have elevated serum and CSF pNfH concentrations when compared to neurologically normal foals. We further hypothesized that vitamin E depletion will be associated with elevated serum and CSF pNfH concentrations in otherwise neurologically healthy foals.

Materials and Methods

Study Design

Case-control study. Samples from foals previously described by Finno et al (2015) were available for analysis in this study.[20] Briefly, thirteen American Quarter Horse foals, born in 2010 (n=4), 2011 (n=7) and 2012 (n=2), were maintained in vitamin E deplete conditions from birth to one year of age. Dietary vitamin E depletion was induced by feeding only hay and a vitamin/mineral concentrate that did not contain

vitamin E to the dams and foals.[20] Foals did not have access to grass pasture. This diet met all the Nutritional Research Council recommendations for growing foals[21] other than vitamin E intake.

Four of these foals (n=1 male, n=3 female) were produced by mares with a high suspicion of eNAD as the mares had neurologic deficits and three of the four dams had previously produced necropsy-confirmed eNAD affected foals. These four foals developed clinical signs consistent with eNAD by six months of age. The remaining nine vitamin E depleted foals, produced by healthy mares, remained neurologically appropriate (n=5 males, n=4 females).
For the first year of life, serum and CSF samples were collected every 1-2 months.

Cerebrospinal fluid was collected via atlantooccipital centesis under general anesthesia. All CSF samples underwent routine fluid analysis (**Supplement Table 3 S1).** Foals that developed clinical signs typical of eNAD (symmetrical proprioceptive deficits) were humanely euthanized to confirm the diagnosis of eNAD by histopathology as described by Aleman et al (2011).[22] Given that age at sampling varied across the three years, post-natal days 30, 60 and 180 were selected for further analysis in order to optimise sample sizes across time points. Samples were stored at -80ºC until analysis.

Since both the original eNAD and healthy foals were raised under vitamin E deplete conditions, CSF and serum samples from an additional cohort of five foals orally supplemented with 10 IU/kg of water dispersible RRR-α-tocopherol (Elevate-WD, Kentucky Equine Research) from birth were collected at post-natal day 30, 60 and 180. Samples were collected from three American Quarter Horse foals born in 2020

Table 3 S1. Cerebrospinal fluid (CSF) fluid analysis summary. Values are median (range). TNCC=total nucleated cell count, RBC=red blood cell. NA=not assessed

	CSF TNCC (cells/μL)			CSF RBC (cells/μL)			CSF Protein (g/dL)		
Group	**Day 30**	**Day 60**	**Day 180**	**Day 30**	**Day 60**	**Day 180**	**Day 30**	**Day 60**	**Day 180**
Healthy Vitamin E Supplemented	0 (0-1)	1 (0-2)	2 (2-7)	0 (0-17)	3 (0-34)	7 (0-112)	75 (56-90)	65 (49-78)	72 (52-90)
Healthy Vitamin E Depleted	0 (0-21)	1 (0-18)	0 (0-4)	2 (1-59)	4 (0-9555)	0 (0-2290)	NA	NA	NA
eNAD	0.5 (0-1)	5.5 (0-12)	0 (0-1)	0 (0-1)	43.5 (0-910)	0.5 (0-176)	NA	NA	NA

(n=1 male, n=2 females) and two male Connemara cross foals born in 2021. These foals were produced by healthy mares. All foals were examined every 1-3 months for the first year of life for evidence of neurologic disease.

Sample size for the vitamin E deplete groups were predetermined as this study used banked samples.[20] The prospective cohort sample size was determined by the maximum number of available foals raised at the same facility and born over a similar time frame as the Vitamin E deplete cohorts.

Serum Alpha-tocopherol Analysis

Serum alpha-tocopherol concentrations were analysed by high-performance liquid chromatography with fluorescence detection as previously described (18).

ELISA Analysis

All samples were analysed in 2020-2021 using a validated sandwich ELISA for pNfH (EnCor Biotechnology Inc, Gainesville, FL) on paired serum and CSF samples from all animals according to Edwards et al (2020).[9] Briefly, paired serum and CSF samples from each animal were analysed on the same plate. The lower and upper limit of detection for this assay is 0.07 ng/ml and 10 ng/ml, respectively. Samples for serum were diluted to 1:2.5 and CSF to 1:5, to fall within the detection range. A standard curve using pNfH derived from lyophilized bovine spinal cord and blank sample diluent was run on every plate. Optical density was measured using a solid-state plate reader (Byonoy, Germany) at 450nm. Samples were analysed in duplicate, with the mean sample OD used to derive the sample value, provided that the sample OD coefficient of variation was $\leq$15%. Quantitative values were determined by linear regression fitting of the standard curve. Where values were below the limit of detection of 0.07 ng/mL, they were imputed to 0.07 ng/mL for analysis (9). These values, where applicable, will be given as not detectable (ND).

Statistical Analysis

Data were analysed for normality using Shapiro-Wilk test. Normally distributed data were evaluated by 2-way repeated measures ANOVA. Non-normally distributed data were analysed by repeated measures ANOVA using the rank aligned transformation method.[23] Values for serum pNFh were not significantly different for depleted foals, regardless of presence of eNAD, therefore this group was also evaluated as pooled

deficient group. Turkey's test was used for post-hoc multiple comparisons with Holm correction for multiple comparisons. A p-value of <0.05 was considered statistically significant. Analysis was performed with RStudio (Rstudio, Inc., Boston, MA). Figures were generated with Prism version 9.2.0 (GraphPad Software Inc, San Diego, CA)

Results

Serum Alpha-tocopherol

Serum alpha-tocopherol concentrations were higher in vitamin E supplemented healthy foals at all time points compared to all other time points from healthy vitamin E depleted foals and eNAD foals (**Fig. 3.1**). A significant group x time interaction was detected (P<0.0001), with group attributing 69.72% of the total variance (P<0.0001).

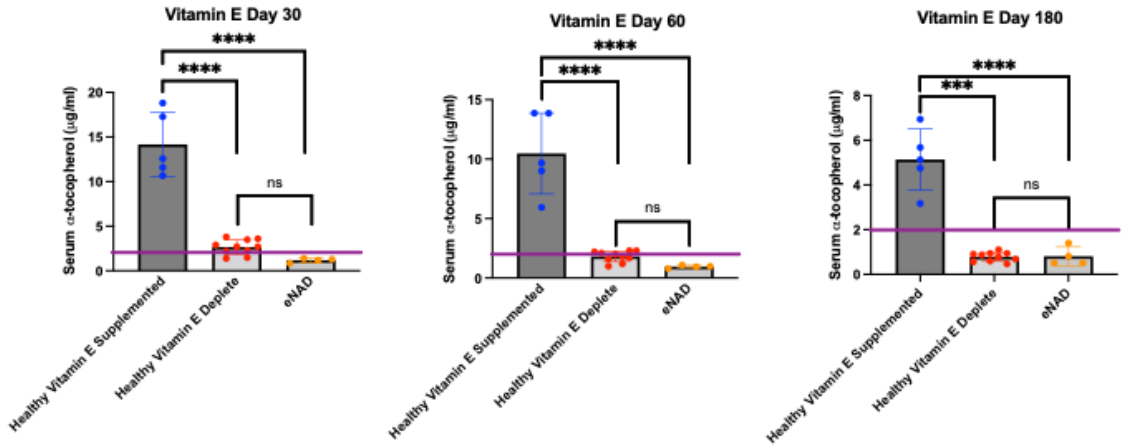

Figure 3.1 Serum vitamin E (alpha-tocopherol) concentrations in juvenile horses. Alpha-tocopherol concentrations were significantly higher in healthy supplemented foals and all other foals at all time points. Healthy vitamin E depleted foals and eNAD foals were not significantly different from one another at any timepoint. Note differing Y-axes between groups. The purple line denotes the lower threshold for normal serum alpha-tocopherol concentrations in normal horses as defined by Finno et al (2015;20). ***p<0.001, ****p<0.0001.

pNfH ELISA Performance

Mean intraassay coefficient of variation (CV) plate standard OD was 9.89% (range 4.81%-14.0%). Mean intraassay CV of OD for serum and CSF samples were 10.4% (range 7.8%-15.9%) and 8.12% (7.45%-9.47%), respectively.

Serum pNfH Concentrations

There was no significant effect of group or age (30, 60 and 180 days) on serum pNfH concentrations (**Supplement Table 3 S2**). All healthy supplemented foals had serum pNfH concentrations below the limit of detection at all time points. When all samples derived from vitamin E depleted foals were pooled and compared to vitamin E supplemented foals, there was no significant effect of vitamin E supplementation on serum pNfH concentrations (P=0.14).

Table 3 S2. Serum phosphorylated neurofilament heavy concentrations in juvenile horses by age at measurement. All values are median (range). Values are in ng/mL. ND=below the limit of detection for the assay

Group	**Day 30**	**Day 60**	**Day 180**
Healthy Vitamin E Supplemented	ND	ND	ND
Healthy Vitamin E Depleted	ND (ND-1.39)	ND (ND-2.87)	ND (ND-5.4)
eNAD	ND	0.17 (ND-0.61)	0.86 (ND-8.16)

CSF pNfH Concentrations

Post-natal day 30, day 60 and day 180 CSF pNfH concentrations were analysed. A significant effect of group (p=0.004), age (p=0.006) and group x age interaction (p=0.02) were detected; however, post-hoc analyses did not identify any significant pairwise comparisons **(Fig. 3.2)**. To examine the effect of vitamin E depletion on

healthy foals with aging, the effect of time was analysed within the healthy vitamin E depleted and healthy vitamin E supplemented groups. Cerebrospinal fluid pNfH concentrations significantly

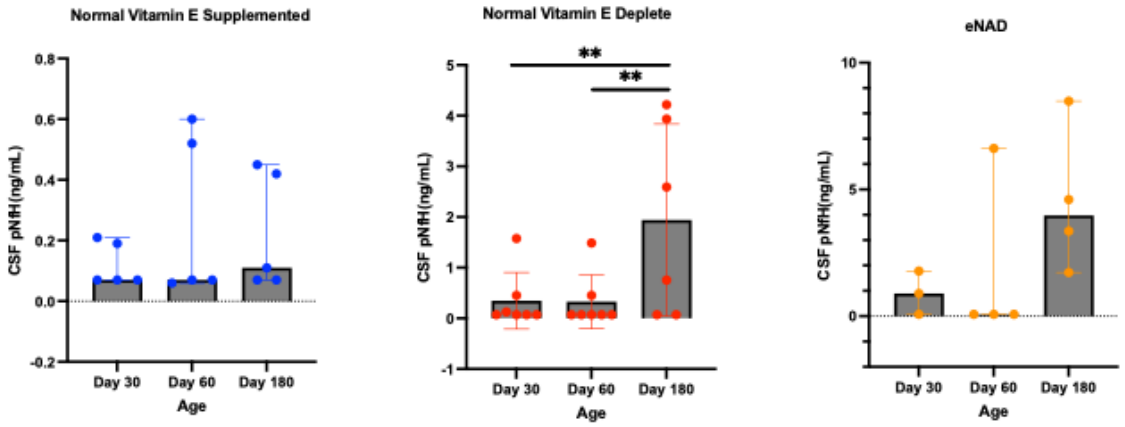

Figure 3.2. Cerebrospinal fluid (CSF) concentration of phosphorylated neurofilament-heavy (pNfH) in healthy vitamin E supplemented juvenile horses, healthy vitamin E depleted juvenile horses and juvenile horses with necropsy confirmed equine neuroaxonal dystrophy. Note differing Y-axes between groups. Repeated measures ANOVA, following rank aligned transformation, identified a significant effect of group *(p*=0.004*),* age (*p*=0.006) and group x age interaction (*p*=0.022); however, post-hoc analyses did not identify any significant pairwise comparisons. Post-hoc testing identified significant increases in CSF pNfH concentrations in healthy vitamin E depleted foals at day 180 when compared to day 30 and day 60. **p<0.01.

increased in healthy vitamin E depleted foals with age **(**p=0.0009; **Fig 3.2)**. Post-hoc testing identified significant increases in CSF pNfH concentrations in healthy vitamin E depleted foals at day 180 when compared to day 30 (P=0.01) and day 60 (P=0.01). This increase in CSF pNfH concentrations was not observed in healthy vitamin E supplemented foals (**Fig 3.2)**. Foals with eNAD demonstrated a decrease in CSF pNfH concentrations at day 60, followed by an increase at day 180; an effect that was not significant (**Fig 3.2)**.

Discussion

The present study demonstrates a distinct temporal association between vitamin E depletion and the axonal damage specific marker pNfH in juvenile horses. By the age of six months, CSF pNfH concentrations in foals definitively diagnosed with eNAD diverge from both healthy vitamin E depleted foals and healthy vitamin E supplemented foals. Importantly, we also demonstrate that vitamin E depleted foals, without apparent neurologic disease, also have elevations in CSF pNfH during the critical window for post-natal development. Thus, subclinical axonal degeneration may occur with vitamin E deficiency early in life.

Neurofilaments as markers of axonal damage are susceptible to several drivers.[17] Previous work in horses demonstrated the same positive association between age and pNfH that has been demonstrated in aging humans.[17,24] Vitamin E depleted foals included in the current study were also included in Edwards et al (2021) however only the samples derived at 4 days of age were included.[9] The current study confirms that foals in vitamin E deficient conditions have an accelerated accrual of pNfH in CSF, with detection as early as two months and consistently by six months of age.

It is of note that, in this study, foals that developed the neurologic consequence of vitamin E deficiency (eNAD) and those that did not both had elevations in CSF pNfH. A similar pattern of elevation has been documented in children with bacterial meningitis, regardless of whether they develop long-term neurologic deficits or have apparent resolution.[12] In both cases, the magnitude of increase is higher and sustained in those individuals with neurologic sequelae. Importantly, it indicates that,

even without detectable neurologic deficits, there may be neuroaxonal damage/degeneration associated with vitamin E depletion. This is a novel finding during post-natal development and indicates that, regardless of susceptibility or long-term consequence, adequate vitamin E concentrations are required in the first months of life to prevent neuroaxonal damage. Speculatively for eNAD, this may indicate that animals that go onto develop clinical disease have a reduced capacity for damage, enabling the progression to clinical disease. This has been suggested previously, where administration of α-tocopherol in genetically susceptible herds reduced the prevalence of clinical disease.[4] Additionally, diagnostic use of this assay in juvenile horses may need to be tempered with the concurrent measurement of serum alpha-tocopherol concentrations. Despite the accelerated accrual, these data do not indicate that an age-specific reference range be developed for CSF when applied to aid in the diagnosis of eNAD.[9]

Many samples failed to reach the limit of detection, especially serum samples. Since neurofilaments are robust to handling conditions, including repeated freeze-thaw, it is likely that this failure is due to the sensitivity of the assay rather than sample quality.[25] In serum, however, either higher levels of fibrin products in circulation or activation of remaining fibrinogen on thawing may directly interfere with antibody-antigen binding.[26,27] Therefore, interference by the sample matrix cannot be completely ruled out. More sensitive platforms, such as single molecule analysis (Simoa), would allow for enhanced detection of pNfH in serum.[17] More accurate assessment of serum concentrations may improve the diagnostic capability of pNfH for eNAD in juvenile horses. While foals with eNAD had higher serum concentrations at 180 days, this study did not have sufficient power to demonstrate a difference

from the other groups. Data from the current study indicates that serum is currently not a suitable matrix for diagnostic use in juvenile horses.

This study controls for the effect of vitamin E through the inclusion of a supplemented group of foals. A weakness, however, is that it does not include eNAD susceptible foals that were also supplemented with α-tocopherol from birth. Given that the putative genetic variant(s) for this condition are yet to be determined in horses, purposeful production of such foals is challenging.[2] The foals in the original study were all from mothers exhibiting clinical traits consistent with eNAD that had previously produced foals with confirmed eNAD. These same mares were too old to safely carry prospective pregnancies at the time pNfH concentrations were evaluated. Additionally, as all progeny with eNAD required histologic diagnosis, there were no younger animals to replace the original breeding population. Therefore, it was not possible to assess pNfH concentrations over time in susceptible foals without the induction of vitamin E depletion. Previous evidence would suggest that the clinical phenotype may be prevented[4], however, whether axonal degeneration occurs despite supplementation remains unknown.

Data from the current study indicates that subclinical axonal degeneration is associated with vitamin E deficiency in juvenile horses. Degeneration was more pronounced in foals that subsequently developed clinical signs consistent with eNAD, and significantly diverged from foals that were supplemented with alpha-tocopherol by six months of age. These finding helps to refine the use of this biomarker in horses, but also has translational implications for human health given the shared molecular mechanisms of degeneration driven by vitamin E depletion.

Acknowledgments

The authors would like to acknowledge the technical staff at the University of California, Davis Center for Equine Health for their care of the animals and assistance with sample collection.

Funding

This study was supported by the University of California, Davis Center for Equine Health with funds provided by the State of California parimutuel fund. Graduate support was provided by private donation. Support for CJF was provided by the National Institutes of Health (NIH) (L40 TR001136).

Abbreviations

α-TOH: Alpha tocopherol

AVED**:** Ataxia with Vitamin E Deficiency

CSF**:** Cerebrospinal Fluid

CV: Coefficient of Variation

eNAD: Equine Neuroaxonal Dystrophy

ND: Not Detectable

OD: Optical Density

pNfH: phosphorylate Neurofilament Heavy

Availability of data and materials

Available by request.

Ethics Approval

All procedures were performed in accordance with the institutional animal care and use committee at the University of California (Protocol #21343).

Consent for Publication

Not applicable.

Competing Interests

Both authors declare no conflict of interest.

Author Contribution

CGD and CJF carried out sample collection. CGD carried our ELISA experiments. CGD and CJF analysed and interpreted data. Manuscript prepared by CGD and edited by CJF. Both authors read and approved the final manuscript.

References

1. Finno, C. J., Valberg, S. J., Shivers, J., D'Almeida, E., & Armién, A. G. (2016). Evidence of the Primary Afferent Tracts Undergoing Neurodegeneration in Horses With Equine Degenerative Myeloencephalopathy Based on Calretinin Immunohistochemical Localization. *Veterinary pathology*, *53*(1), 77–86. https://doi.org/10.1177/0300985815598787
2. Burns, E. N., & Finno, C. J. (2018). Equine degenerative myeloencephalopathy: prevalence, impact, and management. *Veterinary medicine (Auckland, N.Z.)*, *9*, 63–67. https://doi.org/10.2147/VMRR.S148542
3. Mayhew, I. G., deLahunta, A., Whitlock, R. H., & Geary, J. C. (1977). Equine degenerative myeloencephalopathy. *Journal of the American Veterinary Medical Association*, *170*(2), 195–201.
4. Beech J. (1984). Neuroaxonal dystrophy of the accessory cuneate nucleus in horses. *Veterinary pathology*, *21*(4), 384–393. https://doi.org/10.1177/030098588402100404
5. Finno, C. J., Peterson, J., Kang, M., Park, S., Bordbari, M. H., Durbin-Johnson, B., Settles, M., Perez-Flores, M. C., Lee, J. H., & Yamoah, E. N. (2019). Single-Cell RNA-seq Reveals Profound Alterations in Mechanosensitive Dorsal Root Ganglion Neurons with Vitamin E Deficiency. *iScience*, *21*, 720–735. https://doi.org/10.1016/j.isci.2019.10.064
6. Finno, C. J., Bordbari, M. H., Gianino, G., Ming-Whitfield, B., Burns, E., Merkel, J., Britton, M., Durbin-Johnson, B., Sloma, E. A., McMackin, M., Cortopassi, G., Rivas, V., Barro, M., Tran, C. K., Gennity, I., Habib, H., Xu, L., Puschner, B., & Miller, A. D. (2018). An innate immune response and altered

nuclear receptor activation defines the spinal cord transcriptome during alpha-tocopherol deficiency in Ttpa-null mice. *Free radical biology & medicine*, *120*, 289–302. https://doi.org/10.1016/j.freeradbiomed.2018.02.037

7. Finno, C. J., Bordbari, M. H., Valberg, S. J., Lee, D., Herron, J., Hines, K., Monsour, T., Scott, E., Bannasch, D. L., Mickelson, J., & Xu, L. (2016). Transcriptome profiling of equine vitamin E deficient neuroaxonal dystrophy identifies upregulation of liver X receptor target genes. *Free radical biology & medicine*, *101*, 261–271. https://doi.org/10.1016/j.freeradbiomed.2016.10.009
8. Kohlschütter, A., Finckh, B., Nickel, M., Bley, A., & Hübner, C. (2020). First Recognized Patient with Genetic Vitamin E Deficiency Stable after 36 Years of Controlled Supplement Therapy. *Neuro-degenerative diseases*, *20*(1), 35–38. https://doi.org/10.1159/000508080
9. Edwards, L. A., Donnelly, C. G., Reed, S. M., Valberg, S., Chigerwe, M., Johnson, A. L., & Finno, C. J. (2021). Serum and cerebrospinal fluid phosphorylated neurofilament heavy protein concentrations in equine neurodegenerative diseases. *Equine veterinary journal*, 10.1111/evj.13452. Advance online publication. https://doi.org/10.1111/evj.13452
10. Hales, E. N., Habib, H., Favro, G., Katzman, S., Sakai, R. R., Marquardt, S., Bordbari, M. H., Ming-Whitfield, B., Peterson, J., Dahlgren, A. R., Rivas, V., Ramirez, C. A., Peng, S., Donnelly, C. G., Dizmang, B. S., Kallenberg, A., Grahn, R., Miller, A. D., Woolard, K., Moeller, B., … Finno, C. J. (2021). Increased α-tocopherol metabolism in horses with equine neuroaxonal dystrophy. *Journal of veterinary internal medicine*, *35*(5), 2473–2485. https://doi.org/10.1111/jvim.16233

11. Motobayashi, M., Fukuyama, T., Okuno-Yuguchi, J., Tsukahara, K., Nagaharu, S., Hagimoto, R., Kinoshita, T., Nakazawa, Y., & Inaba, Y. (2017). Subclinical Neuroaxonal Damage in Patients With Clinically Mild Encephalitis/Encephalopathy With a Reversible Splenial Lesion. *Pediatric neurology*, *74*, e3–e4. https://doi.org/10.1016/j.pediatrneurol.2017.05.026
12. Matsushige, T., Ichiyama, T., Kajimoto, M., Okuda, M., Fukunaga, S., & Furukawa, S. (2009). Serial cerebrospinal fluid neurofilament concentrations in bacterial meningitis. *Journal of the neurological sciences*, *280*(1-2), 59–61. https://doi.org/10.1016/j.jns.2009.01.021
13. Douglas-Escobar, M., Yang, C., Bennett, J., Shuster, J., Theriaque, D., Leibovici, A., Kays, D., Zheng, T., Rossignol, C., Shaw, G., & Weiss, M. D. (2010). A pilot study of novel biomarkers in neonates with hypoxic-ischemic encephalopathy. *Pediatric research*, *68*(6), 531–536. https://doi.org/10.1203/PDR.0b013e3181f85a03
14. Lee, T., Chikkabyrappa, S. M., Reformina, D., Mastrippolito, A., Chakravarti, S. B., Mosca, R. S., Shaw, G., & Malhotra, S. P. (2018). Ubiquitin C-Terminal Hydrolase 1 and Phosphorylated Axonal Neurofilament Heavy Chain in Infants Undergoing Cardiac Surgery: Preliminary Assessment as Potential Biomarkers of Brain Injury. *World journal for pediatric & congenital heart surgery*, *9*(4), 412–418. https://doi.org/10.1177/2150135118762390
15. McPhillips, L., Kholwadwala, D., Sison, C. P., Gruber, D., & Ojamaa, K. (2019). A Novel Brain Injury Biomarker Correlates with Cyanosis in Infants with Congenital Heart Disease. *Pediatric cardiology*, *40*(3), 546–553. https://doi.org/10.1007/s00246-018-2023-4

16. Goeral, K., Hauck, A., Atkinson, A., Wagner, M. B., Pimpel, B., Fuiko, R., Klebermass-Schrehof, K., Leppert, D., Kuhle, J., Berger, A., Olischar, M., & Wellmann, S. (2021). Early life serum neurofilament dynamics predict neurodevelopmental outcome of preterm infants. *Journal of neurology*, *268*(7), 2570–2577. https://doi.org/10.1007/s00415-021-10429-5
17. Khalil, M., Teunissen, C. E., Otto, M., Piehl, F., Sormani, M. P., Gattringer, T., Barro, C., Kappos, L., Comabella, M., Fazekas, F., Petzold, A., Blennow, K., Zetterberg, H., & Kuhle, J. (2018). Neurofilaments as biomarkers in neurological disorders. *Nature reviews. Neurology*, *14*(10), 577–589. https://doi.org/10.1038/s41582-018-0058-z
18. Traber M. G. (2021). Vitamin E: necessary nutrient for neural development and cognitive function. *The Proceedings of the Nutrition Society*, *80*(3), 319–326. https://doi.org/10.1017/S0029665121000914
19. Péter, S., Friedel, A., Roos, F. F., Wyss, A., Eggersdorfer, M., Hoffmann, K., & Weber, P. (2015). A Systematic Review of Global Alpha-Tocopherol Status as Assessed by Nutritional Intake Levels and Blood Serum Concentrations. *International journal for vitamin and nutrition research. Internationale Zeitschrift fur Vitamin- und Ernahrungsforschung. Journal international de vitaminologie et de nutrition*, *85*(5-6), 261–281. https://doi.org/10.1024/0300-9831/a000281
20. Finno, C. J., Estell, K. E., Katzman, S., Winfield, L., Rendahl, A., Textor, J., Bannasch, D. L., & Puschner, B. (2015). Blood and Cerebrospinal Fluid α-Tocopherol and Selenium Concentrations in Neonatal Foals with Neuroaxonal Dystrophy. *Journal of veterinary internal medicine*, *29*(6), 1667–1675. https://doi.org/10.1111/jvim.13618

21. National Research Council (U.S.). (2007). *Nutrient requirements of horses*. Washington, D.C: National Academies Press.

22. Aleman, M., Finno, C. J., Higgins, R. J., Puschner, B., Gericota, B., Gohil, K., LeCouteur, R. A., & Madigan, J. E. (2011). Evaluation of epidemiological, clinical, and pathological features of neuroaxonal dystrophy in Quarter Horses. *Journal of the American Veterinary Medical Association*, *239*(6), 823–833. https://doi.org/10.2460/javma.239.6.823

23. Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011, May). The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 143-146).

24. Morales Gómez, A. M., Zhu, S., Palmer, S., Olsen, E., Ness, S. L., Divers, T. J., Bischoff, K., & Mohammed, H. O. (2019). Analysis of neurofilament concentration in healthy adult horses and utility in the diagnosis of equine protozoal myeloencephalitis and equine motor neuron disease. *Research in veterinary science*, *125*, 1–6. https://doi.org/10.1016/j.rvsc.2019.04.018

25. Altmann, P., Leutmezer, F., Zach, H., Wurm, R., Stattmann, M., Ponleitner, M., Petzold, A., Zetterberg, H., Berger, T., Rommer, P., & Bsteh, G. (2020). Serum neurofilament light chain withstands delayed freezing and repeated thawing. *Scientific reports*, *10*(1), 19982. https://doi.org/10.1038/s41598-020-77098-8

26. Barton, M. H., Morris, D. D., Crowe, N., Collatos, C., & Prasse, K. W. (1995). Hemostatic indices in healthy foals from birth to one month of age. *Journal of veterinary diagnostic investigation : official publication of the American*

Association of Veterinary Laboratory Diagnosticians, Inc, *7*(3), 380–385. https://doi.org/10.1177/104063879500700314

27. Tate, J., & Ward, G. (2004). Interferences in immunoassay. *The Clinical biochemist. Reviews*, *25*(2), 105–120.

Chapter 4: Cerebrospinal fluid (CSF) and serum proteomic profiles accurately distinguish equine neuroaxonal dystrophy from cervical vertebral compressive myelopathy

Callum G. Donnelly[1], Amy Johnson[2], Steve M. Reed[3] and Carrie J. Finno[1]

[1]Department of Population Health and Reproduction, School of Veterinary Medicine, University of California, Davis, USA

[2]Department of Clinical Sciences, New Bolton Center, School of Veterinary Medicine, University of Pennsylvania, Kennet Sqaure, PA, USA

[3]Rood and Riddle Equine Hospital, Lexington, KY, USA

Keywords: Neurodegeneration, biomarker, machine learning, precision medicine

Abstract

Background

Cervical vertebral compressive myelopathy (CVCM) and equine neuroaxonal dystrophy/degenerative myeloencephalopathy (eNAD/EDM) are leading causes of spinal ataxia in horses. The conditions can be difficult to differentiate, and there is currently no diagnostic modality that offers a definitive antemortem diagnosis.

Objective

To evaluate a novel proteomic technique to protein biomarker discovery. Additionally, to use machine learning algorithms to predict biomarkers that can aid in the antemortem diagnosis of non-infectious spinal ataxia.

Study Design

Case-control study using banked samples.

Methods

Serum and cerebrospinal fluid (CSF) samples from necropsy confirmed adult eNAD/EDM (n=47) and CVCM horses (n=25) formed the two case populations. Neurologically normal adult horses (n=45) formed the control group. Serum and CSF were assayed for 368 neurologically relevant proteins using Olink technology. Data were analyzed using machine learning algorithms to define diagnostic biomarkers.

Results

Of the 368 proteins, 84 were detected in CSF and 146 in serum. In CSF, 18/84 proteins and 30/146 in serum were differentially abundant between the three groups, following correction for multiple testing. Modelling demonstrated that a two-protein test using CSF had the highest accuracy for discriminating between all three groups. CSF R-spondin 1 (RSPO1) and neurofilament-light (NEFL), in parallel, predicted normal horses with an accuracy of 87.2%, CVCM with 84.6% and eNAD/EDM with 73.5%.

Main Limitations

Cross-species platform. Uneven sample size.

Conclusions

Olink technology is a viable way to rapidly screen equine derived sample matrices for potential protein biomarkers. Machine learning approaches analysis allows for the unbiased selection of highly accurate biomarkers from high-dimensional data.

Introduction

Spinal ataxia is a frequently encountered neurodegenerative condition in racing and sport horses. Ataxia, in many cases, may be a career-ending diagnosis. Spinal ataxia affects horses at all levels, with the 1977 Triple Crown winner Seattle Slew famously being affected during his breeding career.[1] Spinal ataxias have several underlying causes, including congenital and acquired cervical vertebral compressive myelopathy (CVCM), equine neuroaxonal dystrophy (eNAD)/equine degenerative myeloencephalopathy (EDM), and equine protozoal myeloencephalitis (EPM). Apart from immunologic diagnosis of EPM, definitively diagnosing the underlying cause of spinal ataxia antemortem is challenging.[2] The physical limitations of the horse preclude sensitive imaging techniques such as magnetic resonance imaging (MRI), with the mainstay of diagnosis relying on radiography and/or myelography and clinical exclusion. These imaging techniques are primarily to exclude CVCM, however, the poor sensitivity and presence of non-significant background lesions limit their utility.[3] Additionally, there is no definitive diagnostic test for the second most common etiology of spinal ataxia, eNAD/EDM).[4,5] As such, the definitive diagnosis in many cases of spinal ataxia can only be achieved through postmortem examination. The inability to accurately diagnose spinal ataxia antemortem is a significant financial burden for horse owners, trainers, and insurance underwriters.

Next-generation proteomic technologies, coupled with detailed phenotyping, provide the opportunity to enhance diagnostic modalities. Spinal ataxia in horses is well positioned to take advantage of these technologies to improve antemortem diagnosis. The application of such techniques to degenerative neurologic disease in humans has already yielded several highly sensitive and specific biomarkers.[6] Of

particular interest is the use of discovery proteomics to rapidly advance the availability of diagnostic and prognostic markers such as neurofilament light chain (NEFL). NEFL is a sensitive and specific marker of neurodegeneration in humans, with serum and cerebrospinal fluid (CSF) concentrations elevated years before the onset of symptoms.[7] The potential to apply these same discovery approaches to equine spinal ataxia are now possible. Therefore, the aim of this study was to define the serum and CSF proteome of horses with spinal ataxia caused by CVCM and eNAD/EDM for a targeted set of proteins curated for neuropathologic conditions. Our primary hypothesis was that horses with spinal ataxia will have a unique proteomic profile in serum and CSF that distinguish them from neurologically normal horses. We additionally hypothesized that CVCM and eNAD/EDM will have distinct proteomic profiles that differentiate each cause of spinal ataxia.

Methods

Pilot Study

Samples of CSF and serum from age- and sex-matched American Quarter Horses were used to validate the use Olink proximity extension assay technology (PEA). Samples were derived from necropsy confirmed eNAD (*n=5*) and neurologically normal horses (*n=5*). Samples were evaluated using the Olink Target 92 Neuro-Exploratory assay. This assay simultaneously evaluates the relative concentration of 92 curated proteins in serum and CSF without the need for additional sample preparation. All assays were performed at Olink (Boston, MA, USA), with samples run on a single plate. Internal plate standardization and quality control were performed by Olink.

Study Cohort

Previously collected paired serum and CSF samples from neurologically normal horses (age range 3-21 y; mean 7.7 ± 5.89 y; *n=45*) formed the control cohort. Paired serum and CSF samples from CVCM (*n=25*) and eNAD/EDM (*n=47*) that were all examined for spinal ataxia and later confirmed by necropsy formed the case populations. All samples were collected as previously describe by Edwards et al (2021) and stored at -80^{o}C until analysis.[8] Samples used for this study had not previously undergone a freeze thaw cycle until the time of analysis.

Sample size was calculated according to the National Cancer Institute's guidelines for classifier development for high-dimensional data.[9,10] Using a standard fold change of 2, with a protein array of 368 analytes and a training set sample tolerance=0.1, at least 19 samples per group would be required to accurately assign groups based on the array. Standard prevalence parameters were used.

Proximity Extension Assay (PEA) Analysis

Study samples quantified using Olink multiplex proximity extension assay (PEA) panels (Olink Proteomics; www.olink.com) according to the manufacturer's instructions and as described by Assarsson et al. (2014).[11] The basis of PEA is a dual-recognition immunoassay, where two matched antibodies labelled with unique DNA oligonucleotides simultaneously bind to a target protein in solution. This brings the two antibodies into proximity, allowing their DNA oligonucleotides to hybridize, serving as template for a DNA polymerase-dependent extension step. This double-stranded DNA, which is unique for specific antigens, will be amplified using P5/P7 Illumina adapters along with sample indexing, which is quantitatively proportional to

the initial concentration of target protein. These amplified targets will finally get quantified by Next Generation Sequencing using Illumina Nova Seq 6000 (Illumina Corporation. San Diego, California). In this study, we have used the Explore 384 Neurology panel, which measures 368 proteins using 1 µl of serum and CSF.

Data Analysis

Relative quantification of protein expression was calculated by read counts normalized to the internal plate controls, as described previously.[11] Read counts are then converted to normalized protein expression units (NPX), providing relative quantification on a $\log_2$ scale.

Proteins were analyzed if they were detectable in $\geq$ 50% of samples. Analysis was performed using the OlinkAnalyze package in R (R project, San Diego, USA). Data was first inspected by principal component analysis (PCA). Analysis of covariance (ANCOVA) was performed for serum and CSF samples, with phenotype (eNAD/EDM, CVCM and normal) considered a fixed effect and age, breed, and sex introduced as covariates. A Bonferroni correction was used to correct for multiple comparisons. A corrected p-value of <0.05 was considered significant.

Proteins that were significantly differentially abundant in serum and CSF were used to train random forest models. Random forest analysis was performed in R using the randomForest package with default parameters. A randomly selected sub-set of cases and controls (80%) were used to train the model. The model was subsequently tested on the remaining cases and controls (20%). Models were constructed for serum alone, CSF alone and a merged serum and CSF dataset.

Results are reported as class error (proportion of incorrect decisions), mean decrease in accuracy and mean decrease Gini index.

To determine the minimum biomarker seta needed to achieve accurate group assignment, conditional inference models were constructed in R using the Partykit package. Default parameters were used for the three classifiers, with the model having the same set of proteins introduced as the random forest model for serum, CSF and merged data. A p-value of <0.05 was used to define binary classification with accuracy and precision of prediction for each group reported.

Results

Pilot Study

All CSF samples and 9/10 serum samples fully passed quality control, with an intra-assay coefficient of variation of 3% and 8%, respectively. In serum, 36/92 proteins were detected in ≥ 50% samples and in CSF 15/92 proteins were detected in ≥ 50% of samples (**Supplementary Table 4 S1***)*. Due to the small sample size, differential abundance analysis was not performed. Our preliminary data indicate that the Olink platform performs favorably with equine derived samples.

Table 4 S1. Proteins identified in ≥ 50% of samples analyzed using the Target 96 Neuro-Exploratory panel for both serum and CSF with eNAD/EDM (*n=5)* and age/sex matched controls (*n=5)* to pilot the technology using equine derived samples. *denotes proteins that were detected in the subsequent study using Olink Explorer 387.

Assay	OlinkID	UniProt	Sample Matrix
CRADD	OID05129	P78560	Serum Only
CETN2	OID05133	P41208	Serum Only
SMOC1	OID05135	Q9H4F8	Serum and CSF
ADGRB3	OID05521	O60242	Serum and CSF
KLB	OID05137	Q86Z14	Serum Only
CDH17	OID05138	Q12864	Serum Only
GPNMB	OID05139	Q14956	Serum Only

ATP6V1F	OID05143	Q16864	Serum Only
ANXA10	OID05147	Q9UJ72	Serum Only
RPS6KB1	OID05150	P23443	Serum Only
CRIP2	OID05154	P52943	Serum and CSF
ADAM15	OID05162	Q13444	Serum and CSF
FGFR2	OID05166	P21802	Serum and CSF
NAA10	OID05168	P41227	Serum Only
SFRP1	OID05171	Q8N474	Serum and CSF*
PRTFDC1	OID05176	Q9NRG1	Serum Only
TBCB	OID05181	Q99426	Serum Only*
NPM1	OID05182	P06748	Serum and CSF
ASGR1	OID05183	P07306	Serum and CSF
COL4A3BP	OID05185	Q9Y5P4	Serum Only
PSG1	OID05186	P11464	Serum Only
PSME1	OID05187	Q06323	Serum Only*
PTS	OID05193	Q03393	Serum Only
FUT8	OID05195	Q9BYC5	Serum and CSF
TPPP3	OID05196	Q9BW30	Serum Only
PFDN2	OID05197	Q9UHV9	Serum Only
AARSD1	OID05202	Q9BTE6	Serum Only
PHOSPHO1	OID05203	Q8TCT1	Serum Only
NEFL	OID05206	P07196	Serum and CSF*
HMOX2	OID05207	P30519	Serum Only
KIF1BP	OID05213	Q96EK5	Serum Only
PPP3R1	OID05214	OID05214	Serum and CSF*
ILKAP	OID05215	Q9H0C8	Serum Only
PMVK	OID05218	Q15126	Serum Only*
WWP2	OID05219	O00308	Serum and CSF*
GGT5	OID05221	P36269	Serum Only*
CLSTN1	OID05158	O94985	CSF Only
CD302	OID05152	Q8IX05	CSF Only
NXPH1	OID05167	P58417	CSF Only*
MAD1L1	OID05191	Q9Y6D9	CSF Only
ISLR2	OID05216	Q6UXK2	CSF Only

Serum Proteins

Quality control parameters were met for 86% of individual assays, with an intra-assay coefficient of variation (COV) of 12%. In >50% of samples, 148 of 368 proteins (40%) were detected. Of those proteins detected in >50% samples, 37 (25%) were differentially abundant amongst the three groups (**Supplementary Table 4 S2***)*.

Despite the relatively large number of differentially abundant proteins identified, the principal component analysis (PCA) for serum did not demonstrate clear clustering by group (**Supplementary Fig. 4.S1**).

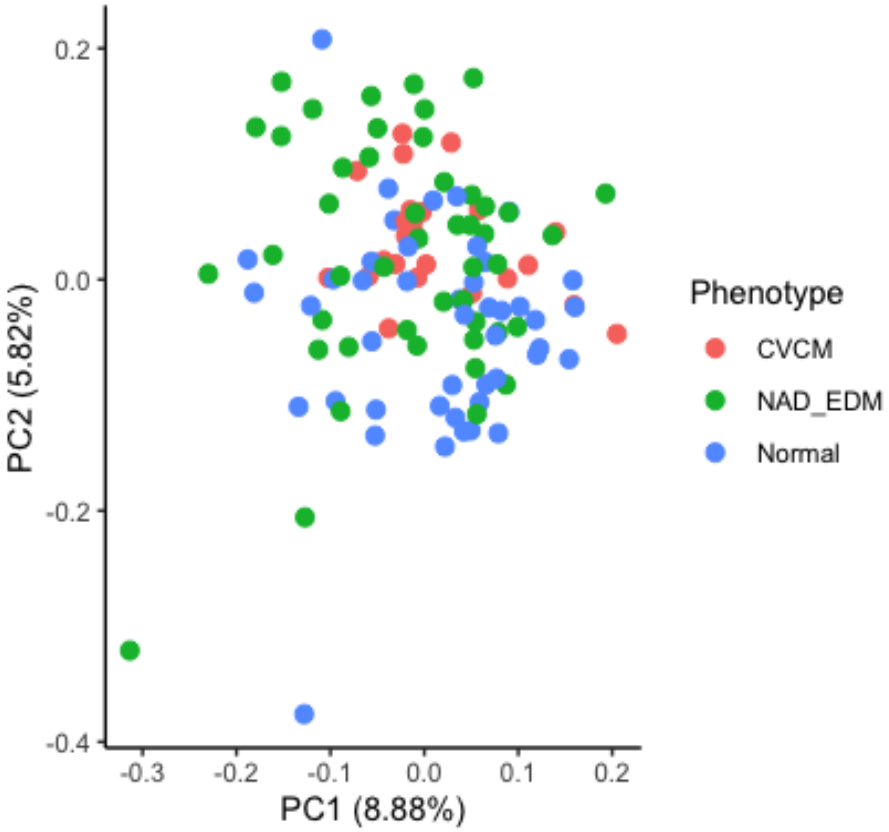

Figure 4 S1 Principal component plot serum proteins showing the first two components (PC1 and PC2). Using proteins with a missingness ≤50% values were converted to eigenvectors and plotted. Each dot represents an individual horse from the CVCM group (n=25; salmon), eNAD/EDM (n=47; blue) or normal (n=45; green).

Table 4 S2. Post-hoc ANCOVA analysis of serum proteins present in ≥ 50% of samples using the Olink Explorer 387 platform that were significantly differently abundant following correction for multiple comparisons. Proteins are listed in order of significance. A Bonferroni adjusted P-value <0.05 was considered significant.

Assay	OlinkID	UniProt	contrast	estimate	conf.low	conf.high	Adjusted_pval
CD164	OID21080	Q04900	NAD_EDM - Normal	0.325567	0.161252	0.489883	2.21E-05
NEFL	OID20871	P07196	CVCM - Normal	1.130536	0.507985	1.753088	1.04E-04
RHOC	OID20950	P08134	CVCM - NAD_EDM	-0.87318	-1.43319	-0.31317	9.69E-04
TARBP2	OID20870	Q15633	CVCM - NAD_EDM	-0.78036	-1.28217	-0.27854	1.00E-03
BIN2	OID21067	Q9UBW5	NAD_EDM - Normal	0.274369	0.094348	0.454391	1.29E-03
THY1	OID21050	P04216	NAD_EDM - Normal	0.285633	0.096942	0.474323	1.40E-03
TBC1D17	OID20844	Q9HA65	CVCM - NAD_EDM	-0.7769	-1.29659	-0.25722	1.63E-03
DKK1	OID21066	O94907	CVCM - NAD_EDM	0.51436	0.168298	0.860421	1.75E-03
CALCA	OID20983	P01258	CVCM - Normal	-0.88985	-1.51492	-0.26479	2.85E-03
PBLD	OID20839	P30039	CVCM - Normal	-0.53554	-0.91672	-0.15435	3.29E-03
DNMBP	OID20956	Q6XZF7	CVCM - NAD_EDM	-0.92552	-1.58511	-0.26593	0.003333
MESD	OID21099	Q14696	CVCM - NAD_EDM	-0.26999	-0.46625	-0.07373	0.004102
TBCC	OID20942	Q15814	CVCM - Normal	-0.56146	-0.97206	-0.15086	0.004363
MITD1	OID20959	Q8WV92	CVCM - NAD_EDM	-0.77114	-1.34233	-0.19996	0.004962
GGT5	OID20909	P36269	CVCM - NAD_EDM	0.646726	0.166617	1.126835	0.005074
THY1	OID21050	P04216	CVCM - Normal	0.340517	0.087234	0.5938	0.005173
C19orf12	OID20804	Q9NSK7	NAD_EDM - Normal	-0.43021	-0.75083	-0.10958	0.005274
PPP3R1	OID20902	P63098	CVCM - NAD_EDM	-0.64736	-1.13947	-0.15526	0.006388
TBCB	OID20993	Q99426	NAD_EDM - Normal	-0.49305	-0.87101	-0.11509	0.006915
DKK1	OID21066	O94907	CVCM - Normal	0.466918	0.104408	0.829428	0.007781
SFRP1	OID20984	Q8N474	CVCM - Normal	1.399865	0.312303	2.487426	0.007828
TST	OID20868	Q16762	CVCM - Normal	-0.70081	-1.24563	-0.15599	0.007876
APRT	OID20927	P07741	CVCM - NAD_EDM	-1.24049	-2.2075	-0.27349	0.008073
GGT5	OID20909	P36269	CVCM - Normal	0.631417	0.128488	1.134345	0.009785
LBR	OID21034	Q14739	NAD_EDM - Normal	0.735311	0.147206	1.323416	0.010144
ALDH1A1	OID21128	P00352	CVCM - Normal	-0.34266	-0.6175	-0.06783	0.010394
NEFL	OID20871	P07196	CVCM - NAD_EDM	0.738595	0.144291	1.3329	0.010687
PTEN	OID20794	P60484	CVCM - NAD_EDM	-0.29912	-0.53998	-0.05827	0.010752
TBCC	OID20942	Q15814	NAD_EDM - Normal	-0.37666	-0.68255	-0.07077	0.011563
TST	OID20868	Q16762	NAD_EDM - Normal	-0.49787	-0.90375	-0.09199	0.011942
C19orf12	OID20804	Q9NSK7	CVCM - Normal	-0.52162	-0.952	-0.09124	0.013195
BAX	OID20856	Q07812	CVCM - NAD_EDM	-0.99399	-1.82104	-0.16694	0.014132

CHMP1A	OID20930	Q9HD42	CVCM - Normal	-0.60938	-1.11805	-0.10072	0.014502
WFIKKN1	OID20939	Q96NZ8	CVCM - Normal	0.455095	0.072153	0.838037	0.015463
MESD	OID21099	Q14696	CVCM - Normal	-0.24383	-0.44942	-0.03825	0.01571
DNMBP	OID20956	Q6XZF7	NAD_EDM - Normal	0.603628	0.088889	1.118366	0.017164
SUSD2	OID21098	Q9UGT4	NAD_EDM - Normal	-0.31126	-0.57764	-0.04489	0.017643
NXPH1	OID20849	P58417	NAD_EDM - Normal	-0.57665	-1.07199	-0.08131	0.018154
BST2	OID21029	Q10589	NAD_EDM - Normal	-0.34222	-0.63817	-0.04627	0.019107
JAM2	OID21064	P57087	CVCM - Normal	0.320304	0.032203	0.608405	0.025436
ADAM22	OID21001	Q9P0K1	CVCM - NAD_EDM	-0.58518	-1.11921	-0.05115	0.028121
APRT	OID20927	P07741	NAD_EDM - Normal	0.818964	0.064323	1.573605	0.030027
MASP1	OID20954	P48740	CVCM - NAD_EDM	-0.27784	-0.53443	-0.02125	0.03048
DUSP3	OID20827	P51452	CVCM - NAD_EDM	0.743752	0.055135	1.432369	0.031002
GLB1	OID20949	P16278	NAD_EDM - Normal	0.717589	0.052848	1.38233	0.03111
PMVK	OID20850	Q15126	CVCM - NAD_EDM	-0.8291	-1.60331	-0.05489	0.032796
PARK7	OID21160	Q99497	NAD_EDM - Normal	0.412715	0.025238	0.800192	0.033972
TFF1	OID21154	P04155	CVCM - NAD_EDM	-0.13723	-0.26736	-0.00711	0.03621
PECAM1	OID21131	P16284	CVCM - NAD_EDM	-0.18414	-0.35929	-0.00898	0.03695
NOS3	OID20834	P29474	CVCM - Normal	0.522493	0.021814	1.023172	0.038699
PSME2	OID20989	Q9UL46	CVCM - Normal	-0.3537	-0.69277	-0.01463	0.038793
NSFL1C	OID21022	Q9UNZ2	NAD_EDM - Normal	0.716216	0.02814	1.404293	0.039322
SCARB1	OID20805	Q8WTV0	NAD_EDM - Normal	-0.45798	-0.90083	-0.01514	0.040925
FMNL1	OID20884	O95466	CVCM - Normal	-0.73643	-1.45242	-0.02044	0.042316
GP6	OID21091	Q9HCN6	NAD_EDM - Normal	0.216845	0.004951	0.42874	0.043631
PAMR1	OID21153	Q6UXH9	CVCM - Normal	0.241835	0.005208	0.478461	0.04398
ITGAM	OID21071	P11215	NAD_EDM - Normal	0.124289	0.002142	0.246437	0.045152
HARS	OID21086	P12081	CVCM - NAD_EDM	-0.39745	-0.78853	-0.00637	0.045485
CXCL8	OID20997	P10145	NAD_EDM - Normal	0.531976	0.007266	1.056686	0.046137
MITD1	OID20959	Q8WV92	NAD_EDM - Normal	0.450839	0.005093	0.896584	0.046793
SERPINB9	OID20932	P50453	CVCM - Normal	-0.62873	-1.25251	-0.00494	0.047757
SOD2	OID21114	P04179	CVCM - NAD_EDM	-0.43211	-0.86292	-0.0013	0.049133

CSF Proteins

Quality control parameters were met for 89% of individual assays, with an intra-assay-COV of 12%. In >50% of samples, 84 of 368 proteins (23%) were detected. Of those proteins detected in >50% samples, 18 (21%) were differentially abundant amongst the three groups (**Supplementary Table 4 S3**). PCA analysis indicated a clear distinction between samples derived from animals with spinal ataxia compared with neurologically normal horses (**Fig. 4.1**). However, there was a high degree of overlap between the two groups of horses with neurodegenerative disease.

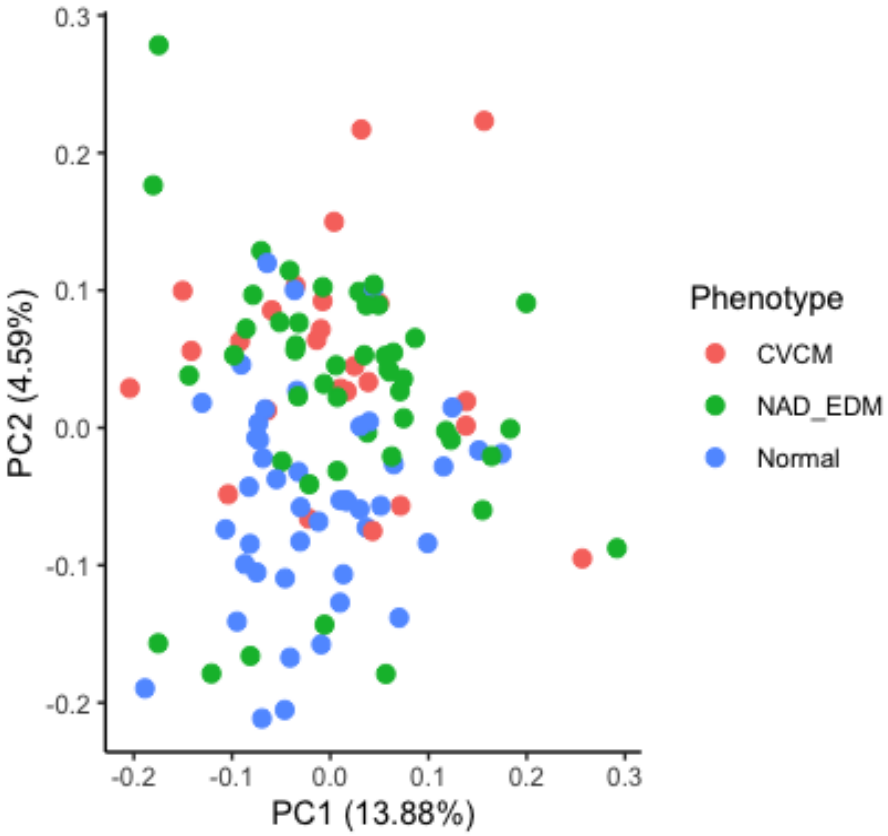

Figure 4.1 Principal component plot of CSF proteins showing the first two components (PC1 and PC2). Using proteins with a missingness ≤ 50%, values were converted to eigenvectors and plotted. Each dot represents an individual horse from the CVCM group (n=25; salmon), eNAD/EDM (n=45; blue) or normal (n=43; green).

Table 4 S3. Post-hoc ANCOVA analysis of serum proteins present in ≥ 50% of samples using the Olink Explorer 387 platform that were significantly differently abundant following correction for multiple comparisons. Proteins are listed in order of significance. A Bonferroni adjusted P-value <0.05 was considered significant.

Assay	OlinkID	UniProt	contrast	estimate	conf.low	conf.high	Adjusted_pval
CD164	OID21080	Q04900	NAD_EDM - Normal	0.325567	0.161252	0.489883	2.21E-05
NEFL	OID20871	P07196	CVCM - Normal	1.130536	0.507985	1.753088	1.04E-04
RHOC	OID20950	P08134	CVCM - NAD_EDM	-0.87318	-1.43319	-0.31317	9.69E-04
TARBP2	OID20870	Q15633	CVCM - NAD_EDM	-0.78036	-1.28217	-0.27854	1.00E-03
BIN2	OID21067	Q9UBW5	NAD_EDM - Normal	0.274369	0.094348	0.454391	1.29E-03
THY1	OID21050	P04216	NAD_EDM - Normal	0.285633	0.096942	0.474323	1.40E-03
TBC1D17	OID20844	Q9HA65	CVCM - NAD_EDM	-0.7769	-1.29659	-0.25722	1.63E-03
DKK1	OID21066	O94907	CVCM - NAD_EDM	0.51436	0.168298	0.860421	1.75E-03
CALCA	OID20983	P01258	CVCM - Normal	-0.88985	-1.51492	-0.26479	2.85E-03
PBLD	OID20839	P30039	CVCM - Normal	-0.53554	-0.91672	-0.15435	3.29E-03
DNMBP	OID20956	Q6XZF7	CVCM - NAD_EDM	-0.92552	-1.58511	-0.26593	0.003333
MESD	OID21099	Q14696	CVCM - NAD_EDM	-0.26999	-0.46625	-0.07373	0.004102
TBCC	OID20942	Q15814	CVCM - Normal	-0.56146	-0.97206	-0.15086	0.004363
MITD1	OID20959	Q8WV92	CVCM - NAD_EDM	-0.77114	-1.34233	-0.19996	0.004962
GGT5	OID20909	P36269	CVCM - NAD_EDM	0.646726	0.166617	1.126835	0.005074
THY1	OID21050	P04216	CVCM - Normal	0.340517	0.087234	0.5938	0.005173
C19orf12	OID20804	Q9NSK7	NAD_EDM - Normal	-0.43021	-0.75083	-0.10958	0.005274
PPP3R1	OID20902	P63098	CVCM - NAD_EDM	-0.64736	-1.13947	-0.15526	0.006388
TBCB	OID20993	Q99426	NAD_EDM - Normal	-0.49305	-0.87101	-0.11509	0.006915
DKK1	OID21066	O94907	CVCM - Normal	0.466918	0.104408	0.829428	0.007781
SFRP1	OID20984	Q8N474	CVCM - Normal	1.399865	0.312303	2.487426	0.007828
TST	OID20868	Q16762	CVCM - Normal	-0.70081	-1.24563	-0.15599	0.007876
APRT	OID20927	P07741	CVCM - NAD_EDM	-1.24049	-2.2075	-0.27349	0.008073
GGT5	OID20909	P36269	CVCM - Normal	0.631417	0.128488	1.134345	0.009785
LBR	OID21034	Q14739	NAD_EDM - Normal	0.735311	0.147206	1.323416	0.010144
ALDH1A1	OID21128	P00352	CVCM - Normal	-0.34266	-0.6175	-0.06783	0.010394
NEFL	OID20871	P07196	CVCM - NAD_EDM	0.738595	0.144291	1.3329	0.010687

PTEN	OID20794	P60484	CVCM - NAD_EDM	-0.29912	-0.53998	-0.05827	0.010752
TBCC	OID20942	Q15814	NAD_EDM - Normal	-0.37666	-0.68255	-0.07077	0.011563
TST	OID20868	Q16762	NAD_EDM - Normal	-0.49787	-0.90375	-0.09199	0.011942
C19orf12	OID20804	Q9NSK7	CVCM - Normal	-0.52162	-0.952	-0.09124	0.013195
BAX	OID20856	Q07812	CVCM - NAD_EDM	-0.99399	-1.82104	-0.16694	0.014132
CHMP1A	OID20930	Q9HD42	CVCM - Normal	-0.60938	-1.11805	-0.10072	0.014502
WFIKKN1	OID20939	Q96NZ8	CVCM - Normal	0.455095	0.072153	0.838037	0.015463
MESD	OID21099	Q14696	CVCM - Normal	-0.24383	-0.44942	-0.03825	0.01571
DNMBP	OID20956	Q6XZF7	NAD_EDM - Normal	0.603628	0.088889	1.118366	0.017164
SUSD2	OID21098	Q9UGT4	NAD_EDM - Normal	-0.31126	-0.57764	-0.04489	0.017643
NXPH1	OID20849	P58417	NAD_EDM - Normal	-0.57665	-1.07199	-0.08131	0.018154
BST2	OID21029	Q10589	NAD_EDM - Normal	-0.34222	-0.63817	-0.04627	0.019107
JAM2	OID21064	P57087	CVCM - Normal	0.320304	0.032203	0.608405	0.025436
ADAM22	OID21001	Q9P0K1	CVCM - NAD_EDM	-0.58518	-1.11921	-0.05115	0.028121
APRT	OID20927	P07741	NAD_EDM - Normal	0.818964	0.064323	1.573605	0.030027
MASP1	OID20954	P48740	CVCM - NAD_EDM	-0.27784	-0.53443	-0.02125	0.03048
DUSP3	OID20827	P51452	CVCM - NAD_EDM	0.743752	0.055135	1.432369	0.031002
GLB1	OID20949	P16278	NAD_EDM - Normal	0.717589	0.052848	1.38233	0.03111
PMVK	OID20850	Q15126	CVCM - NAD_EDM	-0.8291	-1.60331	-0.05489	0.032796
PARK7	OID21160	Q99497	NAD_EDM - Normal	0.412715	0.025238	0.800192	0.033972
TFF1	OID21154	P04155	CVCM - NAD_EDM	-0.13723	-0.26736	-0.00711	0.03621
PECAM1	OID21131	P16284	CVCM - NAD_EDM	-0.18414	-0.35929	-0.00898	0.03695
NOS3	OID20834	P29474	CVCM - Normal	0.522493	0.021814	1.023172	0.038699
PSME2	OID20989	Q9UL46	CVCM - Normal	-0.3537	-0.69277	-0.01463	0.038793
NSFL1C	OID21022	Q9UNZ2	NAD_EDM - Normal	0.716216	0.02814	1.404293	0.039322
SCARB1	OID20805	Q8WTV0	NAD_EDM - Normal	-0.45798	-0.90083	-0.01514	0.040925
FMNL1	OID20884	O95466	CVCM - Normal	-0.73643	-1.45242	-0.02044	0.042316
GP6	OID21091	Q9HCN6	NAD_EDM - Normal	0.216845	0.004951	0.42874	0.043631
PAMR1	OID21153	Q6UXH9	CVCM - Normal	0.241835	0.005208	0.478461	0.04398
ITGAM	OID21071	P11215	NAD_EDM - Normal	0.124289	0.002142	0.246437	0.045152
HARS	OID21086	P12081	CVCM - NAD_EDM	-0.39745	-0.78853	-0.00637	0.045485

CXCL8	OID20997	P10145	NAD_EDM - Normal	0.531976	0.007266	1.056686	0.046137
MITD1	OID20959	Q8WV92	NAD_EDM - Normal	0.450839	0.005093	0.896584	0.046793
SERPINB9	OID20932	P50453	CVCM - Normal	-0.62873	-1.25251	-0.00494	0.047757
SOD2	OID21114	P04179	CVCM - NAD_EDM	-0.43211	-0.86292	-0.0013	0.049133

Machine Learning Biomarker Selection

Random forest classifier models for serum, CSF and a merged CSF/serum dataset were constructed. Two thousand and one trees were used for each sample matrix. In *(***Fig. 4.2, Supplementary Fig. 4.2***).* Testing sets using CSF and merged datasets had the highest accuracy for correctly classifying animals by group. Both datasets performed best for normal and eNAD/EDM horses, correctly classifying with >80% accuracy for each (**Fig. 4.2, Supplementary Fig. 4.3**). Mean decrease accuracy plots indicated a group of proteins that were driving classification in the random forest model.

all three datasets, error values levelled at approximately 500 trees for the training set

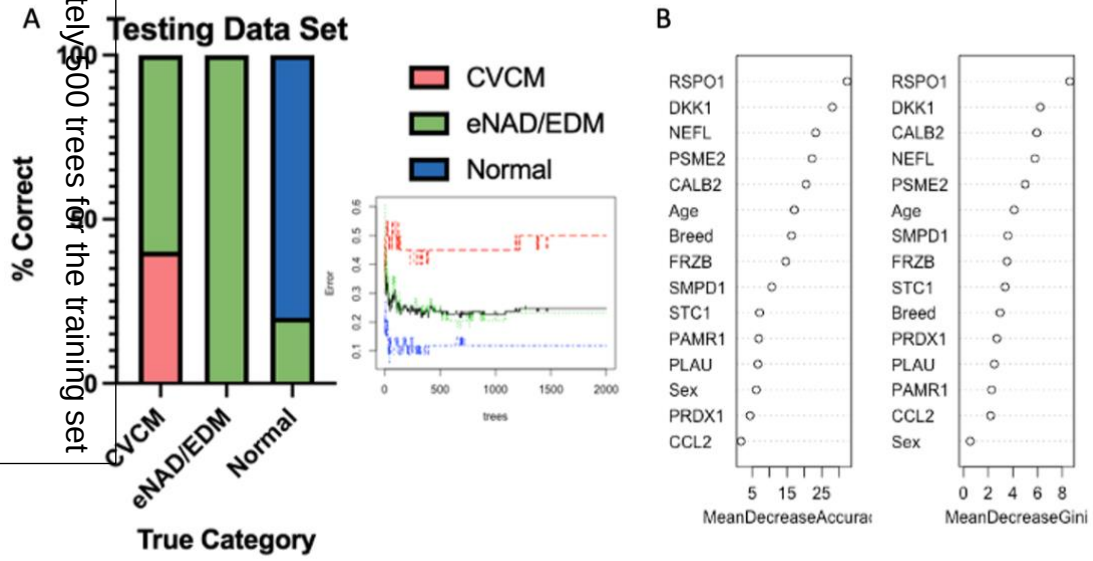

Figure 4.2 Random Forest analysis of CSF. Testing data set (**A**) shows classification of the three groups as percentage of correctly assigned animals. Inset in (**A**) error plot of the training data set for CVCM (red), eNAD/EDM (green), normal (blue) and overall (black). The x-axis indicates the error rate for classification and the y-axis the number of trees used in the classification. Accuracy in the testing set was highest for eNAD/EDM (100%), followed by normal (80%) and CVCM (40%). Importantly, no normal horses were misclassified in the ataxia groups. Mean decrease in accuracy and Gini plots (**B**) ranks the proteins and categorical factors

(age, breed and sex) used in decision by the percentage decrease in accuracy or Gini if that feature were to be removed from the model.

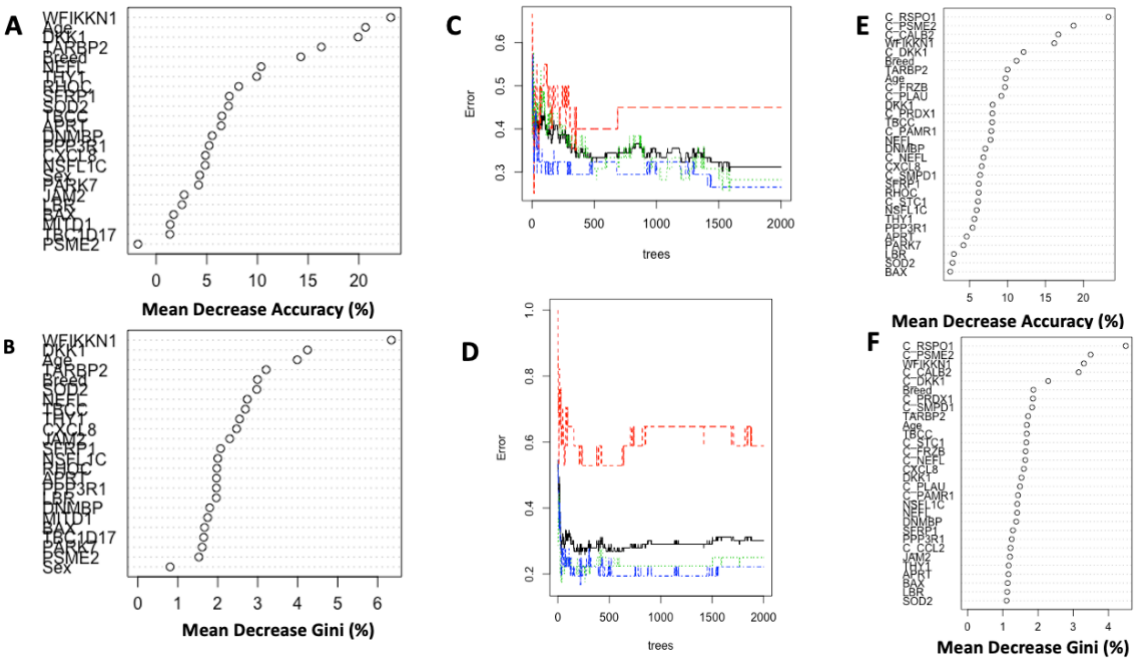

Figure 4 S2. Random Forest Analysis of serum (**A,B,C**) and merged serum/CSF data (**D,E,F**). Error plot (**C,D**) of the training data set for CVCM (red), eNAD/EDM (green), normal (blue) and overall (black). The x-axis indicates the error rate for classification and the y-axis the number of trees used in the classification. Mean decrease in accuracy (**A,E**) and mean decrease Gini ranks (**B,F**) the proteins and categorical factors (age, breed and sex) used in decision by the percentage decrease in accuracy if that feature were to be removed from the model. Proteins with the prefix C_ denotes CSF proteins in the merged data set.

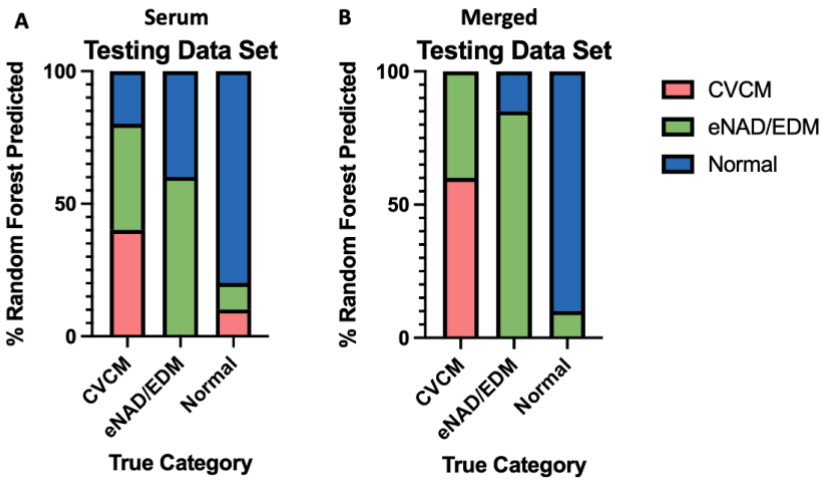

Figure 4 S3 Random Forest model testing. Prediction accuracy determined for serum (**A**) and merged (**B**; serum and CSF) data sets. Merged data performed at a similar accuracy to CSF alone, and again did not misclassify any CVCM cases as eNAD/EDM or normal. Serum performed the least well of all three analyses with misclassification occurring in all groups.

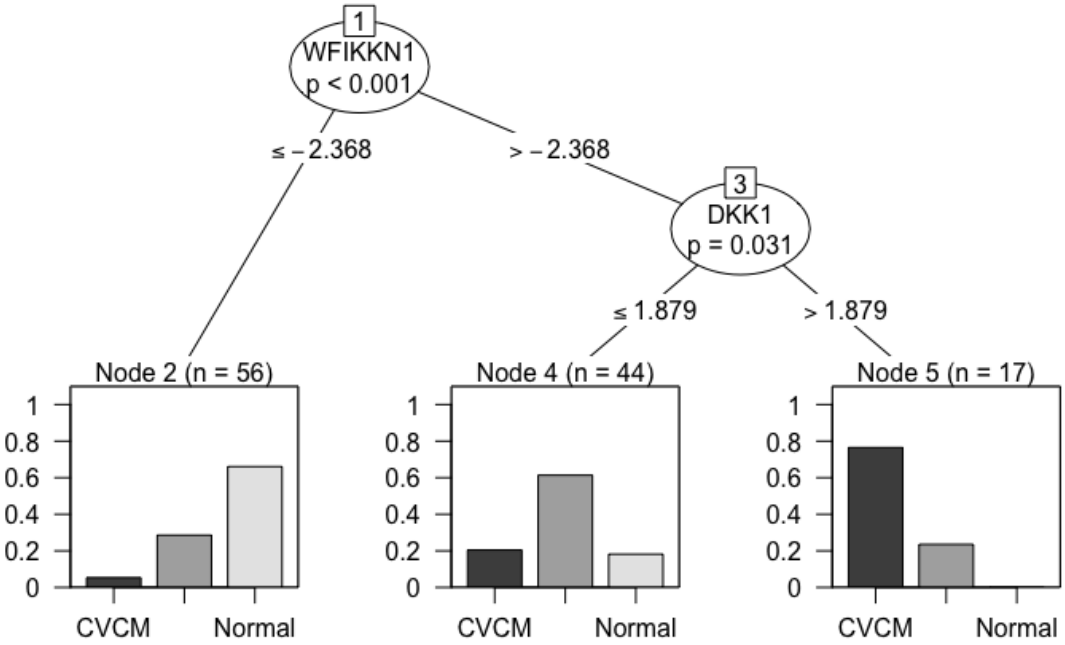

Figure 1 S4 Conditional inference model tree for merged serum/CSF. A three-protein model, with CSF RSPO1, CSF NEFL and serum DKK1 had the highest

accuracy for prediction of normal horses (85.5%), CVCM (85.5%) and eNAD/EDM (72.7%). Abbreviations: DKK1-Dckkoph 1; NEFL-Neurofilament-light; RSPO1-R-Spondin 1.

Conditional inference model selection of biomarkers mirrored the top-ranking proteins as determined by random forest. In serum, two proteins were selected: WAP, follistatin/kazal, immunoglobulin, kunitz and neutrin domain containing 1 (WFKKN1) and Dickkoph-1 (DKK1). In serum, this modelling offered diagnostic accuracies of 67.5% for eNAD/EDM , 87.2% for CVCM and 76.9% for normal horses (**Supplementary Fig. 4S4**). In CSF, a two-protein test (**Fig. 4.*3***) with R-spondin-1 (RSPO1; **Fig. 4.4A**) and neurofilament-light (NEFL; **Fig. 4.4B**) was selected. Profiling these two CSF proteins resulted in a prediction accuracy of 73.5% for eNAD/EDM, 84.6% for CVCM and 87.2% for normal horses. For the merged data set, two CSF proteins (RSPO1 and NEFL) and one serum protein (DKK1; **Fig. 4.4C**) were selected, with prediction accuracies of 72.7% for eNAD/EDM, 84.5% for CVCM and 85.5% for normal horses (**Supplementary Fig. 4S5**).

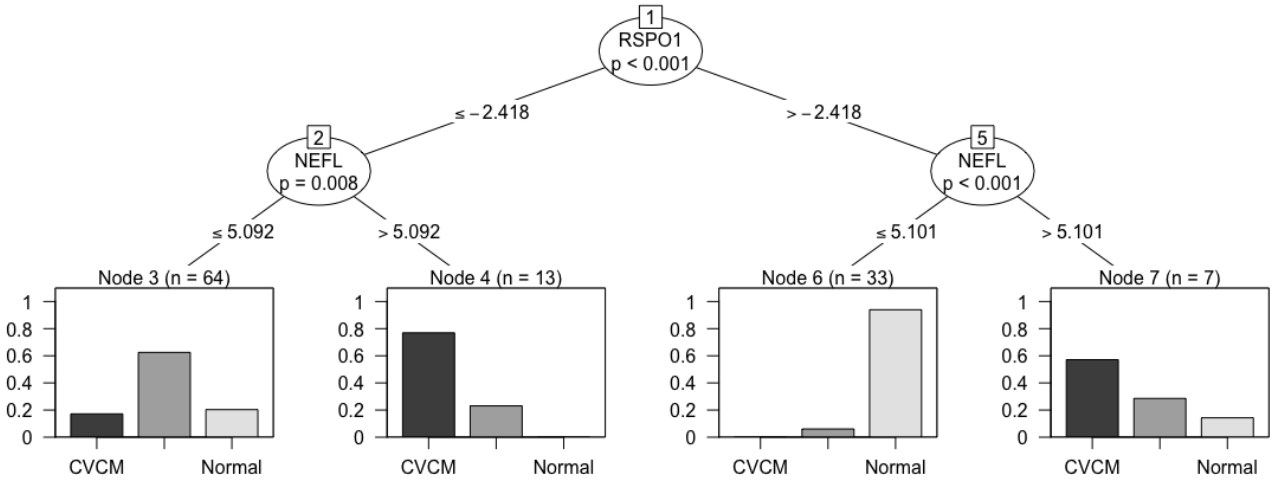

Figure 4.3 Conditional inference model tree for CSF. A two-protein model with RSPO1 and NEFL had the highest accuracy for prediction of normal horses (87.2%), CVCM (84.6%) and eNAD/EDM (73.5%). Abbreviation: NEFL-Neurofilament-light; RSPO1-R-Spondin 1.

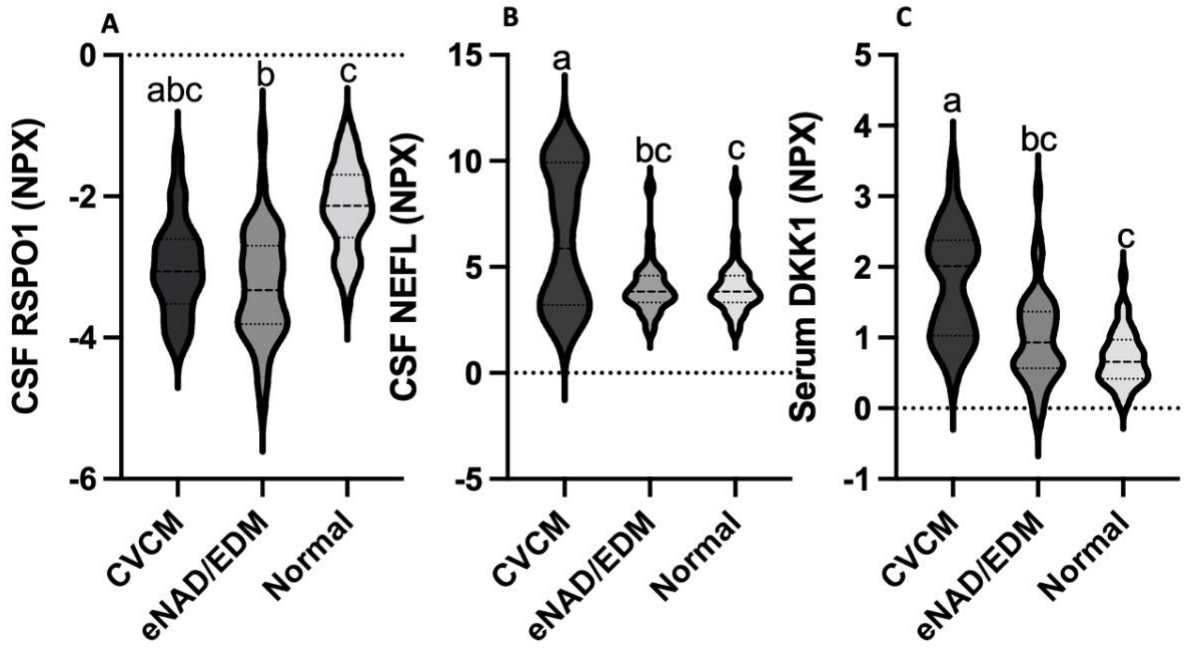

Figure 4.4 Potential markers as defined by random forest and conditional inference models in CSF and merged data sets; CSF RSPO1 (**A**), CSF NEFL (**B**) and serum DKK1 (**C**). Different letters indicate significant difference between groups based on pair-wise post-hoc ANCOVA with correction for multiple testing (P<0.05). Abbreviations: DKK1-Dckkoph 1; NEFL-Neurofilament-light; RSPO1-R-Spondin 1.

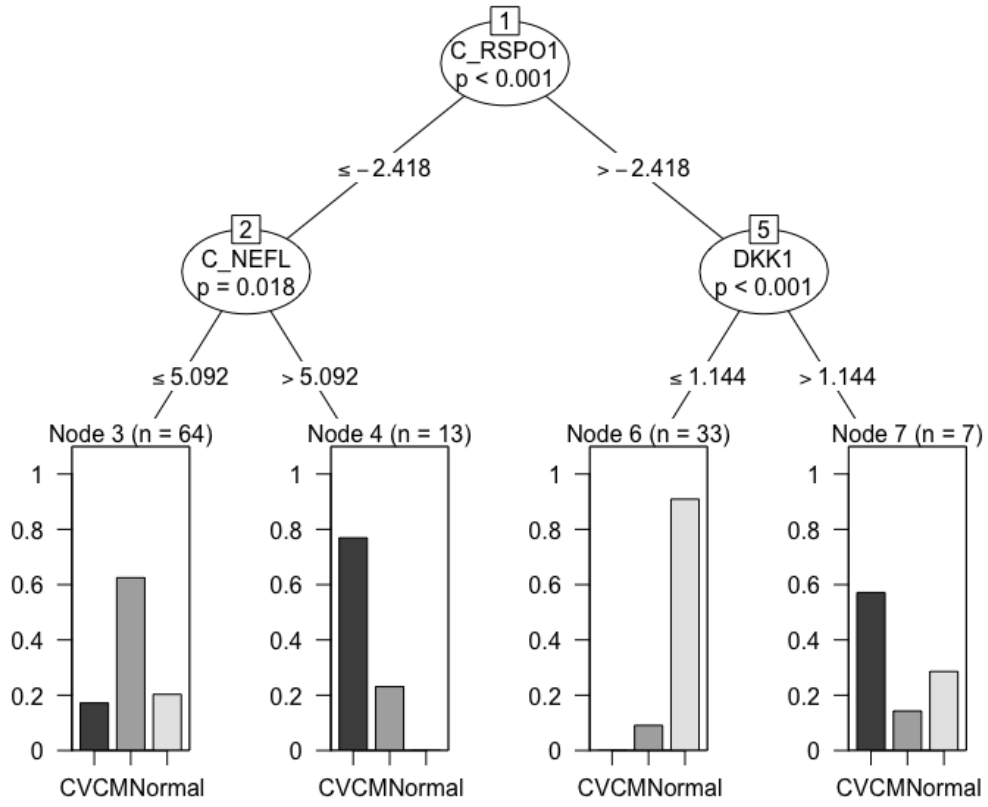

Figure 4 S5. Conditional inference model tree for merged serum/CSF. A three-protein model, with CSF RSPO1, CSF NEFL and serum DKK1 had the highest accuracy for prediction of normal horses (85.5%), CVCM (85.5%) and eNAD/EDM (72.7%). Abbreviations: DKK1-Dckkoph 1; NEFL-Neurofilament-light; RSPO1-R-Spondin 1.

Discussion

For the first time, we demonstrated the effective use of PEA technology using equine serum and CSF sample matrices. Use of this platform allowed for the identification of previously unidentified proteins in equine CSF and serum. Importantly, with the aid of machine learning algorithms, we demonstrated that a limited set of novel CSF proteins could accurately discriminate between not only neurologically normal and abnormal horses, but also by neurodegenerative etiology. These findings add to the catalogue of proteins detectable in equine CSF and serum, while also proposing biomarker candidates that will enhance the diagnostic resolution for equine spinal ataxia.

The equine CSF proteome is a rich, but relatively untapped source of protein biomarkers. Broccardo et al. identified 320 proteins in CSF from six reportedly healthy horses using an 2D-LC-MS/MS approach.[12] None of the proteins identified in the current study were identified in the previous study. While both techniques are highly sensitive, LC-MS/MS is dependent on a high-quality annotation with which to assign protein identity. Given the limitations of the equine proteome annotation, highly abundant proteins are often well characterized, whereas rare proteins may not be identified, even if present.[13] PEA technology is not reliant on protein annotation, nor affected by the presence of highly abundant proteins such as albumin and globulins.[11] As such, it is well suited to detecting and qualitatively defining rare proteins within a sample matrix. The lack of concordance between studies is likely due to the different techniques employed in sample analysis. Additionally, the former study only included healthy horses, and therefore the repertoire of identifiable proteins may have been narrowed by lack of an active pathologic process. While PEA technology allows for identification of proteins with high fidelity, in its current form, it is still a targeted technology and only detects proteins for which the assay has probes. Therefore, the absence of proteins in one study compared to the other is an expected finding. Collectively, both approaches add to the catalogue of known proteins present in equine CSF.

While the PEA assay has probes for 368 proteins, neither CSF nor serum approached the total number of possible detectable proteins. This, in part, may be attributed to the cross-species platform. All probes in this platform were developed for human-derived proteins. Pilot data generated from our laboratory indicated that

proteins with high levels of conservation (>80% protein-protein identity) performed well with this platform. The technology has previously been used for equine synovial samples, where it was accurately able to discriminate between various cyto- and chemokines.[14] Given the dual-antibody design, the risk of false positive detection of proteins is negligible. When used in non-human species, the technology is likely to under-represent the number of detectable proteins in a sample. The current study's findings support this. Additionally, the sample matrix is an important contributor to the available proteins for detection. The discrepancy between CSF and serum was an expected finding, similar to previous work published with human cohorts.[15,16] It is intuitive that that a sample matrix with lower amounts of protein (i.e. CSF) will have a reduced complexity of proteins present.

Machine learning models offer an unbiased and tractable technique to distill high dimensional data into useful diagnostic sets. Random forest and conditional inference, while similar, have some fundamental differences. Namely, random forest takes an ensemble approach, with predictions averaged amongst the sample parameters.[17] Conditional inference selects the feature(s) which best distinguishes between groups in a more purely binary fashion. As such, random forest enables the effect of a feature on prediction to be viewed relative to each other feature (i.e., size of effect of one protein compared to another), whereas conditional inference defines each feature by binary partition. In the current study, random forest helped establish the rank importance of each significant protein in classification of disease phenotype, whereas conditional inference allowed for the discovery of potential diagnostic protein sets. Importantly, the conditional inference sets are informed by the smallest number of proteins to achieve the most accurate classification based on a

statistically significant difference. This has practical significance for future development of these biomarkers as diagnostic tests, with fewer markers aiding in clinical interpretation.[18] Use of the two models resulted in similar classification accuracies for both eNAD/EDM and normal horses, whereas conditional inference was far more accurate for CVCM. This is a result of the smaller sample size of the CVCM group, and the requirement to train the random forest model, further reducing the sample size. By comparison, conditional inference does not introduce a training data set and so the sample size is optimized, and this accounts for the improved performance of the CVCM dataset. Performance of these models suggest that addition of these biomarkers to existing diagnostic modalities would aid in the accurate diagnosis of spinal ataxias resulting from eNAD/EDM or CVCM. While findings from this study are highly encouraging, further work to validate these biomarkers in a replication study is necessary before using these biomarkers in clinical practice.

Neurofilament light (NEFL) was identified as an important protein for group classification in this study, with relative amounts in CSF and serum highest for CVCM. NEFL is part of the neurofilament family of proteins important for the cytoskeletal structure of the axon.[7] Elevations in this protein, both in serum and CSF, have been well characterized in human neurodegenerative diseases including Alzheimer's disease, frontotemporal dementia (FTD) and amyotrophic lateral sclerosis (ALS).[7] Similarly, concentrations in both serum and CSF increase with acute trauma, including traumatic brain injury (TBI) and spinal cord injury (SCI).[19-21] NEFL has not previously been evaluated in horses. Phosphorylated neurofilament heavy (pNfH), a polypeptide from the same family, has been the best characterized

neurofilament in horses with neurologic disease.[8,22-26] Previous work demonstrated elevations of pNfH in CSF of horses both with CVCM and eNAD/EDM. As such, elevations were consistent with neurodegenerative disease but were unable to differentiate etiology. In the current study, NEFL was able to differentiate CVCM from both eNAD/EDM and normal horses. However, before use of NEFL as a diagnostic marker, further work is required to evaluate the protein quantitatively. Ultrasensitive NEFL quantification, such as single molecule analysis (SIMOA), has not yet been reported in the horse. In dogs, use of this technology allowed for the quantification of NEFL in serum/plasma and CSF at picogram concentrations and showed promise as a diagnostic, prognostic, and treatment response biomarker.[27-30]

In addition to NEFL, other proteins that are under evaluation as biomarkers in human neurologic disease were detected. These included Dickkoph-1 (DDK1), an important inhibitor of WNT signaling.[31] Elevations of serum DKK1 have been demonstrated in SCI and may be related to severity of injury.[32] Additionally, DKK1 is under investigation as a target for intervention in Alzheimer's disease.[33] RSPO1 was the highest ranked marker following random forest analysis and conditional inference modelling in CSF. RSPO1 is a positive regulator of canonical WNT/β-catenin signaling and has been shown to increase in pre-symptomatic and affected patients with familial Alzheimer's disease.[34,35] Interestingly, in the current study, it is decreased in both CVCM and eNAD/EDM compared to normal horses. RSPO1 elevation in humans is considered to be a consequence of the pathology associated with Alzheimer's, as spondins has shown protective potential in murine Alzheimer's models.[36] Therefore, despite the difference in trajectory, enrichment in normal horses may indicate a protective function or be reflective of the loss of specific neuron

populations.[37] Serum WAP, follistatin/kazal, immunoglobulin, kunitz and neutrin domain containing 1 (WFIKN1) was an important predictor in the serum only modelling. In CSF, reduced abundance of this protein has been associated with treatment for central nervous system leukemia subsequent to acute lymphocytic leukemia.[38] Using the same Olink panel, elevated plasma WFIKN1 was strongly associated with genetic predisposition to schizophrenia.[39] The role of WFIKN1 in neurologic disease is yet to be fully defined, however, it is a known regulator of TGFβ with this signaling pathway intimately associated with neuronal maintenance, function, and degeneration.[40] Collectively, the consistency of protein signals between the current study to those performed in humans and murine models demonstrate the conservation of molecular mechanisms underpinning neurodegeneration. This gives additional confidence in the use of PEA technology for evaluation in equine samples, with similar signals detected when compared to human and murine data.

Additional potential biomarkers of interest were detected in this study, including CSF calretinin (CALB2) and serum superoxide dismutase-2 (SOD2). CALB2 has previously been used to define the tracts that undergo degeneration in eNAD/EDM.[41] However, detection of this protein in CSF as a marker of neurodegeneration has not previously been demonstrated and represents a novel finding. SOD2 elevations helped define the CVCM horse serum proteome from eNAD/EDM horse proteome. SOD2 is the mitochondrial member of the SOD family, and similar to our findings it is decreased in murine models of SCI.[42] These finding again validates the use of Olink technology for biomarker discovery in non-human species, including horses, as it reflects known molecular components of neurodegeneration in that specie

The findings from this study have highlighted several potential biomarkers for equine neurodegenerative disease. However, the study also has several inherent caveats. First, unequal distribution of disease groups may have affected our machine learning algorithms, particularly with the smaller CVCM sample set. Due to the strict disease category approach that was taken (i.e. that necropsy was required) our available sample pool was reduced. Despite the smaller size, the effect magnitude was greater and somewhat more homogeneous as compared to both the normal and eNAD/EDM groups, allowing for partial abrogation of the sample size deficit. The second caveat is that this study does not include infectious, inflammatory, or injurious etiologies of spinal ataxia. While, diagnostics currently exist for many of the infectious causes of spinal ataxia in horses validation of potential biomarkers will require evaluation of these additional etiologies of spinal ataxia to ensure the diagnostic specificity for neurodegenerative disease. The final caveat is that this study captures a snapshot in time for both disease groups. Temporal dynamics of these biomarkers may play a role in their diagnostic use and therefore prospective repeated sampling studies evaluating this will also be valuable to increase the confidence of the current findings.

In conclusion, this study is the first to report on the use of PEA proteomic technology for equine serum and CSF samples. We demonstrate the effectiveness of this technology that, in tandem with machine learning algorithms, shows promise for highly accurate dual- and tri-plex biomarkers for diagnosis of CVCM and eNAD/EDM. Future work to validate these markers quantitatively and in larger replication cohorts are required before clinical adoption of these biomarkers.

References

1. Larson, E. (2014, June 03). Seattle Slew's Senior Surgeries. Retrieved October 06, 2020, from https://thehorse.com/136004/seattle-slews-senior-surgeries
2. van Biervliet, J., Scrivani, P. V., Divers, T. J., Erb, H. N., de Lahunta, A., & Nixon, A. (2004). Evaluation of decision criteria for detection of spinal cord compression based on cervical myelography in horses: 38 cases (1981-2001). *Equine Veterinary Journal*, *36*(1), 14–20. https://doi.org/10.2746/0425164044864642
3. Furr, M., & Reed, S. M. (Eds.) (2015). Differential diagnosis of equine spinal ataxia. *Equine Neurology (2nd Ed)*, 95-99. John Wiley & Sons
4. Burns, E. N., & Finno, C. J. (2018). Equine degenerative myeloencephalopathy: prevalence, impact, and management. *Veterinary Medicine (Auckland, N.Z.)*, *9*, 63–67. https://doi.org/10.2147/VMRR.S148542
5. Hales, E. N., Habib, H., Favro, G., Katzman, S., Sakai, R. R., Marquardt, S., Bordbari, M. H., Ming-Whitfield, B., Peterson, J., Dahlgren, A. R., Rivas, V., Ramirez, C. A., Peng, S., Donnelly, C. G., Dizmang, B. S., Kallenberg, A., Grahn, R., Miller, A. D., Woolard, K., Moeller, B., … Finno, C. J. (2021). Increased α-tocopherol metabolism in horses with equine neuroaxonal dystrophy. *Journal of veterinary internal medicine*, *35*(5), 2473–2485. https://doi.org/10.1111/jvim.16233
6. Shi, M., Caudle, W. M., & Zhang, J. (2009). Biomarker discovery in neurodegenerative diseases: a proteomic approach. *Neurobiology of Disease*, *35*(2), 157-164.

7. Khalil, M., Teunissen, C.E., Otto, M. *et al.* Neurofilaments as biomarkers in neurological disorders. *Nature Reviews Neurology* **14,** 577–589 (2018). https://doi.org/10.1038/s41582-018-0058-z
8. Edwards, L. A., Donnelly, C. G., Reed, S. M., Valberg, S., Chigerwe, M., Johnson, A. L., & Finno, C. J. (2021). Serum and cerebrospinal fluid phosphorylated neurofilament heavy protein concentrations in equine neurodegenerative diseases. *Equine veterinary journal*, 10.1111/evj.13452. Advance online publication. https://doi.org/10.1111/evj.13452
9. Dobbin, K. K., Zhao, Y., & Simon, R. M. (2008). How large a training set is needed to develop a classifier for microarray data?. *Clinical cancer research : an official journal of the American Association for Cancer Research*, *14*(1), 108–114. https://doi.org/10.1158/1078-0432.CCR-07-0443
10. Pepe, M. S., Li, C. I., & Feng, Z. (2015). Improving the quality of biomarker discovery research: the right samples and enough of them. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, *24*(6), 944–950. https://doi.org/10.1158/1055-9965.EPI-14-1227
11. Assarsson, E., Lundberg, M., Holmquist, G., Björkesten, J., Thorsen, S. B., Ekman, D., Eriksson, A., Rennel Dickens, E., Ohlsson, S., Edfeldt, G., Andersson, A. C., Lindstedt, P., Stenvang, J., Gullberg, M., & Fredriksson, S. (2014). Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PloS one*, *9*(4), e95192. https://doi.org/10.1371/journal.pone.0095192

12. Broccardo, C. J., Hussey, G. S., Goehring, L., Lunn, P., & Prenni, J. E. (2014). Proteomic characterization of equine cerebrospinal fluid. *Journal of Equine Veterinary Science*, *34*(3), 451-458. https://doi.org/10.1016/j.jevs.2013.07.013
13. Chiaradia, E., & Miller, I. (2020). In slow pace towards the proteome of equine body fluids. *Journal of proteomics*, *225*, 103880. https://doi.org/10.1016/j.jprot.2020.103880
14. Löfgren, M., Svala, E., Lindahl, A., Skiöldebrand, E., & Ekman, S. (2018). Time-dependent changes in gene expression induced in vitro by interleukin-1β in equine articular cartilage. *Research in veterinary science*, *118*, 466–476. https://doi.org/10.1016/j.rvsc.2018.04.013
15. Whelan, C. D., Mattsson, N., Nagle, M. W., Vijayaraghavan, S., Hyde, C., Janelidze, S., Stomrud, E., Lee, J., Fitz, L., Samad, T. A., Ramaswamy, G., Margolin, R. A., Malarstig, A., & Hansson, O. (2019). Multiplex proteomics identifies novel CSF and plasma biomarkers of early Alzheimer's disease. *Acta neuropathologica communications*, *7*(1), 169. https://doi.org/10.1186/s40478-019-0795-2
16. Isung, J., Granqvist, M., Trepci, A., Huang, J., Schwieler, L., Kierkegaard, M., Erhardt, S., Jokinen, J., & Piehl, F. (2021). Differential effects on blood and cerebrospinal fluid immune protein markers and kynurenine pathway metabolites from aerobic physical exercise in healthy subjects. *Scientific reports*, *11*(1), 1669. https://doi.org/10.1038/s41598-021-81306-4
17. Breiman, L. (2001). Random Forests. Machine Learning *45*, 5-32. https://doi.org/10.1023/A:1010933404324

18. Prabowo, B. A., Cabral, P. D., Freitas, P., & Fernandes, E. (2021). The challenges of developing biosensors for clinical assessment: a review. *Chemosensors*, *9*(11), 299. **https://doi.org/10.3390/chemosensors9110299**
19. Shahim, P., Politis, A., van der Merwe, A., Moore, B., Chou, Y. Y., Pham, D. L., Butman, J. A., Diaz-Arrastia, R., Gill, J. M., Brody, D. L., Zetterberg, H., Blennow, K., & Chan, L. (2020). Neurofilament light as a biomarker in traumatic brain injury. *Neurology*, *95*(6), e610–e622. https://doi.org/10.1212/WNL.0000000000009983
20. Graham, N., Zimmerman, K. A., Moro, F., Heslegrave, A., Maillard, S. A., Bernini, A., Miroz, J. P., Donat, C. K., Lopez, M. Y., Bourke, N., Jolly, A. E., Mallas, E. J., Soreq, E., Wilson, M. H., Fatania, G., Roi, D., Patel, M. C., Garbero, E., Nattino, G., Baciu, C., … Sharp, D. J. (2021). Axonal marker neurofilament light predicts long-term outcomes and progressive neurodegeneration after traumatic brain injury. *Science translational medicine*, *13*(613), eabg9922. https://doi.org/10.1126/scitranslmed.abg9922
21. Ringger, N. C., Giguere, S., Morresey, P. R., Yang, C., & Shaw, G. (2011). Biomarkers of brain injury in foals with hypoxic-ischemic encephalopathy. *Journal of veterinary internal medicine*, *25*(1), 132-137. **https://doi.org/10.1111/j.1939-1676.2010.0645.x**
22. Kuhle, J., Gaiottino, J., Leppert, D., Petzold, A., Bestwick, J. P., Malaspina, A., Lu, C. H., Dobson, R., Disanto, G., Norgren, N., Nissim, A., Kappos, L., Hurlbert, J., Yong, V. W., Giovannoni, G., & Casha, S. (2015). Serum neurofilament light chain is a biomarker of human spinal cord injury severity

and outcome. *Journal of neurology, neurosurgery, and psychiatry*, *86*(3), 273–279. https://doi.org/10.1136/jnnp-2013-307454

23. Intan-Shameha, A. R., Divers, T. J., Morrow, J. K., Graves, A., Olsen, E., Johnson, A. L., & Mohammed, H. O. (2017). Phosphorylated neurofilament H (pNF-H) as a potential diagnostic marker for neurological disorders in horses. *Research in veterinary science*, *114*, 401–405. https://doi.org/10.1016/j.rvsc.2017.07.020
24. Morales Gómez, A. M., Zhu, S., Palmer, S., Olsen, E., Ness, S. L., Divers, T. J., Bischoff, K., & Mohammed, H. O. (2019). Analysis of neurofilament concentration in healthy adult horses and utility in the diagnosis of equine protozoal myeloencephalitis and equine motor neuron disease. *Research in veterinary science*, *125*, 1–6. https://doi.org/10.1016/j.rvsc.2019.04.018
25. Rojas-Núñez, I., Gomez, A. M., Selland, E. K., Oduol, T., Wolf, S., Palmer, S., & Mohammed, H. O. (2022). Levels of Serum Phosphorylated Neurofilament Heavy Subunit in Clinically Healthy Standardbred Horses. *Journal of Equine Veterinary Science*, *110*, 103861. https://doi.org/10.1016/j.jevs.2021.103861
26. Stratford, C. H., Pemberton, A., Cameron, L., & McGorum, B. C. (2013). Plasma neurofilament pNF-H concentration is not increased in acute equine grass sickness. *Equine Veterinary Journal*, *45*(2), 254-255. **https://doi.org/10.1111/j.2042-3306.2012.00603.x**
27. Yun, T., Koo, Y., Chae, Y., Lee, D., Kim, H., Kim, S., Chang, D., Na, K. J., Yang, M. P., & Kang, B. T. (2021). Neurofilament light chain as a biomarker of meningoencephalitis of unknown etiology in dogs. *Journal of veterinary internal medicine*, *35*(4), 1865–1872. https://doi.org/10.1111/jvim.16184

28. Perino, J., Patterson, M., Momen, M., Borisova, M., Heslegrave, A., Zetterberg, H., Gruel, J., Binversie, E., Baker, L., Svaren, J., & Sample, S. J. (2021). Neurofilament light plasma concentration positively associates with age and negatively associates with weight and height in the dog. *Neuroscience letters*, *744*, 135593. https://doi.org/10.1016/j.neulet.2020.135593
29. Vikartovska, Z., Farbakova, J., Smolek, T., Hanes, J., Zilka, N., Hornakova, L., Humenik, F., Maloveska, M., Hudakova, N., & Cizkova, D. (2021). Novel Diagnostic Tools for Identifying Cognitive Impairment in Dogs: Behavior, Biomarkers, and Pathology. *Frontiers in veterinary science*, *7*, 551895. https://doi.org/10.3389/fvets.2020.551895
30. Fefer, G., Panek, W. K., Khan, M. Z., Singer, M., Westermeyer, H., Mowat, F. M., Murdoch, D. M., Case, B., Olby, N. J., & Gruen, M. E. (2022). Use of Cognitive Testing, Questionnaires, and Plasma Biomarkers to Quantify Cognitive Impairment in an Aging Pet Dog Population. *Journal of Alzheimer's disease : JAD*, 10.3233/JAD-215562. Advance online publication. https://doi.org/10.3233/JAD-215562
31. Park, M. H., Sung, E. A., Sell, M., & Chae, W. J. (2021). Dickkopf1: An Immunomodulator in Tissue Injury, Inflammation, and Repair. *ImmunoHorizons*, *5*(11), 898–908. https://doi.org/10.4049/immunohorizons.2100015
32. Gifre, L., Vidal, J., Carrasco, J. L., Filella, X., Ruiz-Gaspà, S., Muxi, A., Portell, E., Monegal, A., Guañabens, N., & Peris, P. (2015). Effect of recent spinal cord injury on wnt signaling antagonists (sclerostin and dkk-1) and their relationship with bone loss. A 12-month prospective study. *Journal of bone*

and mineral research : the official journal of the American Society for Bone and Mineral Research, *30*(6), 1014–1021. https://doi.org/10.1002/jbmr.2423

33. Sellers, K. J., Elliott, C., Jackson, J., Ghosh, A., Ribe, E., Rojo, A. I., Jarosz-Griffiths, H. H., Watson, I. A., Xia, W., Semenov, M., Morin, P., Hooper, N. M., Porter, R., Preston, J., Al-Shawi, R., Baillie, G., Lovestone, S., Cuadrado, A., Harte, M., Simons, P., … Killick, R. (2018). Amyloid β synaptotoxicity is Wnt-PCP dependent and blocked by fasudil. *Alzheimer's & dementia : the journal of the Alzheimer's Association*, *14*(3), 306–317. https://doi.org/10.1016/j.jalz.2017.09.008
34. Nagano K. (2019). R-spondin signaling as a pivotal regulator of tissue development and homeostasis. *The Japanese dental science review*, *55*(1), 80–87. https://doi.org/10.1016/j.jdsr.2019.03.001
35. Ringman, J. M., Schulman, H., Becker, C., Jones, T., Bai, Y., Immermann, F., Cole, G., Sokolow, S., Gylys, K., Geschwind, D. H., Cummings, J. L., & Wan, H. I. (2012). Proteomic changes in cerebrospinal fluid of presymptomatic and affected persons carrying familial Alzheimer disease mutations. *Archives of neurology*, *69*(1), 96–104. https://doi.org/10.1001/archneurol.2011.642
36. Park, S. Y., Kang, J. Y., Lee, T., Nam, D., Jeon, C. J., & Kim, J. B. (2020). *SPON1* Can Reduce Amyloid Beta and Reverse Cognitive Impairment and Memory Dysfunction in Alzheimer's Disease Mouse Model. *Cells*, *9*(5), 1275. https://doi.org/10.3390/cells9051275
37. Li, C. L., Li, K. C., Wu, D., Chen, Y., Luo, H., Zhao, J. R., Wang, S. S., Sun, M. M., Lu, Y. J., Zhong, Y. Q., Hu, X. Y., Hou, R., Zhou, B. B., Bao, L., Xiao, H. S., & Zhang, X. (2016). Somatosensory neuron types identified by high-

coverage single-cell RNA-sequencing and functional heterogeneity. *Cell research*, *26*(1), 83–102. https://doi.org/10.1038/cr.2015.149

38. Mo, F., Ma, X., Liu, X., Zhou, R., Zhao, Y., & Zhou, H. (2019). Altered CSF Proteomic Profiling of Paediatric Acute Lymphocytic Leukemia Patients with CNS Infiltration. *Journal of oncology*, *2019*, 3283629. https://doi.org/10.1155/2019/3283629
39. Png, G., Barysenka, A., Repetto, L., Navarro, P., Shen, X., Pietzner, M., Wheeler, E., Wareham, N. J., Langenberg, C., Tsafantakis, E., Karaleftheri, M., Dedoussis, G., Mälarstig, A., Wilson, J. F., Gilly, A., & Zeggini, E. (2021). Mapping the serum proteome to neurological diseases using whole genome sequencing. *Nature communications*, *12*(1), 7042. https://doi.org/10.1038/s41467-021-27387-1
40. Tesseur, I., & Wyss-Coray, T. (2006). A role for TGF-beta signaling in neurodegeneration: evidence from genetically engineered models. *Current Alzheimer research*, *3*(5), 505–513. https://doi.org/10.2174/156720506779025297
41. Finno, C. J., Valberg, S. J., Shivers, J., D'Almeida, E., & Armién, A. G. (2016). Evidence of the Primary Afferent Tracts Undergoing Neurodegeneration in Horses With Equine Degenerative Myeloencephalopathy Based on Calretinin Immunohistochemical Localization. *Veterinary pathology*, *53*(1), 77–86. https://doi.org/10.1177/0300985815598787
42. Savikj, M., Kostovski, E., Lundell, L. S., Iversen, P. O., Massart, J., & Widegren, U. (2019). Altered oxidative stress and antioxidant defence in skeletal muscle during the first year following spinal cord injury. *Physiological reports*, *7*(16), e14218. https://doi.org/10.14814/phy2.14218

Concluding Discussion

Unlike other companion species such as dogs, there have been no large-scale attempts to generate an equine resource for use in precision medicine. The Pioneer 100 Horses Health Project (P100HHP) bridges this gap, offering an unparalleled multiomic and deep phenotype resource for the equine health research community. The resource development of the P100HHP, and demonstrating the use of the P100HHP data in improving our understanding of Equine Metabolic Syndrome (EMS), have greatly advanced the field of veterinary precision medicine. Evidence of the benefit of the P100HHP resource is the novel demonstration of bile acid signalling in the pathogenesis of EMS, thereby establishing a direct link between the microbiota and host metabolism that has so far been challenging to define. Only through the deep phenotyping and multiomic approaches developed in the P100HHP have these observations been possible. While a highly impactful finding, subsequent investigation in additional cohorts is required to validate these findings.

In our investigation of equine neurologic disease using a precision medicine approach, we make several contributions. First, in juvenile horses raised in vitamin E deplete conditions, subclinical axonal degeneration is likely occurring. This is despite the lack of clinical progression to spinal ataxia as seen in foals the genetic predisposition to develop equine neuroaxonal dystrophy (eNAD). Second, environmental conditions (i.e. serum vitamin E concentration) are important factors to account for when interpreting the biomarker of axonal damage, phosphorylate-neurofilament heavy in juvenile horses. Based on the need to improve the diagnostic modalities available for use in equine spinal ataxia, we next investigated serum and cerebrospinal fluid (CSF) proteomes for potential biomarkers. Using machine

learning approaches we identified two proteins, R-spondin-1 (RSPO1) and neurofilament-light (NEFL) in CSF that, when measured together, are more than 75% accurate in predicating eNAD and more that 85% accurate in predicting normal horses and horses with cervical vertebral compressive myelopathy. This level of accuracy in necropsy confirmed cases may allow for better diagnosis antemortem in horses with spinal ataxia. Future work is required to quantitate these proteins in validation cohorts before these biomarkers may be used as clinical diagnostic measures.

Development of phenotype, omic and data analysis resources in this thesis will drive the progress towards precision medicine for horses. We have shown the power of combining deep longitudinal phenotyping with multiomic data to advance the understanding of equine disease, with specific reference to equine metabolic syndrome and neurodegenerative diseases. Future adoption of precision medicine approaches like those demonstrated in this thesis will improve the diagnosis and treatment of other equine diseases to the benefit of horse health and welfare.

Addendum: Generation of a biobank from two adult stallions for the Functional Annotation of the Animal Genome (FAANG) initiative

Authors: CG Donnelly[1], RR Bellone[2], EN Hales[3], A Nguyen[1], SA Katzman[4], GA Dujovne[1], KE Knickelbein[4], F Avila[2], TS Kalbfleisch[5], E Giulotto[6], N Kingsley[2], J Tanaka[2], E Esdaile[2], S Peng[1], A Dahlgren[1], A Fuller[7], MJ Mienaltowski[8], T Raudsepp[9], VK Affolter[10], JL Petersen[7], CJ Finno[1]

[1] Department of Population Health and Reproduction, School of Veterinary Medicine, University of California-Davis, Davis, CA, 95616, USA.

[2] Veterinary Genetics Laboratory, School of Veterinary Medicine, University of California-Davis, Davis, CA, 95616, USA.

[3] Morris Animal Foundation, Denver, CO, 80246, USA.

[4] Department of Surgical and Radiological Sciences, School of Veterinary Medicine, University of California-Davis, Davis, CA, 95618, USA.

[5] Department of Biochemistry and Molecular Genetics, School of Medicine, University of Louisville, Louisville, KY, 40292, USA.

[6] Department of Biology and Biotechnology, University of Pavia, via Ferrata 1, Pavia, I-27100, Italy.

[7]Department of Animal Science, University of Nebraska - Lincoln, Lincoln, NE, 68583, USA.

[8]Department of Animal Science, College of Agricultural and Environmental Sciences, University of California-Davis, Davis, CA, 95616, USA.

[9] Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX, 77845, USA

[10]Department of Pathology, Microbiology, and Immunology, School of Veterinary Medicine, University of California-Davis, Davis, CA, 95616, USA.

Reference: Donnelly, C. G., Bellone, R. R., Hales, E. N., Nguyen, A., Katzman, S. A., Dujovne, G. A., Knickelbein, K. E., Avila, F., Kalbfleisch, T. S., Giulotto, E., Kingsley, N. B., Tanaka, J., Esdaile, E., Peng, S., Dahlgren, A., Fuller, A., Mienaltowski, M. J., Raudsepp, T., Affolter, V. K., Petersen, J. L., … Finno, C. J**. (2021).** Generation of a Biobank From Two Adult Thoroughbred Stallions for the Functional Annotation of Animal Genomes Initiative. *Frontiers in genetics*, *12*, 650305. https://doi.org/10.3389/fgene.2021.650305

Abstract

Following the successful creation of a biobank from two adult Thoroughbred mares, this study aimed to recapitulate sample collection in two adult Thoroughbred stallions as part of the Functional Annotation of the Animal Genome (FAANG) initiative. Both stallions underwent thorough physical, lameness, neurologic, and ophthalmic (including electroretinography) examinations prior to humane euthanasia. Epididymal sperm was recovered from both stallions immediately postmortem and cryopreserved. Aseptically collected full thickness skin biopsies were used to isolate, culture and cryopreserve dermal fibroblasts. Serum, plasma, cerebrospinal fluid, urine and gastrointestinal content from various locations were collected and cryopreserved. Under guidance of a board-certified veterinary anatomic pathologist, 102 representative tissue samples were collected from both horses. Whole tissue samples were flash-frozen and prioritized tissues had nuclei isolated and cryopreserved. Spatially contemporaneous samples of each tissue were submitted

for histologic examination. Antemortem and gross pathologic examination revealed mild abnormalities in both stallions. One stallion (ECA_UCD_AH3) had unilateral thoracic limb lameness and bilateral chorioretinal scars. The second stallion (ECA_UCD_AH4) had subtle symmetrical pelvic limb ataxia, symmetrical prostatomegally, and moderate gastrointestinal nematodiasis. DNA from each was whole-genome sequenced and genotyped using the GGP Equine 70K SNP array. The genomic resources and banked biological samples from these animals' augments the existing resource available to the equine genomics community. Importantly we may now improve the resolution of tissue-specific gene regulation as affected by sex, as well as add sex-specific tissues and gametes.

Introduction

Sex bias in animal model-based biomedical research has been a focus area for the National Institutes for Health (NIH).[1] While the influence of sex on experimental outcomes is undeniable, the adoption of sex as a consideration for experimental design has been slow.[2] Genomic regulation is particularly labile to the effect of sex.[3] With the intent of cataloguing the functional and regulatory elements of economically significant animal species, there is need for the Functional Annotation of Animal Genome (FAANG) project to diligently account for sex differences. This is of importance to the equine industry in which sex of animal is an important component of management and use, such as predominance of females in polo competition and males in show jumping.[4] The increased resolution of tissue specific genomic, transcriptional and epigenomic data will be of significant benefit to the progress of equine genomics. In particular, the ability to more precisely characterise regulation of genomic regions associated with traits pertinent to equine health and performance

are made increasingly more possible.[5] The objective of the current study was to augment the existing equine biobank with male samples. We report here the addition of two male Thoroughbred horses to the biobank of two female Thoroughbred horses initially established in 2016.

Materials and Methods

Animals

Two Thoroughbred stallions (aged three and four years; ECA_UCD_AH3 and ECA_UCD_AH4 respectively) were donated for this project. ECA_UCD_AH3 had been in racing training and sustained a career ending musculoskeletal injury prior to donation. ECA_UCD_AH4 had not engaged in racing training, and is a son of the reference genome donor Twilight. Approval for all protocols was granted by the UC Davis Institutional Animal Care and Use Committee (Protocol #21033).

Clinical Assessments

Complete physical examination (including complete blood count and serum biochemistry) was performed and interpreted by two board-certified internists (CJF and CGD) for both stallions. Complete neurologic and lameness examinations were performed by relevant recognised experts (CJF and SAK respectively) for both stallions within 24 hours prior to euthanasia. Complete ophthalmic examination of the anterior and posterior segment (performed by KEK) as well as photopic and scotopic electroretinography (ERG) was performed on both stallions within 48 hours prior to euthanasia.

Clinical Sample Collection

Serum and plasma samples were collected immediately prior to euthanasia via an indwelling intravenous cannula. Heparin and EDTA whole blood was centrifuged at 2000 x g for 10 minutes at 4^0C. Plasma harvested, flash frozen in liquid nitrogen and stored at -80°C. Whole blood in plain tubes was allowed to clot for 30 minutes at room temperature, with serum collected in the same fashion as plasma.

Ejaculated sperm and seminal plasma were collected from ECA_UCD_AH3 at the time of ophthalmic examination. Ejaculated sperm was unable to be collected from ECA_UCD_AH4. Seminal plasma was retrieved by un-cushioned centrifugation at 2000 x g for 15 minutes. The supernatant was separated and centrifuged again before being passed through a 20 µm filter. The supernatant and sperm pellet were flash frozen in liquid nitrogen and stored at -80°C.

Immediately following euthanasia, the left testicle, epididymis and distal ductus deferens were removed from each horse. The epididymes were removed and incised sharply transversely at the cauda epididymis. Commercial semen extender was flushed retrograde via the remaining ductus deferens. Recovered spermatozoa were frozen in 0.5mL polyethylene straws by controlled rate freezing protocol and stored in liquid nitrogen, at 323 and 455.5 million sperm/mL for ECA_UCD_AH3 and ECA_UCD_AH4 respectively.

Synovial fluid was collected aseptically (by centesis) from the left middle carpal radio-carpal joint and the left medial compartment of the femoro-tibial joint from ECA_UCD_AH3 and the right middle carpal radio-carpal joint and left medial

compartment of the femoro-tibial joint from ECA_UCD_AH4 immediately following euthanasia. A veterinary clinical pathologist cytologically evaluated an aliquot of each sample within two hours of collection. The remaining sample was centrifuged at 2000 x g for 10 minutes at 4°C, and the supernatant was flash frozen in liquid nitrogen and stored at -80°C.

Cerebrospinal fluid was collected aseptically via atlanto-occipital centesis. A veterinary clinical pathologist cytologically evaluated an aliquot of each sample within two hours of collection. The remaining sample was then processed in the same manner as synovial fluid.

Peripheral Blood Mononuclear Cell (PBMC) Collection

Whole blood was collected in EDTA tubes 24 hours prior to humane euthanasia. Using density gradient centrifugation, PBMCs were harvested as previously described.[6] Briefly, whole blood was overlayed on a double gradient of 1077 and 1119 histopaque media (Sigma-Aldrich). Then, Samples were centrifuged at 700g for 30 minutes with the enriched PBMC layer carefully removed, flash frozen in liquid nitrogen and stored at -80°C.

Tissue Specific Sampling

Full thickness skin biopsies were aseptically collected from the area overlying the left gluteal muscles for dermal fibroblast isolation and cultured as previously described.[7] Briefly, biopsies were washed in ice cold PBS with the addition of penicillin and streptomycin. Fragments of the dermis (2-3mm^2) were placed into a 24-well tissue-culture-treated plate and covered in complete media (Dulbecco's Minimum Essential

Medium, 20% fetal bovine serum, 2x non-essential amino acids, 2m$_M$ L-glutamine, 2xpenicillin/streptomycin, 2µg/ml amphotericin B and 1µg/ml fluconazole). At confluence, cells were trypsinized, counted, and seeded in a 12-well plate. Cells were passaged and frozen at passage three and four in a DMSO-based cryoprotectant media and stored in liquid nitrogen.[7] DNA isolated from whole blood and from cultured fibroblast cells of each horse were genotyped and compared for fifteen standard genetic markers (fourteen microsatellites and one sex link marker amelogenin) routinely tested for use in identify and parentage testing services as the UC Davis Veterinary Genetics Laboratory.

A total of 102 issue samples were taken from all body systems (**Table A1 S1**). Sample stations were arranged in teams, with a veterinarian overseeing each team to ensure appropriate tissue identification. Additionally, all tissues were examined by a veterinary anatomic pathologist (VKA) for gross abnormalities prior to collection (**Table A1 S2**). Tissue samples were collected as previously described.[8] Briefly, samples from the most representative portion of each tissue were collected and preserved in 10% buffered formalin for histopathology. Samples for nuclei preparation and tissue banking were taken from sites immediately adjacent (proximal and distal) to the representative histopathology sections and flash frozen. Nuclei isolation was performed as similarly to that described for the mare biobank, with the exceptions of testis replacing ovary and C6 spinal cord instead of T1.[8]

Table A1 S1: List of tissues collected for two adult thoroughbred stallions. Asterisk (*) denotes tissues with nuclei isolated.

Musculoskeletal	Cardiovascular	Gastrointestinal
Cartilage • Stifle • Thoracic metacarpo-phalangeal	Left atrium free wall Left ventricle free wall* Right atrium free wall Right ventricle free wall Aortic valve	Liver (left lobe)* Spleen* Tongue Esophagus Stomach

• Pelvic metacarpo-phalangeal Coronary periople (thoracic & pelvic) Deep digital flexor tendon (thoracic & pelvic) Superficial digital flexor tendon (thoracic & pelvic) Suspensory ligament (thoracic & pelvic) Gluteal muscle Longissimus dorsi muscle* Sacrocaudalis lateralis muscle 3rd Metacarpal/tarsal diaphysis Rib bone marrow Femur bone marrow Sesamoid <u>Integument</u> Coronary periople (thoracic & pelvic) Lamina (thoracic* & pelvic) Hoof wall (thoracic & pelvic) Neck skin Unpigmented skin Dorsum skin Gluteal adipose tissue* Abdominal adipose tissue Loin adipose tissue	Mitral valve Pulmonic valve Tricuspid valve <u>Respiratory</u> Trachea Larynx Cricoarytenoid dorsalis Left lung* Mediastinal/bronchiol lymph node* Epiglottis <u>Endocrine</u> Thyroid Adrenal cortex (left & right) Adrenal medulla (left & right) Pancreas <u>Urogenital</u> Kidney cortex (left* & right) Kidney medulla (left & right) Urinary bladder Testis (left & right) Right epididymis • Caput • Cauda • Corpus Bulbourethral gland Prostate Seminal vesicle Ampulla Ductus deferens Corpus spongiosum Corpus caversnosum	Duodenum Jejunum* Ileum Cecum Right dorsal colon* Right ventral colon Left dorsal colon Left ventral colon Small colon <u>Nervous System</u> Dur mater Parietal cortex* Occipital cortex Frontal cortex Temporal cortex Corpus callosum Thalamus Hypothalamus Pituitary Pons Cerebellar vermis Cerebellum lateral hemisphere* Spinal cord • C1 • C6 • T1* • T8 Sciatic nerve <u>Eye</u> Cornea Retina Iris

Table A1 S2: Histopathology findings for two adult thoroughbred stallions		
Tissue	ECA_UCD_AH3	ECA_UCD_AH4
Skin Mid Neck	NSF	NSF
Skin Dorsum	NSF	NSF
Gluteal Adipose	NSF	NSF
Gluteal Muscle	NSF	NSF
Sacrocaudalis Muscle	NSF	NSF
Longissimus Dorsi Muscle	NSF	NSF
Right TL coronary band	NSF	NSF
Right TL & PL Lamina	NSF	NSF
Right PL suspensory lig.	NSF	NSF
Right PL SDFT	NSF	NSF
Left TL & Right PL DDFT	NSF	NSF
Sciatic Nerve	NSF	
Liver	Increased bile ducts/oval cells; mild lymphocytic portal infiltrates;	Rare centrolobular and peripherolobular undefined round cell infiltrate and focal subcapsular fibrosis
Spleen	Active white pulp	NSF
Adrenal Cortex & Medulla	Markedly folded cortex; multifocal; vacuolar change of cortical cells	NSF
Kidney Cortex	NSF	NSF
Kidney Medulla	Mild lymphocytic infiltration	Mild lymphocytic infiltration
Larynx	NSF	Mild subepithelial lympho-plasmocytic infiltrate
CAD Muscle	Minimal interstitial lymphocytes	NSF
Pancreas	Sub capsular fibrosis	NSF
Thyroid	NSF	NSF
Mesenteric Lymph node	Marked cortical (active follicles) and paracortical hyperplasia	Cortical and paracortical hyperplasis
Trachiobronchiol Lymph Node	Marked cortical (active follicles) and paracortical hyperplasia	Cortical and paracortical hyperplasis

Left Lung	Multifocal fibrosis with mild lympho-plasmacellular infiltrates, contracted bronchi with lymphoid infiltrates	Perivascular lymphoid infiltrate; contracted bronchioles
Heart Left Atrium	NSF	Focal perivascular lymphoid myocarditis
Heart Left Ventricle	NSF; 4cm	Small aggregates of undefined round cells; 4.2cm
Heart Right Atrium	NSF	NSF
Heart Right Ventricle	NSF; 4.2cm	Perivascular and interstitial lymph-plasmacytic infiltrate in coronary fat/epicardium; 5.2cm
Ventricular septum	NSF; 4.1xm	NSF; 4.5cm
Mitral Valve	NSF	NSF
Tricuspid Valve	NSF	NSF
Aortic Valve	NSF	NSF
Pulmonic Valve	NSF	NSF
Trachea	NSF	NSF
Tongue	Minimal mutifocal perivascular lymphocytic glossitis	Minimal mutifocal perivascular lymphocytic glossitis
Epiglottis	Minimal lympho-plasmacellular infiltrates in lamina propria	NSF
Esophagus	NSF	NSF
Stomach	Lymphoytic gastritis with a lymph follicle formation	NSF
Duodenum	Mild enteritis	Mild enteritis
Ileum	Mild enteritis	Mild enteritis
Cecum	Lympho-plasmacytic and eosinophilic inflammation with prominent lymphoid tissue and lymphocytic exocytosis	Lympho-plasmacytic and eosinophilic inflammation with prominent lymphoid tissue and lymphocytic exocytosis
Right Ventral Colon	Lympho-plasmacytic and eosinophilic inflammation with prominent lymphoid tissue and lymphocytic exocytosis	Lympho-plasmacytic and eosinophilic inflammation with prominent lymphoid tissue and lymphocytic exocytosis

Left Ventral Colon	Lympho-plasmacytic and eosinophilic inflammation with prominent lymphoid tissue and lymphocytic exocytosis	Lympho-plasmacytic and eosinophilic inflammation with prominent lymphoid tissue and lymphocytic exocytosis
Left Dorsal Colon	Lympho-plasmacytic and eosinophilic inflammation with prominent lymphoid tissue and lymphocytic exocytosis	Lympho-plasmacytic and eosinophilic inflammation with prominent lymphoid tissue and lymphocytic exocytosis
Right Dorsal Colon	Lympho-plasmacytic and eosinophilic inflammation with prominent lymphoid tissue and lymphocytic exocytosis	Lympho-plasmacytic and eosinophilic inflammation with prominent lymphoid tissue and lymphocytic exocytosis
Small Colon	NSF	NSF
Eye (retina, cornea, iris, sclera)	NSF; retina missing	NSF
Frontal Cortex	NSF	Solitary perivascular lymphoid cuff
Occipital Cortex	NSF	NSF
Parietal Cortex	NSF	NSF
Temporal cortex	NSF	Perivascular lymphoid cuff
Pituitary	NSF	NSF
Cerebellum Vermis	NSF	NSF
Cerebellum lateral hemisphere	NSF	NSF
Pons	NSF	Perivascular lymphoid cuff
Thalamus	NSF	NSF
Hypothalamus	NSF	Focal meningeal lymphoid infiltrate
Dura Mater	NSF	NSF
Corpus Callosum	NSF	NSF
Spinal Cord C1	NSF	NSF
Spinal Cord C6	NSF	NSF

Spinal Cord T8	NSF	NSF
Spinal Cord L1	NSF	NSF
Spinal Cord L6	NSF	NSF
Dorsal Root Ganglia C4/5	NSF	NSF
Dorsal Root Ganglia T9/10	NSF	NSF
Urinary Bladder	NSF	NSF
Urethra	NSF	NSF
Prostate	NSF	NSF
Right Testicle	NSF; spermatogenesis present	NSF; spermatogenesis present
Left Testicle	NSF; spermatogenesis present	NSF; spermatogenesis present
Right caput epididymis		NSF
Right cauda epididymis	lymphocytic perivascular epididymitis	NSF
Right corpus epididymis	NSF	Minimal interstitial lymphocytic infiltrates
Seminal vesicle	NSF	NSF
Bulbourethral gland	NSF	NSF
Ampulla	NSF	NSF
Ductus deferens	NSF	NSF
Urethral process/corpus spongiosum	NSF	Lymphocytic balanitis
Body of penis/corpus cavernosum	NSF	NSF
Metacarpal III	NSF	NSF
Sesamoid	NSF	NSF

Genotyping and Whole-Genome Sequencing

Genomic DNA was isolated from whole blood using a previously validated technique.[8] Whole-genome sequencing using an Illumina NovaSeq (150bp paired-end reads) platform was performed at Admera Health, LLC (South Plainfield, NJ), targeting 35X coverage for each horse. Library preparation was performed using the

KAPA library quantification kit for Illumina (Roche Holding AG, Basel, Switzerland). SNP genotyping was performed using the GGP Equine 70k SNP bead chip array (Neogen GeneSeek, Lincoln, NE). Genotype data were merged with those collected from the prior sampling effort of Thoroughbred mares.[8] Marker positions were translated from EquCab2 to EquCab3 using the NCBI Genome Remapping Service (https://www.ncbi.nlm.nih.gov/genome/tools/remap), removing markers that were not mapped to any of the 31 autosomes. Using SNP data, runs of homozygosity (ROH), a measure of diversity, was quantified in DetectRuns (https://github.com/bioinformatics-ptp/detectRUNS/blob/master/detectRUNS/R/run.R) using the method of Marras et al. (2015).[9] Analysis parameters required a minimum of 15 SNPs, maxOpp Run =1 , max MIssRun=1, maxGap=10000000, minLengthBps=100000. Whole-genome sequence data of the stallions and mares were processed and variants called according to the pipeline outlined in Sieck et al. 2020, with the exception of mapping to the EquCab3 reference genome.[8,10] Bi-allelic autosomal SNPs were extracted from the resulting vcf for ROH analysis in the same manner as used for the SNP data.

Additionally, in an effort to provide extensive phenotyping and genotyping for future studies on pigmentation biology, both stallions were genotyped for coat color loci using the commercially available horse full coat color and white pattern panel (https://vgl.ucdavis.edu/panel/full-coat-color-pattern-panel).

Pedigree Analysis

Individual inbreeding coefficients (F) were calculated from five-generation pedigrees of each of the four horses sampled using Pedigraph.[11]

Karyotyping

Sodium heparin stabilized whole blood samples were collect seven days prior to euthanasia for karyotype characterisation. Karyotypes were generated for each horse using the previously validated Pokeweed-stimulated lymphocyte culture.[12] Thirty metaphase cells from ECA_UCD_AH3 and ECA_UCD_AH4 were captured for analysis. Genomic DNA was isolated from lymphocyte cultures for PCR detection of the sex determining region Y *(SRY)* and androgen receptor (*AR)* genes as preciously described.[1]

Results

Clinical Assessments

ECA_UCD_AH3 had several mild abnormalities apparent on clinical examination, including a grade III/V left thoracic limb lameness (AAEP scale[14]) and chorioretinal scarring of both eyes. Scoptopic ERG data for both horses is provided in **Table A1 S3**. No other abnormalities were detected antemortem for ECA_UCD_AH3. ECA_UCD_AH4 did not have clinically detectable lameness, however, a mild grade I/V (Modified Mayhew scale[15]) symmetrical pelvic limb ataxia was detected with a presumptive neuroanatomical localization to the cervical spinal cord. No other abnormalities were detected for ECA_UCD_AH4 on clinical examination. No abnormalities were detected on complete blood count or serum biochemistry for either stallion.

Table A1 S3. a- and b- wave amplitudes and implicit times for scotopic electroretinogrpahy waveforms of two Thoroughbred stallions

	Horse	ECA_UCD_AH3						ECA_UCD_AH4					
	Flash stimulus (cd.s/m2) - white	0.01		3		10		0.01		3		10	
	Eye	OD	OS	OD	OS	OD	OS	OD	OS	OD	OS	OD	OS
a-Wave	amplitude (uV)	16.7	9.0	29.2	91.4	17.2	151.0	5.8	4.5	144.0	110.1	189.3	136.0
	implicit time (ms)	12.8	16.4	12.8	11.8	6.7	10.8	22.5	14.5	14.1	14.6	13.8	14.6
b-Wave	amplitude (uV)	230.8	186.8	358.1	200.8	614.0	375.8	245.6	305.7	364.8	498.1	411.4	444.7
	implicit time (ms)	83.5	91.6	38.9	35.3	69.6	65.5	76.8	67.8	57.7	58.5	61.4	59.2

Clinical Sample Collection

Serum, plasma (collected in EDTA and heparin tubes), buffy coat and urine were collected from both stallions. Seminal plasma and ejaculated spermatozoa were obtained for ECA_UCD_AH3 only. Epididymal recovered spermatozoa from both stallions were successfully isolated and cryopreserved with adequate post thaw kinematic parameters (**Table A1 S4**). Synovial and cerebrospinal fluid samples were cytologically normal (**Table A1 S5 and A1 S6**). PBMCs were successfully collected from both stallions.

Table A1 S4: Epididymal sperm post-thaw analysis for two adult thoroughbred stallions

	ECA_UCD_AH3	ECA_UCD_AH4
Concentration (million/ml)	323.3	455.5
Number of straws		
Rapid progressive%	39.11	52.75
Slow progressive%	0.0	0.0
Non-progressive%	9.43	13.34
Immotile%	51.46	33.91
VCL (μm/sec)	110.2	100.5
VAP (μm/sec)	56.75	51.16
VSL(μm/sec)	43.03	36.59
LIN	0.39	0.36
STR	0.75	0.71
BCF (hertz)	26.23	22.58
ALH (μm)	4.28	4.49

	Table A1 S5: Synovial fluid analysis for two adult thoroughbred stallions					
	ECA_UCD_AH3			ECA_UCD_AH4		
	Left tarso-crural	Right middle carpal	Right radio carpal	Left tarso-crural	Left middle carpal	Left radio carpal
Gross appearance	Yellow/slight haze	Yellow/clear	Yellow/hazy	Yellow/clear	Yellow/hazy	Yellow/clear
Total protein (g/dL)	1.2	1.5	1.5	1.2	2.1	1.7
Mucin clot	Very good	Very good	Very good	Very good	Very good	Very good
Total RBC's (/µL)	<30,000	<30,000	<30,000	Rare	Rare	Rare
Total nucleated cells (/µL)	180	150	180	NA	NA	NA
Neutrophil %	0	1	2	0	1	0
Small mononuclear%	85	77	63	72	78	60
Large mononuclear%	15	22	35	28	21	40
Eosinophil %	0	0	0	0	0	0
Interpretation	No cytological abnormalities	No cytological abnormalities	No cytological abnormalities	No cytological abnormalities	No cytological abnormalities	No cytological abnormalities

	Table A1 S6: Cerebrospinal fluid analysis for two adult thoroughbred stallions	
	ECA_UCD_AH3	ECA_UCD_AH4
Gross appearance	Clear	Clear
Total protein (mg/dL)	72	88
Total RBC's (/µL)	6	<1

Total nucleated cells (/µL)	3	1
Neutrophil %	0	0
Small mononuclear%	95	85
Large mononuclear%	5	15
Eosinophil%	0	0
Interpretation	No cytological abnormalities	No cytological abnormalities

Genotyping

Achieved coverage of the whole-genome sequence data was 38.7 and 37.5X for ECA_UCD_AH3 and ECA_UCD_AH4, respectively. Relative to the reference genome, between 5.4 (ECA_UCD_AH3) and 3.8 (ECA_UCD_AH4) million variants were observed. Data are publicly available through the FAANG consortium. Combining the 70K genotype data of the stallions to that from the two mares and removing variants present on sex chromosome data and on unmapped contigs, 59,823 variants remained. The count of ROH per horse was similar (498 to 503) for all four horses, except ECA_UCD_AH4 (280. However, ECA_UCD_AH4 had a greater number of long (> 8 Mb) ROH than all other horses **(Table A1 S7**). Considering SNP data from WGS, between 1822 (ECA_UCD AH1) and 2867 (ECA_UCD AH4) ROH were observed. Five-generation pedigree-based inbreeding estimates ranged from 0 (ECA_UCD AH1) to 0.344 (ECA_UCD AH4).

Table A1 S7. Runs of homozygosity (ROH) and inbreeding estimate (F_{ROH}) for combined mare and stallion FAANG genome data sets				
Length of ROH (Mb)	ECA_UCD_AH1	ECA_UCD_AH2	ECA_UCD_AH3	ECA_UCD_AH4
0 to 4	466	462	453	220
4 to 8	24	26	38	29
8 to 16	7	11	6	16
16 to 32	3	4	1	11
> 32	0	0	0	4
ROH Genome	765915572	838359933	758961216	1009849409
F_{ROH}	0.336	0.368	0.333	0.444

Genetic profiles from cultured fibroblast cells matched to that of the expected horse (15/15 markers matched for each horse, presented in **Table A1 S8**)

	Table A1 S8. Microsatellite genotyping of cryopreserved fibroblast cell line DNA from two adult thoroughbred stallions compared to whole blood derived genomic DNA			
	Type			
	Whole Blood	Fibroblast	Whole Blood	Fibroblast
Locus	ECA_UCD_AH3	ECA_UCD_AH3	ECA_UCD_AH4	ECA_UCD_AH4
AHT4	JK	JK	K	K
AME	YX	YX	YX	YX
ASB2	N	N	O	O
HMS3	I	I	MP	MP
HMS7	M	M	LM	LM
HTG4	K	K	KM	KM
LEX33	MQ	MQ	Q	Q
HMS2	L	L	L	L
AHT5	J	J	KN	KN
ASB17	GM	GM	NR	NR
ASB23	J	J	J	J
HMS6	MP	MP	MP	MP
HTG10	IO	IO	R	R
LEX3	H	H	N	N
VHL20	I	I	IL	IL

ECA_UCD_AH3 is genetically defined as a bay horse (E/e A/a) with one copy of dominant white 20 (N/W20), while ECA_UCD_AH4 was gray with a bay base coat

(E/e A/A) before graying out and was homozygous for dominant white 20 (W20/W20) (all coat color genotypes are presented in **Table A1 S9**).

Table A1 S9. Color genotyping for two adult thoroughbred stallions		
	ECA_UCD_AH3	ECA_UCD_AH4
Red Factor	E/e	E/e
Agouti	A/a	A/a
Cream	N/N	N/N
Pearl	N/N	N/N
Silver	N/N	N/N
Dun	nd2/nd2	nd2/nd2
Champagne	N/N	N/N
Lethal White Overo	N/N	N/N
Sabino 1	N/N	N/N
Dominant White (W5, W10, W20, W22)	N/W20	W20/W20
Splashed White (SW1, SW3, SW5, SW6)	N/N	N/N
Splashed White (SW2, SW4)	N/N	N/N
Tobiano	N/N	N/N
Leopard	N/N	N/N
Patern-1	N/N	N/N
Gray	N/N	M/G

Karyotype

Both stallions had a normal 64 XY male karyotype with no apparent structural or numerical chromosomal abnormalities. Both stallions were positive for the *SRY* and *AR* genes.

Tissue Specific Sampling

Sample collection commenced 30 minutes following euthanasia and was concluded within three hours for all tissue samples. Samples from 102 tissues were collected and flash frozen (**Table A1 S1**). Sixteen of these tissues also had nuclei isolated and cryopreserved (**Table A1 S1**).

Pathology

Gross examination of ECA_UCD_AH3 was without abnormality, except for a 15x2x1.3cm firm area at the caudal margin of the left caudal lung lobe. This area was therefore avoided for sampling. Histologic examination of the area identified locally extensive fibrosis, Type II pneumocyte proliferation and alveolar histiocytosis, which is consistent with a previous pulmonary inflammatory insult. Mild lymphocytic inflammation was noted at several sections of the gastrointestinal tract, but were considered incidental. Gross examination of ECA_UCD_AH4 revealed marked gastrointestinal nematodiasis throughout the large and small intestinal segments. Gastrointestinal segments had marked lymphocytic and eosinophilic infiltrates consistent with gastrointestinal parasitism. ECA_UCD_AH4 had marked prostatomegally grossly, but was without abnormality on histopathology. There were no gross or histologic findings of the central nervous system of ECA_UCD_AH4 that could explain the mild sensory ataxia. There were no other significant histologic abnormalities. Summary of all pathologic findings are summarized in **Table A1 S2**.

Discussion

We report the successful completion of the male equine biobank for the FAANG consortium. This project has added 102 tissues from two stallions in addition to seven body fluids, one cell line and spermatozoa to the biobank for use by the equine research community. This study recapitulated the guidelines of the original equine tissue collection, demonstrating that this method is repeatable and appropriate for this type of endeavour.

Stringent phenotype information both ante- and postmortem are a critical feature of this biobank and lends strength to the interpretation of results of subsequent studies using these tissues. Similarly, it also establishes the boundaries for the interpretation of data derived from this biobank. This is especially true for the gastrointestinal tissues that had lymphocytic and/or eosinophilic infiltrates. Previous tissue collection from both mares also had a similar infiltrate pattern in the gastrointestinal sections to the stallions described here.[8] While this makes the direct comparison between male and female samples more consistent, it underscores the necessity to perform detailed phenotype characterisation of the tissues. The available phenotype information was expanded in this data set with the addition of ERG measurements. Further, this additional information enhances the ability to select tissue with the most normal function or at minimum document pathology that may affect future assays. In this case, the left retina from ECA_UCD_AH3 with an abnormal waveform was not used for the tissue bank, with the normal right retina retained.

The available whole-genome sequencing and SNP genotyping data from these horses to accompany the biobank and pending tissue-specific regulatory information will assist in future genomic investigations. In this sampling effort, the availability of a stallion closely related to the mare from which the reference genome is based (Twilight) is expected to provide high accuracy mapping and allow for more in depth investigations of achieved versus predicted inbreeding. The intentional inbreeding that resulted in ECA_UCD_AH4 (by breeding record) was corroborated by the frequency and relatively large ROH as identified in the analysis of the SNP genotype data. This also highlights the limitation of this biobank, as it is restricted a single

breed. Future investigation may need to include more diverse biologic replicates to account for between breed variations in gene regulation.[16]

While the equine FAANG project has not progressed to the level of the ENCODE project in humans, important data have already been generated from the equine biobank.[17] Histone modifications as analysed by chromatin immunoprecipitation sequencing have been established for eight tissues from the original two mares.[5] The addition of repressive and enhancing epigenetic marks to the catalogue of equine genome regulation is of great benefit to the equine research community. Sex-biased gene regulation is strongly intertwined with epigenetic regulation, especially as it relates to histone modifications.[18] This reinforces the need for the equine FAANG project to expand the existing biobank with male samples. This undertaking was no small feat and required the sacrifice of two additional animals, the benefit to future equine genomic studies is immense.

Although this study concludes the tissue collection arm of the FAANG project, the work to annotate the equine genome continues in earnest. This addition to the biobank expands the list of tissues available to the equine research community. Groups with specific interest in certain tissues are encouraged to contact the corresponding author for availability, as the ‘adopt-a-tissue’ program successfully used with the original collection is also available for these male samples.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

Clinical examinations were performed by CJF, SAK, KEK and CGD. Semen evaluation and cryopreservation performed by GAD. Karyotyping performed by TR. Tissue collection performed by ENH, AN, KEK, NK, JT, EE, SP, AD,AF, JLP and CGD. All histopathoogy reviewed by VKA. Nuclei preparation performed by RRB and FA. Cell culture and cryopreservation performed by MJM. Genetic analysis performed by JLP. All authors contributed to manuscript preparation.

Acknowledgments

The authors would like to thank all of the technicians at the Center for Equine Health and Anatomic Pathology.

Data Availability

The datasets generated for this study can be found in the FAANG repository fang.org.

References

1. Will, T. R., Proaño, S. B., Thomas, A. M., Kunz, L. M., Thompson, K. C., Ginnari, L. A., Jones, C. H., Lucas, S. C., Reavis, E. M., Dorris, D. M., & Meitzen, J. (2017). Problems and Progress regarding Sex Bias and Omission in Neuroscience Research. *eNeuro*, *4*(6), ENEURO.0278-17.2017. https://doi.org/10.1523/ENEURO.0278-17.2017
2. Zucker, I., Beery, A. Males still dominate animal studies. *Nature* **465,** 690 (2010). https://doi.org/10.1038/465690a
3. Lee S. K. (2018). Sex as an important biological variable in biomedical research. *BMB reports*, *51*(4), 167–173. https://doi.org/10.5483/bmbrep.2018.51.4.034
4. Fenner K, Caspar G, Hyde M, Henshall C, Dhand N, Probyn-Rapsey F, et al. (2019) It's all about the sex, or is it? Humans, horses and temperament. PLoS ONE 14(5): e0216699. https://doi.org/10.1371/journal.pone.0216699
5. Kingsley, N. B., Kern, C., Creppe, C., Hales, E. N., Zhou, H., Kalbfleisch, T. S., MacLeod, J. N., Petersen, J. L., Finno, C. J., & Bellone, R. R. (2019). Functionally Annotating Regulatory Elements in the Equine Genome Using Histone Mark ChIP-Seq. *Genes*, *11*(1), 3. https://doi.org/10.3390/genes11010003
6. Hida N, Maeda Y, Katagiri K, Takasu H, Harada M & Itoh K (2002) A simple culture protocol to detect peptide-specific cytotoxic T lymphocyte precursors in the circulation. Cancer Immunology, Immunotherapy 51, 219–28.
7. Raimondi E, Piras FM, Nergadze SG, Di Meo GP, Ruiz-Herrera A, Ponsa M, Ianuzzi L & Giulotto E (2011) Polymorphic organization of constitutive

heterochromatin in Equus asinus (2n = 62) chromosome 1. Hereditas 148, 110–3.

8. Burns, E. N., Bordbari, M. H., Mienaltowski, M. J., Affolter, V. K., Barro, M. V., Gianino, F., Gianino, G., Giulotto, E., Kalbfleisch, T. S., Katzman, S. A., Lassaline, M., Leeb, T., Mack, M., Müller, E. J., MacLeod, J. N., Ming-Whitfield, B., Alanis, C. R., Raudsepp, T., Scott, E., Vig, S., … Finno, C. J. (2018). Generation of an equine biobank to be used for Functional Annotation of Animal Genomes project. *Animal genetics*, *49*(6), 564–570. https://doi.org/10.1111/age.12717
9. Marras G, Gaspa G, Sorbolini S, Dimauro C, Ajmone-Marsan P, Valentini A, Williams JL, Macciotta NP. Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy. Animal genetics. 2015 Apr 1;46(2):110-21.
10. Sieck, R. L., Fuller, A. M., Bedwell, P. S., Ward, J. A., Sanders, S. K., Xiang, S. H., ... & Steffen, D. J. (2020). Mandibulofacial Dysostosis Attributed to a Recessive Mutation of CYP26C1 in Hereford Cattle. *Genes*, *11*(11), 1246.
11. Garbe, J. R., & Da, Y. (2008). user manual version 2.4. *Department of Animal Science, University of Minnesota*
12. Raudsepp T, Durkin K, Lear TL, Das PJ, Avila F, Kachroo P & Chowdhary BP (2010) Molecular heterogeneity of XY sex reversal in horses. Animal Genetics 41 Suppl 2, 41–52
13. Raudsepp T, Lee EJ, Kata SR, Brinkmeyer C, Mickelson JR, Skow LC, Womack JE & Chowdhary BP (2004) Exceptional conservation of horse-human gene order on X chromosome revealed by highresolution radiation

hybrid mapping. Proceedings of the National Academy of Sciences of the United States of America 101, 2386–91

14. Baxter, G. M. (Ed.). (2020). *Adams and Stashak's lameness in horses*. John Wiley & Sons.
15. Furr, M., & Reed, S. (Eds.). (2015). *Equine neurology*. John Wiley & Sons.
16. Li, M., Wu, H., Luo, Z., Xia, Y., Guan, J., Wang, T., ... & Liu, Y. (2012). An atlas of DNA methylomes in porcine adipose and muscle tissues. *Nature communications*, *3*(1), 1-11.
17. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74
18. Tsai, H. W., Grant, P. A., & Rissman, E. F. (2009). Sex differences in histone modifications in the neonatal mouse brain. *Epigenetics*, *4*(1), 47–53. https://doi.org/10.4161/epi.4.1.7288