Genomic variation in captive deer mouse (*Peromyscus maniculatus*) populations

Matthew D. Lucius[1], Hao Ji [1], Diego Altomare [1], Robert Doran[2], Ben Torkian[2], Amanda Havighorst[1], Vimala Kaza[3], Youwen Zhang[1], Alexander V. Gasparian[1], Joseph Magagnoli[4], Vijay Shankar[5], Michael Shtutman[1], and Hippokratis Kiaris[1,3]*

[1]Department of Drug Discovery and Biomedical Sciences, College of Pharmacy, University of South Carolina, SC, USA.

[2]Research Computing, Division of Information Technology, University of South Carolina, SC, USAa
[3]*Peromyscus* Genetic Stock Center, University of South Carolina, SC, USA

[4]Department of Clinical Pharmacy and Outcomes Sciences, College of Pharmacy, University of South Carolina, Columbia, SC, USA

[5]Center for Human Genetics, College of Science, Clemson University, SC, USA

***, Correspondence**: H. Kiaris (hk@sc.edu)

Abstract

Background: Deer mice (genus *Peromyscus*) are the most common rodents in North America. Despite the availability of reference genomes for some species, a comprehensive database of polymorphisms, especially in those maintained as living stocks and distributed to academic investigators, is missing. In the present study we surveyed two populations of *P. maniculatus* that are maintained at the *Peromyscus* Genetic Stock Center (PGSC) for polymorphisms across their 2.5x10^9 bp genome.

Results: High density of variation was identified, corresponding to one SNP every 55bp for the high altitude stock (SM2) or 207bp for the low altitude stock (BW) using snpEff (v4.3). Indels were detected every 1157bp for BW or 311bp for SM2. The average Watterson estimator for the BW and SM2 populations is 248813.70388 and 869071.7671 respectively. Some differences in the distribution of missense, nonsense and silent mutations were identified between the stocks, as well as polymorphisms in genes associated with inflammation (NFATC2), hypoxia (HIF1a) and cholesterol metabolism (INSIG1) and may possess value in modeling pathology.

Conclusions: This genomic resource, in combination with the availability of *P. maniculatus* from the PGSC, is expected to promote genetic and genomic studies with this animal model.

Introduction

Mammals of the genus *Peromyscus* (deer mice) are the most abundant rodents of North America (1-3). Deer mice play important roles in public health as they have been identified as a natural reservoir for various infectious agents such as Hantaviruses (4,5) and the arthropod- transmitted spirochete *Borrelia burgdorferi* (6,7) that cause Lyme disease, babesiosis, anaplasmosis, viral encephalitis and others. More recently *P. maniculatus* was also shown to be sensitive to SARS-CoV2 infection implying that it may also function as a secondary reservoir for the coronavirus that causes the COVID-19 pandemic (8,9). Because of their abundance and their characteristics, *Peromyscus* species are being used extensively as animal models for studies ranging from evolution, physiology, infectious diseases, metabolism, genetics, aging and behavior (2,3,10-15). In example, P. leucopus lives up to 8 years in captivity as compared to the other *Peromyscus* species that reportedly live up to 4 years providing models for aging studies (16). P. californicus and P. polionotus are strictly monogamous species that are used in studies on behavior (15,17). P. eremicus is adapted for life in the desert providing a model to explore adaptation at extreme environments (2). *P. maniculatus* are being used for a wide array of studies ranging from metabolism and the regulation of stress response to altitude adaptation (2).

A major limitation in understanding better the impact of deer mice in public health , and in exploiting their utility as a research model is the lack of comprehensive genomic variation data in reference populations that are readily accessible to outside users (18,19, 20). The *Peromyscus* Genetic Stock Center at the University of South Carolina maintains different species of deer mice that are maintained as closed, genetically diverse colonies since the original caption of the original colony founders, and distributes them to outside investigators. Among them, 2 populations of *P. maniculatus* are being maintained as outbred, genetically diverse stocks: the BW stock (*Peromyscus* *maniculatus bairdii*) bred in captivity since 1948 and descended from 40 ancestors wild-caught near Ann Arbor, MI, and the SM2 stock (*Peromyscus maniculatus sonoriensis*) derived from about 50 animals, wild-caught by Jack Hayes in 1995 near White Mountain Research Station, CA.

Several *Peromyscus* species, have been sequenced (21,22) while for others~~,~~ including *P. maniculatus*, chromosomal assembly level reference genomes and annotations are available, providing strong foundation for genomic analyses, as opposed to scaffold-level assemblies (23-26). Baylor College of Medicine provided a scaffold *P. maniculatus* reference genome in 2013 with a scaffold N50 of 3,760,915 and a contig N50 of 36,367 and the Hoekstra laboratory at Harvard University and HHMI provided a chromosome level *P. maniculatus* reference genome in 2018 (HU_Pman_2.1) which had a scaffold N50 of 115,033,041 and a contig N50 of 30,111 (27, 28). Nevertheless, a genome-wide database on polymorphisms of *P. maniculatus* is lacking, restricting greatly their usage and their exploitation as genetic models. This limitation is especially pertinent to populations that are purposely maintained as outbred stocks in a stock center and are readily accessible to outside investigators. In this study we performed whole genome sequencing (WGS) to discover polymorphisms in *P. maniculatus* and initiate the establishment of a robust polymorphism database for *P. maniculatus*. Our analyses involved individuals from both the SM2 and BW populations, as an attempt to characterize these 2 distinct, yet highly relevant evolutionarily, subspecies and to record the dynamics on genomic diversity in these closed populations that are maintained in captivity for several decades.

Materials and Methods

Samples and Library Preparation

Animals were sacrificed under isoflurane anesthesia. DNA samples were isolated from the liver by using the DNeasy kit (Qiagen) and quantified with the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen, Cat. No. P7589). To prepare libraries, the TruSeq DNA PCR-free Kit (Illumina, Cat. No. 20016327) was used according to manufacturer recommendations (29). Briefly, genomic DNA (1.1 µg) was diluted to 55 µl with Resuspension Buffer. Samples were sonicated using a Covaris M220 Focused-ultrasonicator with appropriate settings to generate 350 bp fragments. Fragmented DNA was quality controlled using an Agilent 2100 Bioanalyzer and the DNA 1000 Kit (Cat. No. 5067-1504) (30). Fragmented DNA was cleaned up with magnetic beads, end-repaired, and size selected using different ratios of the Sample Purification Beads (SPB). DNA fragments were 3' end adenylated, TruSeq DNA Sgl index adapters (Illumina, Cat. No. 20016329 and 20016330) were ligated to the fragments' ends, and libraries were cleaned up with SPB. Libraries were quantified by qPCR using the NEBNext Library Quant Kit for Illumina (New England Biolabs, Cat. No. E7630S) and fragment size was assessed using an Agilent 2100 Bioanalyzer and the High Sensitivity DNA Kit (Cat. No. 5067-4626) (30). Libraries were pooled and sequenced with NovaSeq S4 (Illumina San Diego, CA) 150 bp Pair ends, by Psomagen (Rockville, MD, USA).

SNP and Indel Calling

The following pipeline used was adapted from the Genome Analysis Toolkit (GATK) Best Practices (31). Paired-end FASTQ files of each sample were aligned to the HU_Pman_2.1 *P. maniculatus* reference genome (GCA_003704035.1, Ensembl release-96) using BWA-MEM (v0.7.17-r1188) (27, 32). The resulting SAM files were sorted by coordinate into BAM file format using Picard (v2.18.15; http://broadinstitute.github.io/picard/) (33). Alignment metrics and duplicate metrics are in Supplemental Tables. Using GATK (v4.0.5.1) HaplotypeCaller variants were called with default parameters and then SNPs and Indels were selected using SelectVariants default parameters. SNPs and Indels were filtered using GATK VariantFiltration with the "filter-expression" parameter with the following limits for SNPs and Indels respectively: "(QD < 2.0) || (FS > 60.0) || (MQ < 40.0) || (MQRankSum < -12.5) || (ReadPosRankSum < -8.0) || (SOR > 3.0)" ; "(QD < 2.0) || (FS > 200.0) || (ReadPosRankSum < -20.0) || (SOR > 10.0)". The GATK BaseRecalibrator tool was run and applied with the filtered SNPs and Indels followed by a repeat of variant calling and filtration with the newly recalibrated bam files. SNPs and Indels are then annotated using snpEff (v4.3); a local snpEff database was built for *P. maniculatus* using the HU_Pman_2.1.96 annotation in GTF file format (34). The resulting variants can be accessed on the European Variation Archive (EVA) with the accession ID PRJEB41333.

ANGSD

The thetas for BW and SM2 thetas were calculated using ANGSD (v0.930) (35). The methods used to calculate the thetas were done according to ANGSD's website page on Thetas, Tajima, and Neutrality tests (http://www.popgen.dk/angsd/index.php/Thetas,Tajima,Neutrality_tests) (36, 37). The SFS estimation was calculated using 500,000,000 bp buckets. The average of the SFS estimations was then calculated across all buckets and used for the calculation of the theta values.

Validation of SNPs

INSIG1 and HIF1a SNPs were validated via PCR and sanger sequencing of the fragments. . Gene-specific primers were ordered from Integrated DNA Technologies (Coralville, Iowa . The primer sequences for INSIG1 are as follows:
Forward Primer: AA-TAATACGACTCACTATAGGG-TTGCCAATAATGTCCAACTG
Reverse Primer: AA-CAGGAAACAGCTATGAC-GAGTGATCAGCGTAGCTAGG
The primer sequences for HIF1a are as follows:
Forward Primer: AA-TAATACGACTCACTATAGGG-TGCCACCACCACCACTACTG
Reverse Primer: AA-CAGGAAACAGCTATGAC-GGCTTTTGCGAGTTTGTTTG
Froward primers have T7 sequence extension and reverse primers have the M13 extensions attached for the following Sanger sequencing with the corresponded T7 and M13 sequencing primers. INSIG1 primers were annealed at 66°C and HIF1a primers were annealed at 72°C. Samples were gel-purified gel purification kit (Zymo Research, Irvine, CA) followed by Sanger sequencing.

Results

Overall diversity in SM2 and BW stocks

Ten (10) individuals from two *P. maniculatus* stocks were subjected to whole genome sequencing (WGS), *Peromyscus maniculatus sonoriensis* (SM2) and *Peromyscus maniculatus bairdii* (BW). The SM2 stock was established by animals captured near the White Mountain Research Station, CA in 1995 and the BW population was captured near Ann Arbor, Michigan in 1948 (2). These 2 populations were continuously maintained isolated, as different stocks, since their original acquisition by the PGSC. The HU_Pman_2.1 reference genome was established by an individual of the BW stock (https://www.ncbi.nlm.nih.gov/genome/browse/#!/eukaryotes/11397/) but provides a decent reference for the SM2 subspecies as well; the disadvantage to using this reference genome for SM2 individuals is divergent reads in SM2 samples may not map properly to the BW reference or may not be mapped altogether. An alternate SM2 reference genome will need to be established to create accurate polymorphic calls. These two populations differ by the altitudes they are found in the wild, with SM2 being found at higher altitudes and BW being found at lower altitudes (50). Paired-end WGS analysis with an average 34X coverage depth, a standard deviation of 5.73X, a minimum of 26.98X coverage and a maximum of 48.47X coverage indicated that BW exhibited an average of about 12.1 million single nucleotide polymorphisms (SNPs) while each SM2 had 42-46 million SNPs, across their $2.5\text{x}10^9$ bp genome (Figure 1a). Using chromosome 1 as a sample of coverage of the genome, each sample had an average of 92.36% coverage for all bases with more than 10 reads per base. Each sample's chromosome 1 had a coverage range of 2.44% and had a coverage standard deviation of 1.08%. All variant data can be found on the European Variation Archive (EVA) under the project ID PRJEB41333. The BW have a range of 2.1-2.2 million insertions/deletions (indels) and the SM2 have 7.4-8.1 million indels (Figure 1a). Each sample in Figure 1 is indicated by the order of their ID in the PGSC. There is an average variant rate of a SNP every 55 bp and an indel every 311 bp in SM2 according to snpEff. BW have a SNP approximately every 207 bp and an indel every 1157 bp according to snpEff. Although the reference genome is aligned for *P. maniculatus bairdii* instead of *P. maniculatus sonoriensis*, the range of SNPs and indels found in SM2 is much wider than that of BW (Figure 1b). The total amount of BW SNPs in each individual had a lower range of 11.79 million SNPs and an upper range of 12.36 million SNPs with a total count of 17.52 million SNPs for the entire sample size. The total amount of SNPs found in each individual SM2 however had a lower range of 43.01 million SNPs and upper range of 46.06 million SNPS with a total count of 48.36 SNPs in the SM2 sample size. The Watterson's estimator of theta for each chromosome in BW and SM2 samples was found using ANGSD. The average theta of each chromosome for BW samples is 248,813.70. The minimum theta was in chromosome 13 with a value of 49,976.30 and the maximum theta was in chromosome 2 with a value of 480,638.38. The average theta of each chromosome for SM2 samples is 869,071.77. The minimum theta was in chromosome 22 with a value of 444,356.96. The maximum theta was in chromosome 1 with a value of 1,537,521.37. The missense and nonsense polymorphisms have an average heterozygosity of 0.0013 with a standard deviation of 8.48×10^{-5} in SM2 and an average heterozygosity of 0.0019 with a standard deviation of 5.29×10^{-5} in BW (Figure 2a). The synonymous polymorphisms have an average heterozygosity of 0.0023 with a standard deviation of 0.0003 in SM2 and an average heterozygosity of 0.0045 with a standard deviation of 0.0007 in BW (Figure 2b).

Variation in the incidence of missense, nonsense and silent mutations between SM2 and BW.

Between the stocks, the distribution of missense, nonsense and silent mutations exhibited differences: When expressed as a fraction of total polymorphisms identified, silent mutations prevailed in SM2 while missense and nonsense mutations in gene coding regions were significantly higher in the individuals of the BW stock (Figure 3).

The shared SNPs and indels between SM2 and BW populations have also been investigated (Figure 4). The correlations were found by finding the number of SNPs/indels that matched between each sample pairing and normalizing to the average number of SNPs/indels for each respective subspecies. Each *P. maniculatus* pairing had a correlation between 0.6 and 0.85.

The total correlation between all samples was shown to demonstrate the relationship for both SM2 and BW (Figure 5). In figure 5 dendrograms are used to shows the SM2 and BW samples are distinctly separated with clustered polymorphisms for both SNPs and indels.

Indels, insertions and deletions in BW and SM2

The distribution of indels in relation to coding regions were almost identical in both stocks and was highest in the intergenic and intronic regions, followed by areas upstream and downstream of coding sequences and being minimal in 5’ and 3’ UTR, exonic sequences and splice donor and acceptor sites (Figure 6). Surprisingly, in 5' and 3' UTR no indels were detected in BW and only a small number of indels were detected in SM2, despite that in exon regions indels ranged to the levels of about $12x10^3$ and $4x10^3$ per specimen in SM2 and BW respectively (Figure 7). This is opposed to human UTR regions which contain multiple indels (38). A number of insertions and deletions, ranging from 1 bp upwards to about 500bp were detected in both stocks. Their distribution followed in both SM2 and BW, an exponentially declininag pattern and especially the deletions, exhibited a transient peak, at around 180bp corresponding to 100 incidences per specimen (Figures 8 and 9).

Occurrence of SNPs and Indels Across Individual Chromosomes

A quick glance at each individual chromosome shows there are differing variant rates in each chromosome. A random sample was taken for BW and SM2 and the variant rate in each chromosome was calculated based off the number of variant occurrences and the length of the chromosome (Supplementary Tables 1,2). SNPs had a quicker variant rate than Indels throughout all respective chromosomes. For SNPs, sample 8 (35706) had a lower bound variant rate of 1 variant every 880 base pairs in chromosome 13 and an upper bound variant rate of 1 variant every 116 base pairs in chromosome 17. The indel variant rate in sample 8 had a lower bound of 1 variant every 4309 base pairs in chromosome 13 and an upper bound of 1 variant every 661 base pairs in chromosome 17. The SM2 SNPs in sample 4 (10736) had a lower bound variant rate of 1 variant per 40 base pairs in chromosome 18 and an upper bound variant rate of 1 variant per 103 base pairs in the X chromosome. The indel variant rates in sample 4 ranged from 1 variant per 221 base pairs in chromosome 18 to 1 variant per 513 base pairs in the X chromosome.

Coverage of the genome for SNPs and indels was also shown by the SNPs and indels in chromosomes with the least and the greatest number of polymorphisms for each subspecies. BW samples were represented by sample 8 (Figure 10) and SM2 samples were represented by sample 4 (Figure 11). The peaks shown in figures 10 and 11 are the number of polymorphisms found in 10,000 base pair bins. SM2 samples show greater coverage than BW samples due to greater numbers of polymorphisms.

Missense and nonsense mutations of potential biomedical value

Of note is a roster of specific mutations that were identified indicating the existence of polymorphisms in disease-associated genes. Some of them were seen only in one stock while others in both stocks were assessed.

Missense Mutations

Many of the polymorphisms found in *P. maniculatus* created missense mutations. Some of the representative missense mutations found were in the NFATC2, and HIF1α genes. NFATC2 (nuclear factor of activated T cells) 2, is part of the T cells transcription complex which plays a role in gene transcription during an immune response (39, 40). NFATC2 contained a missense mutation which caused a threonine to alanine substitution (T133A). This predicted change from a non-polar amino acid to a polar amino acid may cause substantial changes in NFATC2 conformation (41). This missense mutation was found in 3 out of 4 SM2 samples and 2 out of 6 BW samples. 2 of the SM2 are homozygous whereas all other samples with this mutation were heterozygous.

HIF1α, hypoxia induced factor 1 alpha subunit, is part of a heterodimeric structure which takes role during the hypoxia response (42). There are two missense mutations within the HIF1α gene, S630A and V662I. These polymorphisms were only seen in SM2 stock, with an allelic frequency of about 0.75, and were validated in 20 additional individuals per stock. Given the role of HIF1a in regulating the response to hypoxia it is plausible that these polymorphisms, if they existed in the original founders of the colony, are related to the high-altitude adaptation of the SM2 animals.

Stop Codons

Along with missense mutations found in *P. maniculatus*, there were also nonsense mutations leading to stop codons causing premature termination of translation. These nonsense mutations could be used to create natural knockout models in the context of a naturally existing wild type population. A few representative nonsense mutations found were in INSIG1, POLQ, and LRP5.

INSIG1, also known as insulin induced gene 1, is an ER protein responsible for regulating cholesterol metabolism, lipogenesis, and glucose homeostasis, mainly through the binding of SREBP cleavage-activating protein (SCAP) (39, 43). The protein is a transmembrane 259 amino acids long with the mutation creating a stop codon on amino acid 190 (R190*). The active site of INSIG1 is Aspartic Acid 187 on the end of the 4th transmembrane domain (44, 45). Although the active site of INSIG1 is before the nonsense mutation, this still creates a truncated protein and changes the conformation. Despite the high prevalence of the mutant allele in our stocks, no homozygous animals were identified implying lethality. This was confirmed in assaying 37 randomly selected *P. maniculatus* individuals in which allelic frequencies exhibited significant deviation from Hardy-Weinberg equilibrium using Fisher's exact test (P = 0.045). 13 *P. maniculatus* had a homozygous wildtype genotype whereas 24 *P. maniculatus* had a heterozygous genotype and no *P. maniculatus* had a homozygous mutant genotype. Furthermore, breeding of heterozygous animals failed to produce homozygous mutant offspring.

POLQ is the gene for DNA Polymerase Theta, a polymerase necessary for microhomology-mediated end joining (MMEJ) (39, 46). A stop codon was identified at amino acid 330 (R330*) out of 2550 amino acids. This was seen in 2 heterozygous SM2 and appeared homozygous wildtype in all other samples.

LRP5, low-density lipoprotein receptor-related protein 5, plays a role in affecting bone mass accrual during development and skeletal homeostasis (39, 47). A stop codon was identified at E273*. This was only seen in SM2 *Peromyscus*, all of which were heterozygous, and homozygous wildtype in all BW *Peromyscus* samples.

Discussion

In the present study we report a comprehensive roster of polymorphisms detected in two populations of *P. maniculatus* that are maintained in the PGSC. The analysis covered 2.5X10^9 bp of the *P. maniculatus* genome and revealed the presence of about 17.5X10^6 SNPs and 2.1X10^6 indels for BW stock, against the publicly available genome assembly. Variation was about 4 times higher in the animals of the SM2 stock. This high density of polymorphisms, especially if individuals of the two stocks interbreed, provides tremendous genetic power in mapping loci of interest.

The higher polymorphism count in SM2 is likely due to the genetic divergence of the two populations and the fact that SM2 samples were aligned to the BW genome, due to a lack of an SM2 reference genome. Nevertheless, the fact that the range of variation within animals of the same stock was higher for the SM2, despite only 4 SM2 as opposed to 6 BW individuals were sequenced, may suggest that SM2 stock could have higher allelic diversity than BW, but a larger sample size would be needed to support this argument. Heterozygosity was lower though in the SM2, implying lower intrapopulation diversity as compared to BW which can be due to the different history of the stocks in our facilities: Both stocks were originally established by similar methods, 40-50 wild caught animals and this discrepancy between BW and SM2 stocks may reflect the diversity of the original founders. In addition, BW are utilized at a higher degree than the SM2 and the BW colony was established about 40 years earlier. Therefore BW are more actively bred at the PGSC which may occasionally result in a series of bottlenecks that compromises their diversity as recorded today. Regardless of SM2 having higher occurrences of variants than BW, both BW and SM2 have full coverage of variants across each chromosome. These rates of genetic variation are comparable in magnitude to those reported for conventional laboratory mice (Mus) at which 71 X 10^6 SNPs and 12 X 10^6 M indels have been identified across 13 inbred mouse strains (48)

In comparing the distribution of the polymorphisms in coding regions between the stocks, a noteworthy observation was made, related to the bias seen in SM2 for the type of mutations identified: In the SM2 stock, the fraction of missense and nonsense polymorphisms expressed as a proportion of the total SNPs identified, was lower than that of BW, while synonymous polymorphisms were more common in SM2. The small population size, the differences in the breeding programs and history of the two stocks in captivity due to differences in the demand of the stocks, and the original extraction of variation data by alignment of the SM2 individuals to BW genome, highly restricts the extraction of evolutionarily relevant conclusions.

In addition to its value in characterizing the genomic architecture of these commonly used stocks of *P. maniculatus*, the present analyses also revealed a roster of loss of function alleles and mutant alleles that could be used to generate animals with desired genotypes in the context of a natural population. Those included in example a truncated form of the cholesterol biosynthesis regulator INSIG1, and HIF1a polymorphisms. In the case of INSIG1 it has been shown that INSIG1/INSIG2 knockouts in *Mus musculus* still lead to viable offspring that produce much higher levels of cholesterol as a result (49). In *P. maniculatus* though, truncation of INSIG1 in homozygosity apparently leads to lethality. Whether this is related to *Peromyscus* physiology, or toxicity of the truncated INSIG1 allele remains to be established. The high frequency though of this allele in heterozygosity implies some advantages in the individuals that cause its stabilization in the population. HIF1a variation may also be of special value since they were seen, in high frequency in the SM2 animals only. Whether this polymorphism possesses evolutionary relevance remains to be established since the animals were bred for several generations in captivity and this SNP may be a de novo mutation that emerged in our colony.

Besides its value in describing the landscape of diversity in captive *Peromyscus* stocks and in pointing to specific, functional polymorphisms of potential biological value, this resource has an additional significance: All individual animals at the PGSC, including those supplied to investigators worldwide, are pedigreed and can be traced back to their original ancestors. Thus, this resource provides abundant baseline genetic data that refer to a reference, genetically diverse population. This in turn greatly facilitates both retrospective analyses of specimens derived in the past and studies that can be implemented in the future using animals with comparable genetic make-up. Finally, even though the genetic variation data reported here do not accurately reflect the variation of *P. maniculatus* in the wild, the present results may be of use to investigators addressing the genomic diversity of wild-caught *P. maniculatus*.

Declarations

Ethics approval and consent to participate: Studies were approved by IACUC of UofSC. Protocol no., 2356-101506-042720. All methods were carried out in accordance with relevant guidelines and regulations.

Consent for publication: Not applicable

Availability of data and materials: Animals are available from the *Peromyscus* Genetic Stock Center located at the University of South Carolina. DNA sequences have been deposited to the European Variation Archive (EVA) under the project ID PRJEB41333 (https://www.ebi.ac.uk/ena/browser/view/PRJEB41333?show=analyses).

Competing interests: None

Funding: This study was supported by NSF (OIA-1736150). The PGSC is supported by a grant from NSF (DBI-1755670).

Authors' contributions: ML, MS and HK wrote the paper; DA JH and ML performed experiments; ML, RD, BT, VS, MS and HK analyzed data; AH, VK,YZ, and AG, curated and provided specimens; MS, and HK supervised studies

Acknowledgements: N/A

Literature Cited

1. Crossland, J. and A. Lewandowski. 2006. *Peromyscus* - A fascinating laboratory animal model. Techtalk, 11:1-2.
2. Havighorst, A., Crossland, J., & Kiaris, H. (2017, January). *Peromyscus* as a model of human disease. In Seminars in cell & developmental biology (Vol. 61, pp. 150-155). Academic Press.
3. Bedford, N. L., & Hoekstra, H. E. (2015). The natural history of model organisms: *Peromyscus* mice as a model for studying natural variation. Elife, 4, e06813.
4. Luong LT, Vigliotti BA, Campbell S, Comer JA, Mills JN, Hudson PJ. Dynamics of hantavirus infection in Peromyscus leucopus of central Pennsylvania. Vector Borne Zoonotic Dis. 2011;11(11):1459-1464. doi:10.1089/vbz.2010.0255
5. Botten J, Mirowsky K, Kusewitt D, Bharadwaj M, Yee J, Ricci R, Feddersen RM, Hjelle B. Experimental infection model for Sin Nombre hantavirus in the deer mouse (*Peromyscus maniculatus*). Proc Natl Acad Sci U S A. 2000 Sep 12;97(19):10578-83. doi: 10.1073/pnas.180197197. PMID: 10973478; PMCID: PMC27067.
6. Bunikis J, Tsao J, Luke CJ, Luna MG, Fish D, Barbour AG. Borrelia burgdorferi infection in a natural population of Peromyscus Leucopus mice: a longitudinal study in an area where Lyme Borreliosis is highly endemic. J Infect Dis. 2004 Apr 15;189(8):1515-23. doi: 10.1086/382594. Epub 2004 Mar 30. PMID:
7. Barbour AG, Bunikis J, Fish D, Hanincová K. Association between body size and reservoir competence of mammals bearing Borrelia burgdorferi at an endemic

site in the northeastern United States. Parasit Vectors. 2015;8:299. Published 2015 May 30. doi:10.1186/s13071-015-0903-5

8. Bryan D. Griffin, Mable Chan, Nikesh Tailor, Emelissa J. Mendoza, Anders Leung, Bryce M. Warner, Ana T. Duggan, Estella Moffat, Shihua He, Lauren Garnett, Kaylie N. Tran, Logan Banadyga, Alixandra Albietz, Kevin Tierney, Jonathan Audet, Alexander Bello, Robert Vendramelli, Amrit S. Boese, Lisa Fernando, L. Robbin Lindsay, Claire M. Jardine, Heidi Wood, Guillaume Poliquin, James E. Strong, Michael Drebot, David Safronetz, Carissa Embury-Hyatt, Darwyn Kobasa. North American deer mice are susceptible to SARS-CoV-2. bioRxiv 2020.07.25.221291; doi: https://doi.org/10.1101/2020.07.25.221291
9. Anna Fagre, Juliette Lewis, Miles Eckley, Shijun Zhan, Savannah M Rocha, Nicole R Sexton, Bradly Burke, Brian Geiss, Olve Peersen, Rebekah Kading, Joel Rovnak, Gregory D Ebel, Ronald B Tjalkens, Tawfik Aboellail, Tony Schountz. SARS-CoV-2 infection, neuropathogenesis and transmission among deer mice: Implications for reverse zoonosis to New World rodents. bioRxiv 2020.08.07.241810; doi: https://doi.org/10.1101/2020.08.07.241810
10. Borniger JC, Nelson RJ. Photoperiodic regulation of behavior: Peromyscus as a model system. Semin Cell Dev Biol. 2017 Jan;61:82-91. doi: 10.1016/j.semcdb.2016.06.015. Epub 2016 Jun 23. PMID: 27346738.
11. Steinman MQ, Trainor BC. Sex differences in the effects of social defeat on brain and behavior in the California mouse: Insights from a monogamous rodent. Semin Cell Dev Biol. 2017 Jan;61:92-98. doi: 10.1016/j.semcdb.2016.06.021. Epub 2016 Jun 30. PMID: 27375045; PMCID: PMC5201444.

12. Schweizer RM, Velotta JP, Ivy CM, Jones MR, Muir SM, Bradburd GS, Storz JF, Scott GR, Cheviron ZA. Physiological and genomic evidence that selection on the transcription factor Epas1 has altered cardiovascular function in high-altitude deer mice. PLoS Genet. 2019 Nov 7;15(11):e1008420. doi: 10.1371/journal.pgen.1008420. PMID: 31697676; PMCID: PMC6837288.
13. Natarajan C, Inoguchi N, Weber RE, Fago A, Moriyama H, Storz JF. Epistasis among adaptive mutations in deer mouse hemoglobin. Science. 2013 Jun 14;340(6138):1324-7. doi: 10.1126/science.1236862. PMID: 23766324; PMCID: PMC4409680.
14. Havighorst A, Zhang Y, Farmaki E, Kaza V, Chatzistamou I, Kiaris H. Differential regulation of the unfolded protein response in outbred deer mice and susceptibility to metabolic disease. Dis Model Mech. 2019;12(2):dmm037242. Published 2019 Feb 27. doi:10.1242/dmm.037242.
15. Bendesky, A., Kwon, Y. M., Lassance, J. M., Lewarch, C. L., Yao, S., Peterson, B. K., ... & Hoekstra, H. E. (2017). The genetic basis of parental care evolution in monogamous mice. Nature, 544(7651), 434-439.
16. Vrana PB, Shorter KR, Szalai G, Felder MR, Crossland JP, Veres M, Allen JE, Wiley CD, Duselis AR, Dewey MJ, Dawson WD. Peromyscus (deer mice) as developmental models. Wiley Interdisciplinary Reviews: Developmental Biology. 2014 May 1;3(3):211-30.
17. Kowalczyk AS, Davila RF, Trainor BC. Effects of social defeat on paternal behavior and pair bonding behavior in male California mice (Peromyscus californicus). *Horm Behav*. 2018;98:88-95. doi:10.1016/j.yhbeh.2017.12.010

18. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloğlu A, Ozen S, Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A. 2009 Nov 10;106(45):19096-101. doi: 10.1073/pnas.0910672106. Epub 2009 Oct 27. PMID: 19861545; PMCID: PMC2768590.

19. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009 Sep 10;461(7261):272-6. doi: 10.1038/nature08250. Epub 2009 Aug 16. PMID: 19684571; PMCID: PMC2844771.

20. Kenney-Hunt, J., Lewandowski, A., Glenn, T. C., Glenn, J. L., Tsyusko, O. V., O'Neill, R. J., ... & Shorter, K. R. (2014). A genetic map of *Peromyscus* with chromosomal assignment of linkage groups (a *Peromyscus* genetic map). Mammalian genome, 25(3-4), 160-179.

21. Long AD, Baldwin-Brown J, Tao Y, Cook VJ, Balderrama-Gutierrez G, Corbett-Detig R, Mortazavi A, Barbour AG. The genome of *Peromyscus leucopus*, natural host for Lyme disease and other emerging infections. Sci Adv. 2019 Jul 24;5(7):eaaw6441. doi: 10.1126/sciadv.aaw6441. PMID: 31355335; PMCID: PMC6656541.

22. Colella, J. P., Tigano, A., & MacManes, M. D. (2020). A linked-read approach to museomics: Higher quality de novo genome assemblies from degraded tissues. *Molecular ecology resources*, *20*(4), 856-870.

23. Vij, S., Kuhl, H., Kuznetsova, I. S., Komissarov, A., Yurchenko, A. A., Van Heusden, P., ... & Saju, J. M. (2016). Chromosomal-level assembly of the Asian seabass genome using long sequence reads and multi-layered scaffolding. PLoS genetics, 12(4).
24. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., ... & Gocayne, J. D. (2001). The sequence of the human genome. science, 291(5507), 1304-1351.
25. Damas, J., O'connor, R., Farr√©, M., Lenis, V. P. E., Martell, H. J., Mandawala, A., ... & Larkin, D. M. (2017). Upgrading short-read animal genome assemblies to chromosome level using comparative genomics and a universal probe set. Genome research, 27(5), 875-884.
26. Brown, J., Crivello, J., & O'Neill, R. J. (2018). An updated genetic map of *Peromyscus* with chromosomal assignment of linkage groups. Mammalian Genome, 29(5-6), 344-352.
27. Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M. R., Armean, I. M., ... & Cummins, C. (2019). Ensembl 2019. Nucleic acids research, 47(D1), D745-D751.
28. Assembly [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 1988–2020. Accession No. GCA_003704035.1, HU_Pman_2.1; 2021 Jan 14. Available from: https://www.ncbi.nlm.nih.gov/assembly/GCA_003704035.1/

29. Huptas, C., Scherer, S., & Wenning, M. (2016). Optimized Illumina PCR-free library preparation for bacterial whole genome sequencing and analysis of factors influencing de novo assembly. *BMC research notes*, *9*(1), 269.

30. Harrington, C. A., Winther, M., & Garred, M. M. (2009). Use of bioanalyzer electropherograms for quality control and target evaluation in microarray expression profiling studies of ocular tissues. *Journal of ocular biology, diseases, and informatics*, *2*(4), 243-249.

31. Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., ... & Banks, E. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics, 43(1), 11-10.

32. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.

33. Broad Institute. (Accessed: 2018/02/21; version 2.17.8). "Picard Tools." Broad Institute, GitHub repository. http://broadinstitute.github.io/picard/

34. Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., ... & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly, 6(2), 80-92.

35. Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. BMC bioinformatics, 15(1), 1-13.

36. Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data.

37. Korneliussen, T. S., Moltke, I., Albrechtsen, A., & Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. BMC bioinformatics, 14(1), 1-14.

38. Clark, T. G., Andrew, T., Cooper, G. M., Margulies, E. H., Mullikin, J. C., & Balding, D. J. (2007). Functional constraint and small insertions and deletions in the ENCODE regions of the human genome. *Genome biology*, *8*(9), R180.

39. Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., ... & Sirota-Madi, A. (2010). GeneCards Version 3: the human gene integrator. *Database*, *2010*.

40. Peng, S. L., Gerth, A. J., Ranger, A. M., & Glimcher, L. H. (2001). NFATc1 and NFATc2 together control both T and B cell activation and differentiation. Immunity, 14(1), 13-20.

41. Broome, B. M., & Hecht, M. H. (2000). Nature disfavors sequences of alternating polar and non-polar amino acids: implications for amyloidogenesis. Journal of molecular biology, 296(4), 961-968.

42. Weidemann, A., & Johnson, R. S. (2008). Biology of HIF-1 α. Cell Death & Differentiation, 15(4), 621-627.

43. Yang, T., Espenshade, P. J., Wright, M. E., Yabe, D., Gong, Y., Aebersold, R., ... & Brown, M. S. (2002). Crucial step in cholesterol homeostasis: sterols promote

binding of SCAP to INSIG-1, a membrane protein that facilitates retention of SREBPs in ER. Cell, 110(4), 489-500.

44. Dong, X. Y., & Tang, S. Q. (2010). Insulin-induced gene: a new regulator in lipid metabolism. Peptides, 31(11), 2145-2150.

45. Feramisco, J. D., Goldstein, J. L., & Brown, M. S. (2004). Membrane topology of human insig-1, a protein regulator of lipid synthesis. Journal of Biological Chemistry, 279(9), 8487-8496.

46. Brambati, A., Barry, R. M., & Sfeir, A. (2020). DNA polymerase theta (Polθ)–an error-prone polymerase necessary for genome stability. Current Opinion in Genetics & Development, 60, 119-126.

47. Gong, Y., Slee, R. B., Fukai, N., Rawadi, G., Roman-Roman, S., Reginato, A. M., ... & Zacharin, M. (2001). LDL receptor-related protein 5 (LRP5) affects bone accrual and eye development. Cell, 107(4), 513-523.

48. Doran, A.G., Wong, K., Flint, J. et al. Deep genome sequencing and variation analysis of 13 inbred mouse strains defines candidate phenotypic alleles, private variation and homozygous truncating mutations. Genome Biol 17, 167 (2016). https://doi.org/10.1186/s13059-016-1024-y

49. Engelking, L. J., Liang, G., Hammer, R. E., Takaishi, K., Kuriyama, H., Evers, B. M., ... & Brown, M. S. (2005). Schoenheimer effect explained–feedback regulation of cholesterol synthesis in mice mediated by Insig proteins. The Journal of clinical investigation, 115(9), 2489-2498.

50. Hammond, K. A., Roth, J., Janes, D. N., & Dohm, M. R. (1999). Morphological and physiological responses to altitude in deer mice Peromyscus maniculatus. *Physiological and Biochemical Zoology*, *72*(5), 613-622.

Figure Legends

Figure 1 Comparison of SNP and Indel counts found in BW and SM2 Counts of SNPs and Indels (**a**) were taken for each sample. Each sample is listed in the order of their ID at the PGSC. Each count is measured in millions. SM2 samples are marked by the color blue while BW samples are marked by the color red. In **b** the SNP counts of SM2 and BW are organized into boxplots to compare the range of SNP counts for both SM2 and BW. The counts are measured in millions where SM2 SNP counts have a range of 3 million and BW SNP counts have a range of 0.5 million.

Figure 2 Heterozygosity of *Peromyscus maniculatus*. BW show higher heterozygosity than SM2. The missense and nonsense polymorphisms (**a**) have an average heterozygosity of 0.0013 with a standard deviation of $8.48x10^{-5}$ in SM2 and an average heterozygosity of 0.0019 with a standard deviation of $5.29x10^{-5}$ in BW. The synonymous polymorphisms have a higher heterozygosity than the missense and nonsense polymorphisms. The synonymous polymorphisms (**b**) have an average heterozygosity of 0.0023 with a standard deviation of 0.0003 in SM2 and an average heterozygosity of 0.0045 with a standard deviation of 0.0007 in BW.

Figure 3 *Peromyscus maniculatus* SNP functional class. In (**a)** the percentage of the functional class of all SNPs in each sample is shown. In (**b)** the difference between

the percentage of each SNP functional class for BW and SM2 samples are shown. SM2 samples are shown in blue whereas BW samples are shown in red.

Figure 4 *Peromyscus maniculatus* Interspecies Polymorphism Correlation. In the upper and lower panels the correlation of polymorphisms between each SM2 samples and BW samples are shown. SNPs and Indel correlations are shown in left and right respectively. The sample numbers are found on the upper and left border of each table with the correlation percentage in the cross between two samples. The range of the correlation for each sample pair ranged from 0.6 to 0.85. Each correlation was normalized to the mean number of respective polymorphisms across all samples of the respective subspecies.

Figure 5 *Peromyscus maniculatus* Polymorphism Clustering. The matching polymorphisms were taken between each sample and clustered to show the relationship between the SM2 and BW *Peromyscus*. The top two clusters of each dendrogram show a separation between SM2 and BW based on both SNPs (**a**) and indels (**b**). Each sample ID is shown with their color correlating with the subspecies (BW: red, SM2: blue). The y-axis shows the relative distance between each sample.

Figure 6 Gene regions where SM2 and BW indels are found. The percentage of the gene regions where indels were found is shown here. Even though SM2 have significantly higher counts of indels than BW, the percentage of indels found in a region

are identical. The samples shown were randomly chosen to represent SM2 and BW samples. Sample 1 (10683) is an SM2 and sample 5 (35060) is a BW.

Figure 7 Indels found in each gene region across all samples. Each gene region is separated to show indel count in SM2 (blue) and BW (red) samples. A majority of indels are found in intergenic and intronic regions.

Figure 8 Count of insertions in all samples. The graphs here show the count of insertions in each sample as related to their length. Samples 1-4 are SM2 and samples 5-10 are BW. Insertions in the SM2 samples reach up to at least 450 bp whereas insertions in the BW samples reach up to at least 250 bp.

Figure 9 Count of deletions in all samples. The graphs here show the count of deletions in each sample as related to their length. Samples 1-4 are SM2 and samples 5-10 are BW. Deletions in the SM2 samples reach up to at least 250 bp whereas insertions in the BW samples reach up to at least 200 bp.

Figure 10 Variant Coverage in a BW Peromyscus. Sample 8 (35706) was used as a random example BW to show the range of coverage of SNPs and Indels in an individual chromosome. Chromosome 13 had the least amount of SNPs and Indels (Left) whereas chromosome 17 had the greatest amount of SNPs and Indels (Right). Variants were counted in bins of 10,000 bp.

Figure 11 Variant Coverage in an SM2 Peromyscus. Sample 4 (35706) was used as a random example SM2 to show the upper and lower bounds of coverage of SNPs and Indels in an individual chromosome for SM2 Peromyscus. The X chromosome had the least amount of SNPs and Indels (Left) whereas chromosome 18 had the greatest amount of SNPs and Indels (Right). Variants were counted in bins of 10,000 bp.