8-16-2019 1:30 PM

Cross-species utility of the Mouse Diversity Genotyping Array in assaying single nucleotide polymorphisms

Rachel Kelly, *The University of Western Ontario*

Supervisor: Hill, Kathleen A., *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Biology

Recommended Citation

Kelly, Rachel, "Cross-species utility of the Mouse Diversity Genotyping Array in assaying single nucleotide polymorphisms" (2019). *Electronic Thesis and Dissertation Repository*. 6368.
https://ir.lib.uwo.ca/etd/6368

Abstract

In the study of genetic diversity in non-model species there is a notable lack of the low-cost, high resolution tools that are readily available for model organisms. Genotyping microarray technology for model organisms is well-developed, affordable, and potentially adaptable for cross-species hybridization. The Mouse Diversity Genotyping Array (MDGA), a single nucleotide polymorphism (SNP) genotyping tool designed for *M. musculus*, was tested as a tool to survey genomic diversity of wild species for inter-order, inter-family, inter-genus, and intra-genus comparisons. Application of the MDGA cross-species provides genetic distance information that reflects known taxonomic relationships reported previously between non-model species, but there is an underestimation of genetic diversity for non-Mus samples. The number and types of samples included in sets genotyped together must be considered in cross-species hybridization. The number of loci with heterozygous genotypes mapped to published genome sequences indicates potential for cross-species MDGA utility.

Keywords: Mouse Diversity Genotyping, MDGA, cross-species genotyping, cross-species, genetic diversity, single nucleotide polymorphism.

Summary for Lay Audience

There is a need for a tool that can assay DNA sequence differences in species that are understudied and for which there is little or no DNA sequence information available. One method of analyzing differences in DNA sequences in species with well-understood genomes is through a genotyping microarray, a technology with demonstrated utility cross-species. This tool is capable of examining the DNA sequence information at hundreds of thousands of sites across the genomes of well-studied organisms in a single assay. The Mouse Diversity Genotyping Array (MDGA) is a tool that was designed to examine known differences at 493,290 sites across the genome of the house mouse, *Mus musculus*. Given that the MDGA was designed for the house mouse, and that closely-related organisms share genetic similarity, the MDGA was tested for utility in identifying genome variation in other wild (feral) mice and rodents. The MDGA was tested on 44 DNA samples from inbred laboratory mice and wild species that last shared a common ancestor millions of years ago. Variation identified from more distantly-related species that were not of the same genus as the laboratory house mouse was an underestimate of the true amount of variation present in the genome of wild species. The utility of the MDGA for use with DNA from wild species is best suited to mice from the same genus as the house mouse. Identifying changes in genetic variation within populations of wild rodents can help researchers understand the links between specific genome changes and the ability to adapt to pressures in the environment, as well as better understand the evolution of rodents. The MDGA is a cost-effective tool for rapidly identifying genetic variation in wild rodent species until the cost of sequencing the genomes of understudied species is reduced.

Co-authorship Statement

Rachel Kelly completed this work under the supervision and financial support of Dr. Kathleen Allen Hill. Rachel completed the work presented in the thesis with assistance in genotyping five sample sets in this study as well as assistance with the R scripts required to generate distance matrices, SNP trees, and rainfall plots. Maja Milojevic assisted in genotyping the inter-order, inter-family, inter-genus, and intra-genus test sets as well as the four naked mole rats utilized in a case study using the Affymetrix Power Tools software. Dr. Kathleen Hill assisted in running Fisher's Exact tests of significance using Cytel software. Marjorie E. Osborne Locke created the R code used to generate SNP-based genetic distance matrices and SNP phylogenetic trees of relatedness. Functional association analysis of shared SNPs between model and non-model organisms was made possible by a list of Ensembl gene IDs associated with Mouse Diversity Genotyping Array probes provided by Freda W. Qi.

Acknowledgements

I would first like to express my deep gratitude to my supervisor Dr. Kathleen Hill, whose guidance and encouragement helped me grow and learn as a researcher within her lab. Her support and patience in teaching has been instrumental throughout each process of my graduate research. Her creativity in approaching research problems and attention to detail are skills I hope to carry with me in the future. I would also like to thank my advisors Dr. Jamie Kramer and Dr. Anthony Percival-Smith for their advice and constructive feedback on my research during this project's journey.

This research was funded by a Natural Sciences and Engineering Research Council of Canada Discovery Grant awarded to Dr. Kathleen Hill. I would like to thank the Western School of Graduate and Postdoctoral Studies for the Western Graduate Research Scholarship that helped support my graduate studies. I would also like to thank the Department of Biology and the Environmental Mutagenesis and Genomics Society for travel awards that allowed me the unique opportunity to present my research at the EMGS 49th Annual Meeting in San Antonio, Texas.

I would like to thank a number of people for their expert input and assistance throughout my research project. Thank you to Dr. Bin Luo, Dr. Reg Kulperger, and Dr. Camila de Souza for their technical and statistical expertise. Thank you to Dr. Lucien Ilie and his lab team for teaching and assistance in E-MEM analyses.

A huge thank you to my family, who have always supported me following my passion of developing a career in science and are my biggest cheerleaders. A special thank you to my lab family and friends for their help and support, and for all the laughs we've shared together.

Table of Contents

List of Figures

List of Tables

List of Appendices

List of Abbreviations

BLAST	Basic Local Alignment Search Tool
bp	base pair; kb (kilobase); Mb (Megabase)
BRLMM-P	Bayesian Robust Linear Model with Mahalanobis distance classifier – perfect match
DAVID	Database for Annotation, Visualization, and Integrated Discovery
DNA	deoxyribonucleic acid
E-MEM	efficient computation of maximal exact matches
FTP	File transfer protocol
ID	identifier
KEGG	Kyoto Encyclopedia of Genes and Genomes
MDGA	Mouse Diversity Genotyping Array
MGI	Mouse Genome Informatics
MRCA	most recent common ancestor
MUGA	Mouse Universal Genotyping Array
MYD	millions of years divergence
NCBI	National Center for Biotechnology Information
nt	nucleotide
PASA	PCR amplification of specific alleles
PCR	polymerase chain reaction
RADseq	restriction site associated DNA sequencing
RFLP	Restriction Fragment Length Polymorphism
SNP	single nucleotide polymorphism
SNPSTeP	single nucleotide polymorphism spatial-temporal plot
USD	United States Dollar

1 Introduction

1.1 Non-model organisms lack tools to survey genomic diversity

There is a lack of knowledge and resources for population geneticists to use in assaying and characterizing genetic diversity genome-wide within non-model species, or species that are not traditionally used in genetic research (DeMay et al., 2017; Grant and Grant, 2002; Razgour et al., 2019). There is a bias for the study of the human genome and analytical methods to study human diversity (Lander et al., 2001; Sherry et al., 2001). Given challenges in the direct study of samples from humans, there is a historical reliance on model organisms that act as a proxy for the human genome (Keane et al., 2011; Zhao et al., 2004). In sum, there is a lack of genomic sequence information available for non-model species and a lack of tools to assay genomic diversity in understudied organisms (Hoffman et al., 2013; More et al., 2019; Ogden et al., 2012). Custom tools for assaying genomic diversity are needed, but the creation of these tools is time-consuming and expensive. There is a need to explore existing technologies designed for use with human and model species and the effectiveness of the existing technologies for cross-species application.

1.2 Model organisms are convenient proxies but remain approximations

Model organisms are species that are chosen to act as a proxy for a system that is more complex and more challenging to study. A few key benefits to using model organisms in genetics research include ease of breeding and maintaining the species in captivity, short generation times, ability to mimic the effects of human disease, and perhaps most importantly, the ability to manipulate the genomes of model organisms with greater ease (Aditi et al., 2016; Kuperwasser et al., 2005; Mungall et al., 2015; Styczyńska-Soczka et al., 2017; Zeef et al., 2012). The ability to genetically manipulate model organisms is aided by the vast wealth of genomic information available for these species. The information available includes fully sequenced genomes, and annotations on the location and effect of genomic variation (Eppig et al., 2015; Millburn et al., 2016; Shimoyama et al., 2016; Zhu et al., 2015). Despite the clear benefits of using model organisms to understand the links between genotype and the resulting phenotype, for certain research objectives model organisms often do not represent the vast genetic diversity of related wild or feral organisms.

1.3 There is untapped genetic research potential in non-model organisms

There are numerous species that are currently not considered model organisms but represent untapped avenues of research regarding the effects of the nature, utility, and impact of genetic diversity. Wild species are typically non-model organisms that could become useful models in the context of human health if greater genomic information was available. One example is the elephant, an interesting potential model of cancer resistance in a large mammal with a long lifespan (Abegglen et al., 2015). Population genetic studies of wild species would also benefit from a greater range of organisms that have a fully sequenced genome with gene annotations available (Harris et al., 2013; Montana et al., 2017). Analyzing genomes of non-model species can help elucidate more precise divergence times and landmark events in the evolution of mammals (Bennett et al., 2017; Li et al., 1990). Non-model organisms living in close proximity to humans can act as environmental sentinels, providing data on genomic changes caused by environmental mutagens (Rodríguez-Estival and Smits, 2016). A key motivation for this study is the immediate opportunity and need for tools to assay genetic diversity in wild rodent species (Harris et al., 2013; Rodríguez-Estival and Smits, 2016).

1.4 Genome variation tells a story of past, present, and future

Genome variation, or differences in the DNA sequence between two or more organisms of interest, can inform researchers about the health of an organism. Understanding the links between differences in genotype and phenotype is paramount in determining the genetic root of disease aetiology (Lander et al., 2001). Deductions of the genetic cause of current phenotypic states and estimations of future disease risk can be identified by studying genetic variation of organisms (Wray et al., 2007). The ability to monitor genetic variation of a population allows a new degree of information to be gleaned from species of interest. Allelic differences between populations of the same species that separated geographically over time can be used to understand the effects of environmental pressures on the genomes of organisms (Coop et al., 2009; Natarajan et al., 2015). It is also possible to track the effects of environmental mutagens within the genomes of individuals in a population over time (Bickham et al., 2000; Štambuk et al., 2013). Alleles at proximal loci in the genome that are inherited together are known as haplotypes. Haplotypes can be used to track the evolutionary history of a species (Johnson et al., 1998; Vonholdt et al., 2010).

1.5 Mammalian comparative genomics offers benefits to humans

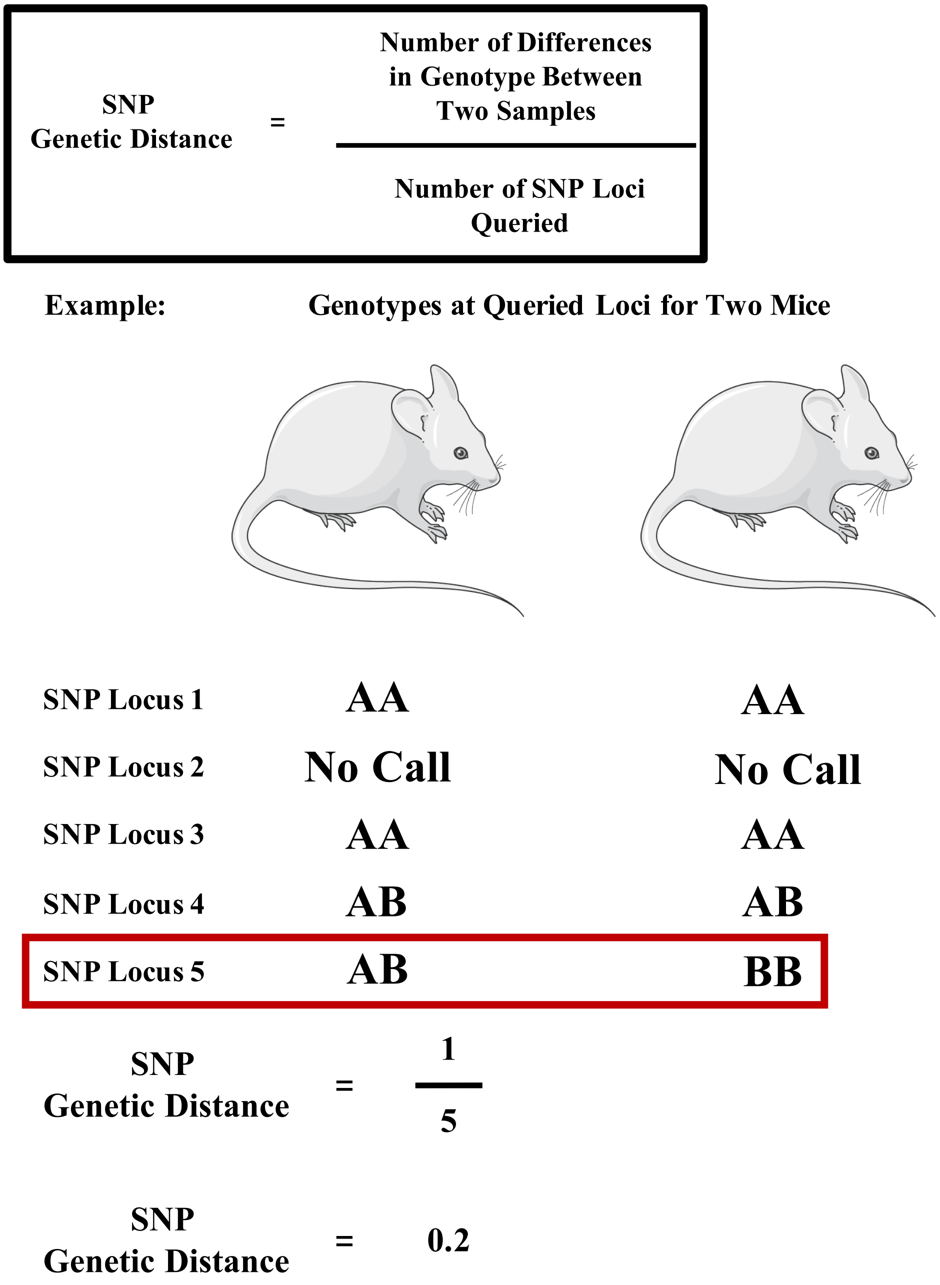
The primary benefit to humans of mammalian comparative genomics is that humans are members of the class Mammalia and share distinctive developmental characteristics with other mammals that other classes of organisms do not experience. The genes that humans and other mammals inherited from a common ancestor are known as homologs, and homologous genes are potential new targets for disease research and evolutionary studies (O'Brien et al., 1999). Mammalian comparative genomics can aid in mapping the location of genes of different species and in identifying syntenic regions. Syntenic regions of the genome between two or more species have a similar inherited linkage of genes due to common ancestry (Waterston et al., 2002). Perhaps the most significant comparative genomic study of its time was the comparison of the mouse genome to the human genome after sequencing. Through comparison of the human and mouse reference genomes, a large amount of synteny between human and mouse genomes that make mice tractable for human genetic studies was discovered (Waterston et al., 2002). There are also key biological similarities between humans and other mammals including reproductive and developmental pathways that are not shared between humans and non-mammalian species (Luis Villanueva-Cañas et al., 2017). In a key comparative genomic study published this year, researchers identified a genetic basis underlying the evolution of inner ear development in mammals (Pisciottano et al., 2019).

1.6 Single nucleotide polymorphisms are targets for comparative genomic analysis

Single nucleotide polymorphisms, commonly referred to as SNPs, are single base positions in the genome that are variable in genotype for individuals in a population. The minor, or less common, variant of a SNP allele must be present in at least 1% of a population (Wang et al., 1998). SNPs are the most abundant type of variation in the genome, making it an excellent target for comparative genomic studies (Marth et al., 1999; Wang et al., 1998). Hundreds of thousands of SNPs can be conveniently assayed concurrently across the genome of a model organism with the advent of genotyping microarrays (Gunderson et al., 2005; LaFramboise, 2009). According to a 2018 price quotation from ThermoFisher Scientific (Applied Biosystems), the average price of purchasing and using a mouse genotyping array that assays hundreds of thousands of key SNPs is approximately $600 USD per sample for older array models like the Mouse Diversity Genotyping Array (Yang et al., 2009). Newer array models are even more cost effective, with a price of about $75.5 USD per sample. The cost of sequencing a whole mouse genome as of February 2019 was approximately $1,300 USD per sample, making the sequencing option approximately 17 times more expensive than the latest SNP genotyping technologies (Sivashankari and Shanmughavel 2007; Wetterstrand K 2019). Sequencing is cost prohibitive for population studies and large sample sets. Using SNPs for comparative genomics provides a large amount of genomic information in one application of the genotyping array, and the associated bioinformatics analysis is simpler and faster compared to traditional next generation sequencing methods. If the genome sequence and SNP genotypes are known, custom genotyping arrays can be created to assay specific SNP loci of interest (Keating et al., 2008; Voight et al., 2012).

SNPs are a useful type of genome variation to target for comparative genomics because SNP loci are numerous and widespread in the genome. Trees that reflect the relative divergence times of species studied can be made through different types of genomic information, but are referred to as SNP trees when generated from values known as SNP-based genetic distances (Coll et al., 2014; Libiger et al., 2009; Locke et al., 2015). SNP-based genetic distances are calculated by dividing the total number of SNP loci that have different genotypes between two organisms by the total number of SNP loci that are queried (Figure 1.6.1). A minimum SNP-based genetic distance value of zero reflects that at the loci queried, the two organisms have identical genotypes. can be A maximum SNP-based genetic distance value of one reflects that at the loci queried, the two organisms have different genotypes at every locus (Locke et al., 2015). Using a neighbour-joining method of clustering samples with smaller SNP-based genetic distance comparisons between them, SNP trees reflecting genetic relatedness can be constructed (Saitou and Nei, 1987). Assessing SNP variation is informative for phylogenetic, evolutionary, and population genetic studies (Libiger et al., 2009; Locke et al., 2015; McCue et al., 2012).

SNPs genotypes are informative when examined in the context of spatial position across the genome of the species analyzed. Spatial analysis of SNP genotypes can be used to distinguish populations of species from one another (Lah et al., 2016). SNP loci can be classified according to how the SNP genotypes change within a population for a particular locus (Hannigan et al., 2017; Morin et al., 2004). SNP loci that are variable in genotype within a population or SNP loci that are invariant in genotype for all individuals in a population can be visualized across the genome of the model species to identify trends in conservation and spatial position. To accomplish this, rainfall plots can be adapted for use to display the chromosomal distribution of SNP genotypes (Figure 1.6.2; Domanska et al., 2017; Nik-Zainal et al., 2012). Rainfall plots are scatterplots generated where genomic position is the x-axis and inter-SNP locus distance is the y-axis. Each



$$\text{SNP Genetic Distance} = \frac{\text{Number of Differences in Genotype Between Two Samples}}{\text{Number of SNP Loci Queried}}$$

Example: **Genotypes at Queried Loci for Two Mice**

SNP Locus 1	**AA**	**AA**
SNP Locus 2	**No Call**	**No Call**
SNP Locus 3	**AA**	**AA**
SNP Locus 4	**AB**	**AB**
SNP Locus 5	**AB**	**BB**

$$\text{SNP Genetic Distance} = \frac{1}{5}$$

$$\text{SNP Genetic Distance} = 0.2$$

Figure 1.6.1 SNP genetic distance values based on genotype differences between individuals reflect genetic relatedness
Single nucleotide genetic distances are calculated by dividing the number of genotypic differences between two individuals by the total number of loci queried. In the example, five loci are queried in two mice with a single difference in genotype between the two mice highlighted in a red box. The genetic distance derived from SNP genotypes between them is 0.2. A SNP-based genetic distance value of 0 indicates the individuals compared are genetically identical, and a SNP-based genetic distance value of 1 indicates that the individuals compared are genetically dissimilar. This determination was made for the 493,290 loci assayed by the Mouse Diversity Genotyping Array.

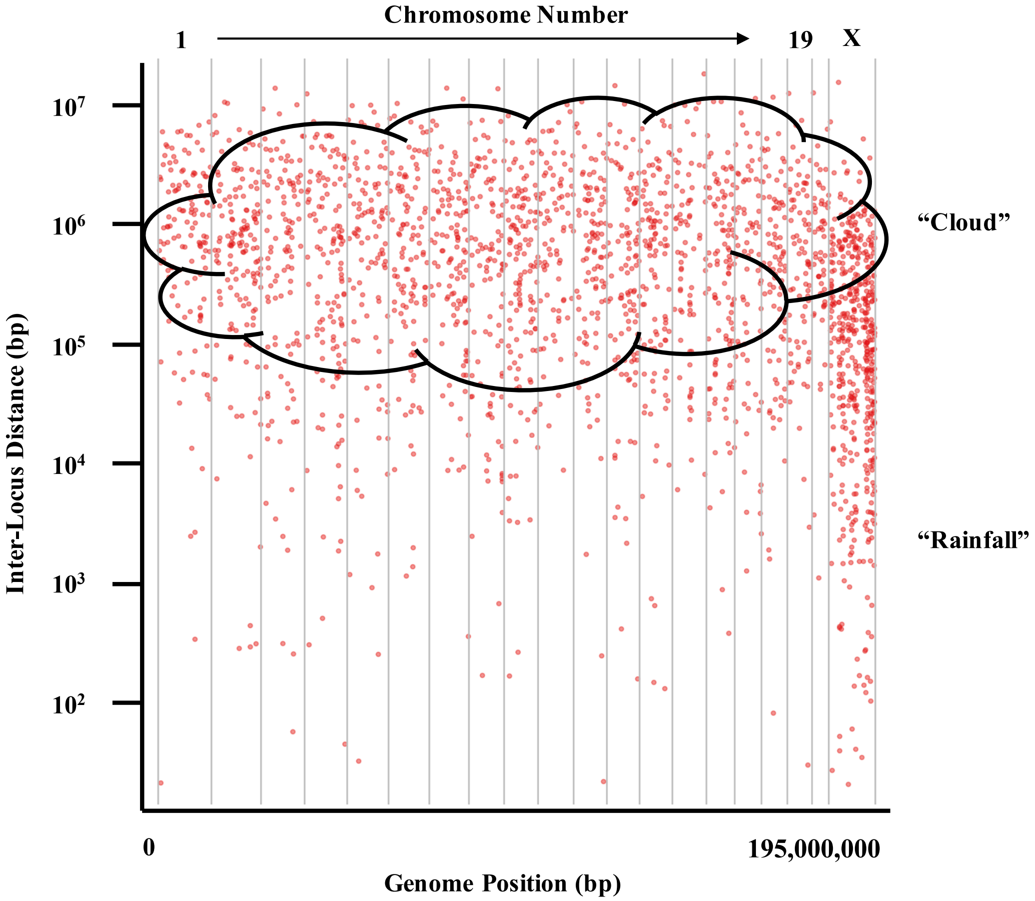


Figure 1.6.2 Rainfall plot used to visualize inter-locus distance between SNP genotypes
SNP genotypes of interest are displayed as red dots on the plot. Genomic position of SNPs and the inter-locus distance between them are displayed in base pairs (bp). SNP genotypes with a large inter-locus distance from the last queried locus in the mouse genome are represented by a dot positioned higher in the plot than a locus with a very small inter-locus distance from the previous queried locus. The first queried SNP genotype in the genome is not plotted as each dot represents a SNP locus plotted with respect to the inter-locus distance with a previous SNP locus. SNP genotypes that are uniformly distributed (inter-locus 10^6 bp spacing) across the genome compose the "cloud" of the rainfall plot, and SNP loci that are closely interspaced are the "rainfall".

data point in the scatterplot represents a SNP locus of interest, and regions of the genome with proximal clusters of SNP loci have smaller inter-locus distances, appearing as ‘rainfall’ from the cloud of SNP loci with greater inter-locus distances. SNP genotypes can also be utilized to analyze mutational signatures that are characteristic of the effect of environmental mutagens (Kucab et al., 2019). Analyzing the changes in SNP genotypes of an organism before and after exposure to a mutagen can indicate the nature of the mutagen and the mutational mechanism in environmental surveys of species. Mutational signatures examine DNA changes in a trinucleotide context (i.e. with consideration of the upstream and downstream adjacent nucleotide of the mutated nucleotide), with different possible transitions or transversions combining to create a unique signature (Figure 1.6.3; Alexandrov et al., 2013; Kucab et al., 2019; Nik-Zainal et al., 2012). Analyzing changes in SNP genotype signatures as a method of comparative genomics would allow for identification and analysis of mutagenic effects on the genome (Nik-Zainal et al., 2015).

Previous research provides evidence for the ability to assess SNP diversity cross-species. SNP diversity has been evaluated in agricultural species to assess genetic welfare of populations maintained and manage breeding strategies (Wang et al., 2018; Williams et al., 2010). SNPs have also been utilized in identifying genomic sequence for non-model species and the creation of draft genomes (Miller et al., 2015; Ogden et al., 2012). SNP diversity is an important factor to consider for conservation genetic strategies where the identification of heterozygous SNP loci or SNP loci genotyped for both alleles in a population is key to surveying genome diversity in non-model species. (Hoffman et al., 2013; McCue et al., 2012; Ogden et al., 2012).

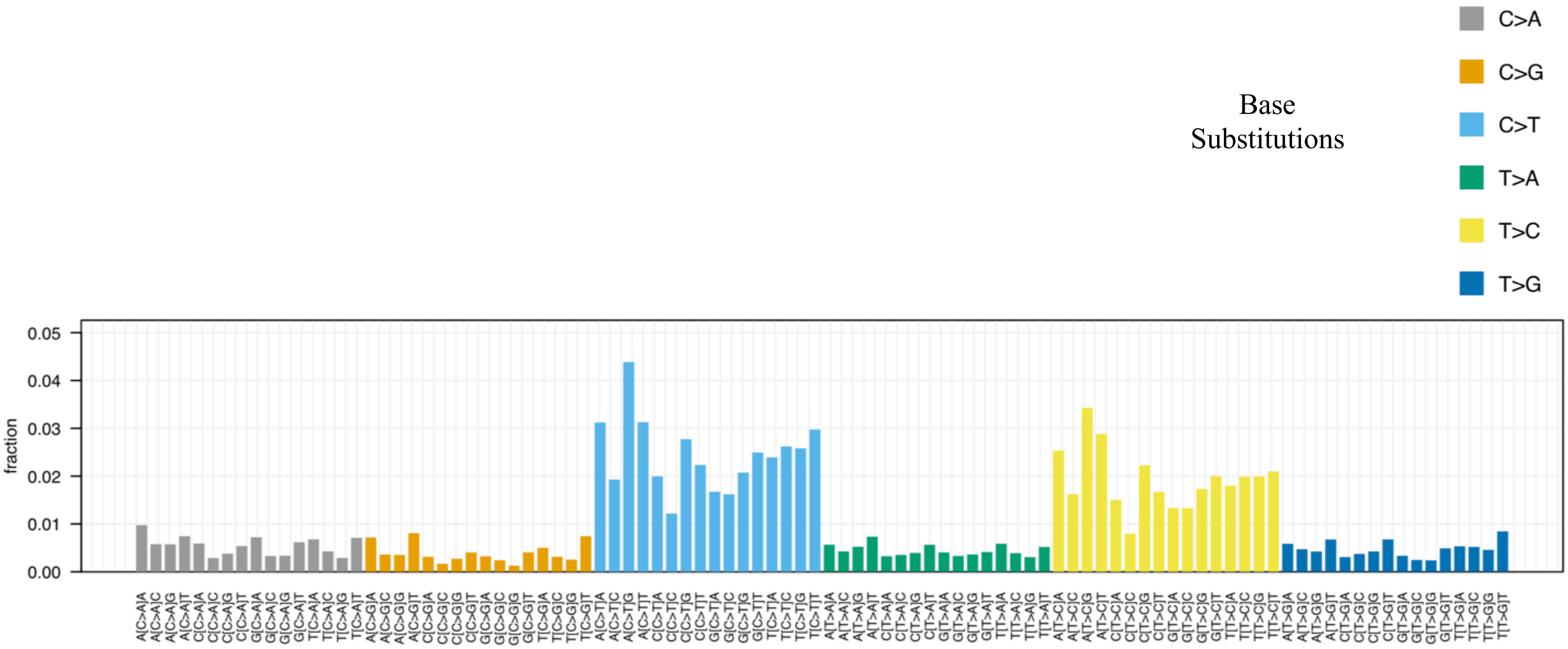


Figure 1.6.3 Mutational signature plot to visualize trends in transitions and transversions at queried SNP loci Mutation signatures analyze transitions and transversions in a trinucleotide context, taking into consideration the nucleotide upstream and downstream from the mutation. As an example, the notation A[C>G]A indicates a transversion of a C to a G in the context of A nucleotides at the 5' and 3' locations (5' NNN 3'). There are a total of 96 transitions and transversions with respect to the possible adjacent nucleotides that contribute to a mutational signature. This mutation signature plot shows the relative proportions of the base substitutions detectable by the probe sequences on the Mouse Diversity Genotyping Array.

1.7 Tools for cross-species SNP genotyping are limited

Researchers studying model organisms benefit from a wide range of tools and technologies optimized to identify SNP variation within the species of interest. Chief among the potential approaches to identifying SNPs are the genotyping microarray technologies. Other approaches to SNP identification such as restriction fragment length polymorphism (RFLP) utilize specific restriction enzymes to target loci of interest. PCR-based methods like PCR Amplification of Specific Alleles (PASA) amplify a single locus based on the presence of a specific SNP. Genotyping arrays surpass RFLP and PASA techniques because arrays can analyze hundreds of thousands of SNP loci at one time, a great many more SNP loci per assay than RFLP or PASA (Locke et al., 2015; Saifullah and Tsukahara, 2018; Yang et al., 2009; Ye et al., 2001). Lower-cost sequencing approaches like restriction-site associated DNA sequencing (RADseq) relies on restriction site-associated digestion of DNA to create libraries of specific sequence lengths. The sequence libraries can be used in conjunction with genotyping arrays to identify the potential hundreds of thousands of known SNPs that can be obtained in one genotyping assay (Wang et al., 2018; Zhao et al., 2018).

Until prices of whole genome sequencing are lowered, genotyping microarrays are the clear choice for identifying SNP genotypes in a high-throughput and cost-effective manner (Wetterstrand K 2019). There are the obvious benefits of low cost and high number of loci queried per sample with use of microarrays, but there are a few challenges. Microarray-based genotyping is dependent on hybridization of test DNA to probe sequences affixed to the array slide. Suboptimal hybridization conditions can result in false genotyping results or fewer loci genotyped depending on the array hybridization conditions and quality of the sample DNA (Bumgarner, 2013; Draghici et al., 2006). Hybridization of DNA from non-model species to genotyping arrays made for model organisms can be affected negatively if optimal hybridization conditions are very different from the model. Another issue to consider with utilizing genotyping arrays cross-species are the challenges presented during the genotyping process. A number of genotyping algorithms are employed to analyze raw microarray data and provide a genotype for each locus (Lamy et al., 2011; Rabbee and Speed, 2006). The genotyping algorithms often use a training set of samples that are separate from the test samples that are analyzed in a study (Lamy et al., 2011; Pounds et al., 2009). The purpose of the training set is to teach the genotyping algorithm to read typical raw array data and allow for greater accuracy in genotyping loci of test samples (Pounds et al., 2009). However, training sets should reflect the genetic diversity of the test set of samples. Most microarrays are made for specific model species and genotyping cross-species is a challenge. When genotyping cross-species, non-model organisms typically have greater genetic diversity that would exceed the maximum genetic diversity of the training set. The greater genetic diversity of the test set can result in false genotype assignments to occur. Underestimates or overestimates of the true number of SNP loci that are present in a non-model species and the diversity detected at SNP loci can also occur (Hong et al., 2008; Miclaus et al., 2010).

1.8 There is a precedence for utilizing SNP genotyping arrays cross-species

Researchers have previously explored the possibility of cross-species application of several genotyping arrays (Figure 1.8.1). The primary types of genotyping arrays

Published Cross-Species Comparisons

A

		Publications organized by increasing divergence time of cross-species hybridization comparisons	Array technology tested for cross-species utility	Number of loci queried by each technology
1	○	Miller *et al.* (2018)	Ovine Infinium HD SNP Beadchip	777k loci
2	○	vonHoldt *et al.* (2010)	Canine v2 SNP array	48k loci
3	●	Pertoldi *et al.* (2010)	Bovine Beadchip	53k loci
4	●	Michelizzi *et al.* (2010)	Bovine Beadchip	54k loci
5	●	Ogden *et al.* (2012)	Bovine Beadchip	54k loci
6	●	Kharzinova *et al.* (2015)	Bovine & Ovine Beadchip	55k and 54k loci
7	●	Moravčíková *et al.* (2015)	Bovine Beadchip	55k loci
8	●	Haynes & Latch (2012)	Bovine Beadchip	55k loci
9	●	Miller *et al.* (2012)	Ovine Beadchip	49k loci
10	●	More *et al.* (2019)	Bovine HD Genotyping Beadchip	777k loci
11	●	Hoffman *et al.* (2013)	Canine HD Beadchip	172k loci
12	●	McCue *et al.* (2011)	Equine Array	55k loci
13	○	Kelly *et al.* (2019)	Mouse Diversity Genotyping Array	493k loci

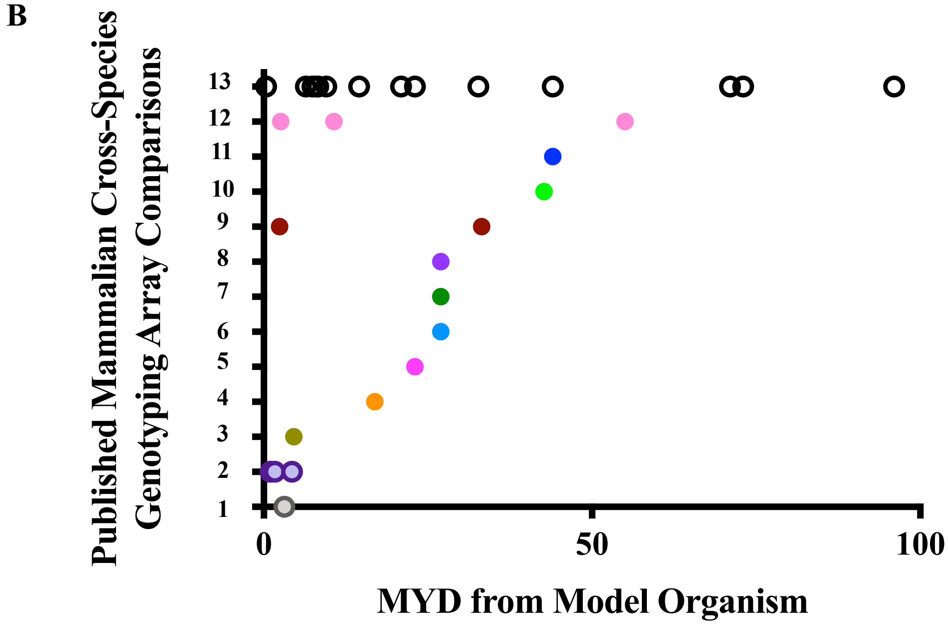


Figure 1.8.1 Summary of published research on mammalian cross-species genotyping using SNP genotyping microarrays
(A) Published research is organized in increasing order of genetic divergence in millions of years divergence (MYD) of non-model test samples from the model reference organism. Authors, publication year, genotyping microarray technology, and approximate number of loci queried (in thousands) are listed for each publication. (B) The sample of publications on mammalian cross-species array studies with the 13th representing the contributions of this thesis to the cross-species genotyping array field.

previously used in 12 published cross-species genotyping studies are designed for agricultural species and arrays designed for domestic breeding purposes including bovine, ovine, canine, and equine array technologies (Kharzinova et al., 2015; Miller et al., 2012, 2018; Moravcikova et al., 2015; More et al., 2019; Ogden et al., 2012; Pertoldi et al., 2010; vonHoldt et al., 2010). The Bovine SNP50 genotyping array designed to identify over 50,000 SNPs in the genome of cows, was applied to two species of oryx which diverged from the modern cow 23 million years ago (Ogden et al., 2012). The oryx antelope species evolved to thrive in the desert, and wild populations have declined drastically due to poaching and habitat loss. With a single application of the Bovine SNP50 array, 148 SNPs were identified in the scimitar horned oryx (*Oryx dammah*), and 149 SNPs were identified in the Arabian oryx (*Oryx leucoryx*). The novel loci discovered in the oryx species will be valuable in determining diversity and relatedness of oryx populations and aid in conservation efforts (Ogden et al., 2012).

Recently, researchers have attempted to apply a genotyping array designed for an agricultural species to non-model species that are important economically. The Bovine SNP50 array was utilized cross-species with the alpaca (*Vicugno pacos*), a species with hair fibre valued economically (More et al., 2019). Though the cow and alpaca diverged from one another approximately 42.7 million years ago, researchers identified over 6,700 alpaca SNPs that could be useful in managing breeding strategies to maximize the amount of high-quality alpaca fibre produced. This can be achieved by screening the genomes of alpaca, and breeding alpaca that have genomes enriched with SNPs identified as being linked to high quality hair fibre (More et al., 2019).

A domestic ovine SNP genotyping array has also been used cross-species to identify sexually-selected traits in the bighorn sheep (*Ovis canadensis*) in an effort to better understand the genetic underpinnings of fitness in this wild species (Miller et al., 2018). Over 3000 SNP loci were genotyped in a population of bighorn sheep, and one particular locus was found to be associated with body mass as a sexually-selected trait. Researchers concluded it was likely that sexually-selected traits were polygenic and this study marked a first step in better understanding associations of single nucleotide variation with fitness in the bighorn sheep (Miller et al., 2018).

A final example demonstrating the applicability of genotyping arrays cross-species is from a landmark study that utilized the Canine HD Beadchip genotyping array with DNA of the Antarctic fur seal, *Arctocephalus gazella* (Hoffman et al., 2013). The canine array used was designed to assay over 22,000 SNPs in diverse domestic dog breeds. While researchers vonHoldt *et al.* (2010) had previously applied a canine genotyping array to wolf species, researchers Hoffman *et al.* (2013) attempted to use the canine array to characterize SNP variation within fur seal populations, which had become endangered due to effects of climate change. Despite a 44 million-year divergence time between dogs and Antarctic fur seals, 173 SNPs were identified as being conserved between these species. The conserved SNPs were associated with genes involved in energy metabolism and may become relevant in future studies that aim to understand the types of polymorphisms that are retained over vast evolutionary timespans (Hoffman et al., 2013).

1.9 Rodents are candidates for cross-species SNP genotyping

While there are studies that utilize genotyping arrays that were designed for agricultural or economic breeding purposes for cross-species genotyping in related species, there is a lack of research that explores cross-species genotyping within rodents. Rodents are extremely fecund and live in a multitude of different environments across the globe. A number of rodents live commensally alongside humans in environment with unique selective pressures created by human influence that affect the genomes of rodents (Hulme-Beaman et al., 2016). Rodents adapting to rapidly changing human environments offer a unique opportunity to examine accelerated evolution (Harris et al., 2013; Hulme-Beaman et al., 2016). Wild rodents living commensally with humans are exposed to similar environmental mutagens that humans are exposed to, and therefore are prime candidates for monitoring mutagenesis caused by environmental agents (da Silva et al., 2000; Silva et al., 2000).

There are many non-model rodent species that are of special interest for genetic research. One example is that of the naked mole rat, *Heterocephalus glaber*. This species of rodent lives in subterranean tunnels of the African desert and is one of the two eusocial mammals on Earth (Jarvis, 1981). Naked mole rats are the longest-lived rodents (Csiszar et al., 2007; Sahm et al., 2018) and also have very low cancer incidence rates (Seluanov et al., 2018; Tian et al., 2013). There has been preliminary work done in sequencing the genome of the naked mole rat, but the genome is currently composed of unplaced genomic scaffolds from shotgun sequencing (Keane et al., 2014). Further development and annotation of the naked mole rat genome is required to facilitate use of this unique organism in research.

Other interesting potential candidates of non-model rodents are species from the genus Peromyscus. Peromyscus species are referred to as deer mice, although they diverged from the house mouse (*Mus musculus*) over 30 million years ago (Bedford and Hoekstra, 2015; Hedges et al., 2015). Deer mice live dispersed across all of North America in very diverse environments (Bedford and Hoekstra, 2015). Peromyscus species have been previously used in studies of environmental monitoring at Alberta oil sand sites as sentinels of the effects of environmental contaminants (Rodríguez-Estival and Smits, 2016). The study focused on morphological differences caused by environmental contaminants found in tissues of deer mice, as there is a lack of genomic data available for Peromyscus species. Transcriptomic sequence changes from a group of deer mice living in urban and rural environments were analyzed by researchers who discovered evidence of rapid evolution (Harris et al., 2013). Peromyscus species of interest to both evolutionary and environmental monitoring studies would benefit from fully sequenced and annotated genomes to facilitate future research.

Rodents compose over 2,000 species on Earth, live in diverse environments, and have a large amount of genomic diversity (Krubitzer et al., 2011). Rodent genomes seem to undergo rapid evolution in certain cases, with new exons frequently being created (Wang et al., 2005). Peromyscus is a genus of species that are shown to undergo rapid genomic evolution, making them an interesting model for adaptation and population genetic studies (Harris et al., 2013; Ramsdell et al., 2008). Interestingly, though Peromyscus species are referred to as deer mice, they show greater genetic similarity to species of rats than species of the genus Mus (Ramsdell et al., 2008). Linkage group analysis between genomes of Rattus, Mus, and Peromyscus species established that Mus species have undergone more recent genome rearrangements than rats or deer mice (Ramsdell et al., 2008). The recent genomic rearrangements in the house mouse introduce a challenge when determining the spatial location of conserved variation in the genome that may be present between the model house mouse compared to distant relatives like deer mice.

1.10 Estimations of divergence times for rodents are derived from molecular data

The divergence time is the number of years from the most recent common ancestor (MRCA) between two species. Divergence time can be determined from geological data including fossil records (Tavaré et al., 2002). Molecular data such as ribosomal sequences (Guterres et al., 2018), highly conserved mitochondrial coding sequences (Nicolas et al., 2012; Rudra et al., 2016), and Y chromosomal sequences (Eusebi et al., 2017) can be used to determine divergence time in conjunction with fossil record data. Divergence times may also be estimated using comparisons of inherited repetitive stretches of sequences known as microsatellites that are located throughout the genome (Sun et al., 2009). More recently, researchers have worked to create more precise estimates of divergence times between organisms from an amalgamation of molecular and geological data. The online public knowledge-base ‘Timetree - the timescale of life’ provides estimates of relative divergence times between taxa and draws this information from over 3,900 studies that represent over 97,000 species (Hedges et al., 2015).

1.11 Mouse Diversity Genotyping Array is a candidate tool for cross-species study

The house mouse, *M. musculus*, has been used widely in genetic studies and has a fully sequenced and annotated genome developed through wide use of this organism in research. There are many tools that have been created to conveniently assess SNP diversity within the genome of the house mouse. The Mouse Diversity Genotyping Array (MDGA) is a tool designed to survey hundreds of thousands of SNP loci across the genome of the house mouse and was specifically created to maximize the amount of SNP diversity that can be identified within laboratory mouse strains and crosses (Yang et al., 2009). The MDGA has better genome coverage than many other array technologies. Another array technology available that was designed to characterize SNP variation in lab mouse strains and crosses is the Mouse Universal Genotyping Array (MUGA) which can detect up to 141,090 SNPs (Morgan et al., 2016). By comparison, the MDGA is advertised as capable of detecting over 600,000 SNP genotypes in the genomes of laboratory mice, and the majority of the SNP genotypes detected are located in non-coding regions of the genome (Yang et al., 2009). After testing and the removal of poorly performing SNP probes, the MDGA was found to genotype 493,290 SNP loci within the genome of the house mouse (Locke et al., 2015). The MDGA identifies hundreds of thousands more SNPs than MUGA, making the MDGA an attractive tool to test applicability cross-species with rodents.

1.12 MDGA is a hybridization-based genotyping array technology

The MDGA is a hybridization-based genotyping tool that relies on complementary binding of target DNA to interrogating probes affixed to the array slide. The MDGA is capable of assaying SNP genotypes at 493,290 SNP loci within laboratory strains of mice, and also contains over 900,000 probes that query copy number variants. In detecting SNP genotypes, there are eight probes that are located at different positions on the array slide that all target the same SNP locus. Of the eight probes, four target the major or most common SNP allele, and four target the minor or less common SNP allele (Figure 1.12.1). The eight probes that target the two alleles are offset in genome sequence from one another to increase accuracy of genotyping. The redundancy of the probe design on the MDGA provides greater confidence in determining a genotype at each locus. The signal from all eight DNA fragments that have attached adaptors are amplified by PCR and a pool of the amplified DNA is created through purification using polystyrene beads (Figure 1.12.2). The pool of amplified DNA is fragmented and labelled with a fluorescent signal. Labelled DNA is applied to the array and given time to hybridize to array probes. After hybridization, the array is washed to remove unbound DNA and raw fluorescence intensities are read by a scanner. An image file of raw genotype signals, known as a CEL file, is produced. The CEL file is produced can be used with a genotyping algorithm to assign genotypes at each queried locus.

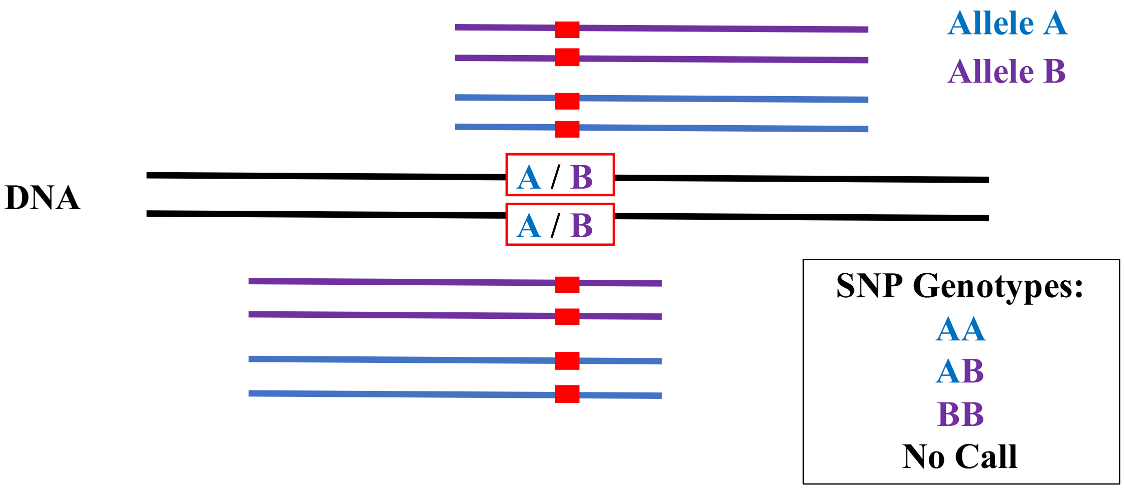


Figure 1.12.1 Redundancy of MDGA probe design to target major and minor SNP alleles
Four SNP probes target allele A (blue) and four SNP probes target the B allele (Purple). In total, 8 probes on the MDGA target each SNP locus of interest (red box) in the house mouse (*M. musculus*). Hybridization intensities are averaged across all 8 probes in determining a genotype at a particular SNP locus queried.

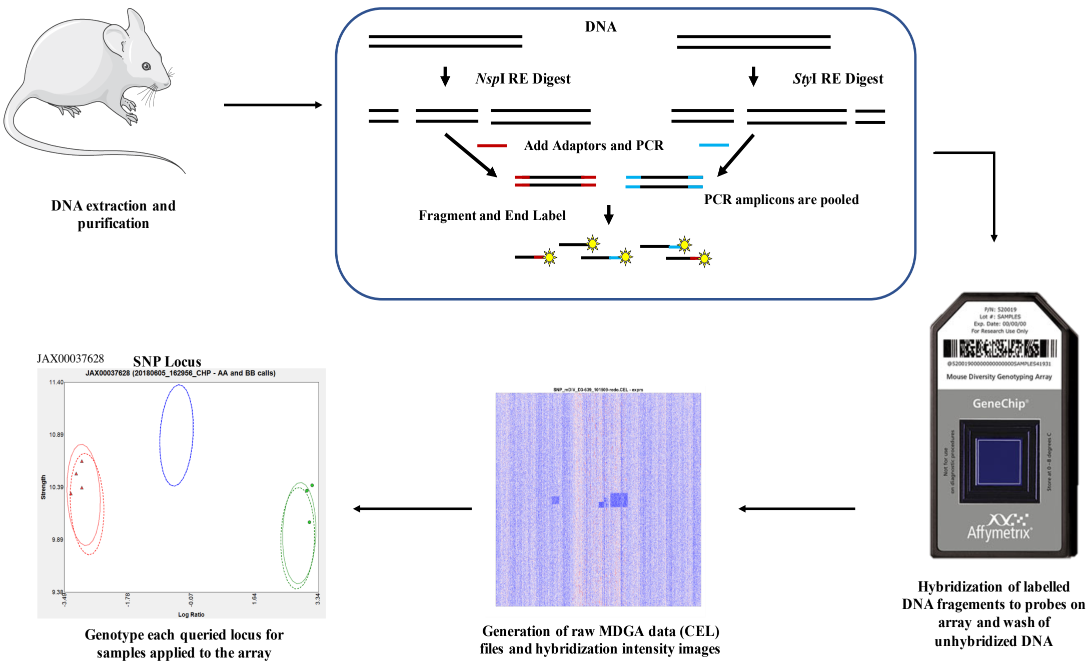

DNA extraction and purification
DNA
NspI RE Digest
StyI RE Digest
Add Adaptors and PCR
PCR amplicons are pooled
Fragment and End Label
Mouse Diversity Genotyping Array
GeneChip
Affymetrix
Hybridization of labelled DNA fragements to probes on array and wash of unhybridized DNA
Generation of raw MDGA data (CEL) files and hybridization intensity images
JAX00037628
SNP Locus
JAX00037628 (20180605_162956_CHP - AA and BB calls)
11.40
10.89
10.39
9.89
9.38
Strength
-3.49
-1.78
-0.07
1.64
3.34
Log Ratio
Genotype each queried locus for samples applied to the array

Figure 1.12.2 Mouse Diversity Genotyping Array SNP genotyping process
DNA is extracted, purified, and prepared for hybridization to the MDGA. After hybridization, the array is washed, and hybridization intensity images are generated. Genotyping of SNP loci of DNA samples applied to the MDGA is performed with Affymetrix Power Tools (APT) Release 1.16.0.

1.13 Bioinformatic resources and tools to analyze SNPs in mice

Mice are well-established model organisms and have a wealth of SNP information available through the Mouse Genome Informatics international database (Zhu et al., 2015). The mouse genome database can be used to mine *in silico* information to search for phenotypic effects of SNP loci queried by the MDGA (Eppig et al., 2015). *In silico* validation of genotype assignments made from MDGA array data can be done with bioinformatic tools like the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990). BLAST is a useful tool for aligning a relatively small number of sequences to a publicly available genome, but this tool is computationally taxing and slow when attempting to align hundreds of thousands of SNP array probe sequences to publicly available genomes. A new software tool 'efficient computation of maximal exact matches' (E-MEM) is capable of aligning hundreds of thousands of unique SNP array probe sequences to genomes of interest. By mapping array probe sequences to non-model genome sequences available online, SNP loci that are genotyped using the array technology can be cross-validated as being present in non-model organisms (Khiste and Ilie, 2015).

1.14 Central Hypothesis

Given that there is greater genetic identity between organisms of the same species than between species and beyond a genus, family, and order, it is hypothesized that the Mouse Diversity Genotyping Array will have greater utility with non-model Mus species than for non-Mus organisms. The MDGA contains probe sequences that are complementary to 493,290 unique loci that contain known single nucleotide variation within 12 classical inbred and 7 wild-derived strains of mice, and wild (feral) mice. It is hypothesized that application of the MDGA to wild rodent DNA samples will help elucidate potential polymorphic loci, or the number of loci that can detect both the A and B allele in a population, and that can be used cross-species.

1.15 Experimental Aims

The first experimental aim has three steps. The first step is to define the limits for cross-species utility of the MDGA for publicly available samples organized at four levels of taxonomic classification that have different maximum divergence times from the reference house mouse. A test set is a set of samples from different organisms that are genotyped together in a group. The number of samples and the types of organisms affects the genotyping results. The first aim was accomplished by analyzing SNP genotypes of test sample sets including an intra-genus test set (9.5 MYD, n = 27), an inter-genus test set (32.7 MYD, n = 37), an inter-family test set (73 MYD, n = 31), and an inter-order test set (96 MYD, n = 40). A pairwise comparison of the differences in SNP genotypes between samples of different test sets and the reference house mouse was used to construct trees of genetic relatedness based on SNP genotypes at MDGA queried loci. It was predicted that genotyping results for wild rodents would reflect what would be expected for each species based on published determinations of species divergence from *M. musculus*. SNP trees of genetic relatedness are expected to reflect the known patterns of divergence established in literature.

As a method of validating the experimental genotyping results obtained using the MDGA, the second step of aim one is to map MDGA probe target sequences to available online genomes of test samples. An *in-silico* search was performed using the program E-MEM to search for MDGA target sequences that are present a single instance in the available genomes of wild rodent species. It was predicted that the number of unique matches will decrease as divergence time of non-model species from *M. musculus* increases. MDGA loci that are genotyped experimentally in wild samples using the array and are also mapped a single time in the available online genome are candidate SNP loci that may represent conserved SNP variation between the reference house mouse and wild rodent. In particular, SNP loci with heterozygous genotypes may represent potential polymorphic loci that can identify both the A and B alleles in non-model species.

The third step of experimental aim one is to take the candidate loci that are genotyped using the array and can be mapped to an online genome and determine functional pathways that are shared between the non-model species and the reference *M. musculus*. Ensembl gene IDs associated with candidate loci will be analyzed with the functional annotation tool of the Database for Annotation, Visualization, and Integrated Discovery (DAVID). It is predicted that pathways involved in already recognized, highly conserved functions will be shared between the house mouse and wild rodent species.

The second experimental aim is to examine the potential for inter-genus and inter-family cross-species utility of the MDGA in two case studies of species of interest. The rodents examined in the case studies are the naked mole rat (*H. glaber*) which differs from *M. musculus* in taxonomic family (73 MYD), and species of the genus Peromyscus. Peromyscus species are commonly known as deer mice and are from a different genus than the house mouse (32.7 MYD). Four naked mole rat samples were examined in the first case study, and it was determined whether genotyping results can be utilized to recapitulate the known relationships between the samples based on sex and colony of the organism. It was expected that at a divergence time of 73 million years from the house mouse, the array may have an insufficient number of informative SNP probe sequences to detect differences in source colony and sex of the organism. It is more likely that the major genetic differentiation between naked mole rats will be by sex differences. Conserved functional pathways between naked mole rats and the house mouse were also investigated. In the second case study, seven samples of deer mice composed of six different species (with one species *P. maniculatus* represented by two subspecies) were examined to determine if genotyping results produced from raw array data can be used to differentiate samples according to known divergence patterns established in literature. MDGA SNP loci that were genotyped in *P. maniculatus* subspecies were cross-validated using an *in silico* search for the unique presence of SNP loci in the genome assembly. Functional pathway associations shared between deer mice and the house mouse were examined for cross-genus conservation. It was expected that the genotyping results produced from MDGA data can be used to differentiate species according to established divergence times. Given that conserved variation was detected between the dog and seal at 44 MYD by Hoffman *et al.* (2013), it was expected that conserved variation between the house mouse and Peromyscus species at 32.7 MYD would be detected.

The third experimental aim is broken down into two steps. The first step was to examine the genomic distribution of variation genotyped within all 27 Mus samples across the 19 chromosomal pairs of autosomes with the reference house mouse. Loci were classified according to the degree to which a genotype changes across all 27 Mus samples (is variable), or remains a constant (or invariant) genotype for all samples. Loci with genotypes that are variable and invariant across test samples were plotted at the genomic position of the reference house mouse and analyzed for specific patterns or clustering of loci across the genome. It was expected that there will be fewer variable loci on the X chromosome than the autosomes due to the high degree of conservation associated with this sex chromosome.

The second step was to create a visualization of genotype changes spatially across a chromosome and temporally for different Mus species differing in evolutionary divergence. The analysis is restricted to Mus species that contain 19 autosomes, the same number as the reference house mouse. It was expected that there will be an increase in the number of AB and BB genotypes in wild samples compared to the reference house mouse. There was also an expectation that changes in the genotypes across the genome and across Mus species will not be random. It was expected some SNP loci will be variable in genotype and that others will be invariant in genotype. Particular genotypes that remain unchanged at loci across species of different evolutionary divergence times may indicate conserved SNP variation between the model house mouse and its wild relatives.

2 Materials and Methods

2.1 Assessing limits of cross-species applicability of the MDGA across forty samples of wild mammalian species.

Forty publicly available MDGA raw data (CEL) files were downloaded from the Center for Genome Dynamics at the Jackson Laboratory (2012, The Jackson Laboratory; ftp.jax.org/petrs/MDA/). The forty samples consist of twenty-seven Mus CEL files, two Rattus CEL files, seven Peromyscus (deer mouse) CEL files, one Apodemus (wood mouse) CEL file, and CEL files representing more highly diverged species including a squirrel, mountain tapir, and African Black Rhino (Table 2.1.1). The forty samples downloaded were grouped into different test sets that produced different results to analyze after genotyping. One additional *M. musculus* CEL file was utilized as a reference for the house mouse genome on which the MDGA was designed (Table 2.1.1).

CEL file raw array intensity images were analyzed for quality control purposes and hybridization abnormalities in array images were noted (Appendix A). Two CEL files were noted for having an abnormal spot with uneven DNA hybridization to the array that is referred to as a “coffee ring” formation (Jose M. Moran-Mirabal et al. 2006; Hu and Larson 2006). The abnormal samples were not removed from the analysis due to the redundancy of probe design across the MDGA. There are also technical replicates for the abnormal *M. saxicola* and *M. nannomys orangiae* CEL files that are devoid of hybridization abnormalities included in the study. There were no exclusions of data from the sample sets utilized.

Table 2.1.1 Forty publicly available MDGA data (CEL) files of the present study

CEL File[a]	**Sex of Organism**	**Scientific Name**	**Common Name**	**Divergence Time**[b] **from *Mus musculus* (MYD)**
SNP_mDIV_A7-7_081308.CEL	Male	*Mus musculus*	House Mouse Reference	0
SNP_mDIV_D3-639_101509-redo[c]	Female	*Mus musculus castaneus*	Southeastern Asian House Mouse	0.35
SNP_mDIV_D3-639_91809	Female	*Mus musculus castaneus*	Southeastern Asian House Mouse	0.35
SNP_mDIV_D9-647_101509-redo	Male	*Mus dunni*	Earth-Colored Mouse	6.4
SNP_mDIV_D9-647_91809	Male	*Mus dunni*	Earth-Colored Mouse	6.4
SNP_mDIV_D4-640_101509-redo	Male	*Mus famulus*	Servant Mouse	6.4
SNP_mDIV_D4-640_91809	Male	*Mus famulus*	Servant Mouse	6.4
SNP_mDIV_D8-474_012209	Male	*Mus famulus*	Servant Mouse	6.4
SNP_mDIV_D5-642_101509-redo	Male	*Mus fragilicauda*	Sheath-Tailed Mouse	6.4
SNP_mDIV_D5-642_91809	Male	*Mus fragilicauda*	Sheath-Tailed Mouse	6.4

[a] MDGA data (CEL) files were downloaded from the Center for Genome Dynamics at the Jackson Laboratory.

[b] Divergence time is given in millions of years from the reference house mouse, *Mus musculus* (timetree.org).

[c] "redo" files are a technical replicate of the CEL file with the same sample identifier code. Ex: SNP_mDIV_D3-639_101509-redo is a technical replicate of SNP_mDIV_D3-639_91809, where D3-639 is the sample identifier.

SNP_mDIV_D6-643_101509-redo	Male	*Mus fragilicauda*	Sheath-Tailed Mouse	6.4
SNP_mDIV_D6-643_91809	Male	*Mus fragilicauda*	Sheath-Tailed Mouse	6.4
SNP_mDIV_D7-644_101509-redo	Male	*Mus caroli*	Ryukyu Mouse	7.41
SNP_mDIV_D7-644_91809	Male	*Mus caroli*	Ryukyu Mouse	7.41
SNP_mDIV_D6-472_012209	Male	*Mus caroli*	Ryukyu Mouse	7.41
SNP_mDIV_D8-646_101509-redo	Male	*Mus cervicolor*	Fawn-Coloured Mouse	7.41
SNP_mDIV_D8-646_91809	Male	*Mus cervicolor*	Fawn-Coloured Mouse	7.41
SNP_mDIV_A2-645_102109	Male	*Mus cookii*	Cook's Mouse	7.41
SNP_mDIV_A3-648_102109	Male	*Mus platythrix*	Flat-Haired Mouse	8.1
SNP_mDIV_A4-649_102109	Male	*Mus platythrix*	Flat-Haired Mouse	8.1
SNP_mDIV_A5-650_102109	Male	*Mus saxicola*	Rock-Loving Mouse	8.1
SNP_mDIV_A6-651_102109	Male	*Mus saxicola*	Rock-Loving Mouse	8.1
SNP_mDIV_D7-473_012209	Male	*Mus pahari*	Shrew Mouse	8.29
SNP_mDIV_D11-653_101509-redo	Male	*Mus nannomys minutoides*	African Pygmy Mouse	9.5
SNP_mDIV_D11-653_91809	Male	*Mus nannomys minutoides*	African Pygmy Mouse	9.5

SNP_mDIV_D10-652_101509-redo	Male	*Mus nannomys orangiae*	Orange Mouse	9.5
SNP_mDIV_D10-652_91809	Male	*Mus nannomys orangiae*	Orange Mouse	9.5
SNP_mDIV_A7-654_102109	Male	*Mus nannomys mattheyi*	Matthey's Mouse	9.5
SNP_mDIV_B8-1190_082410	Male	*Apodemus sylvaticus*	Wood Mouse	14.5
SNP_mDIV_A9-656_102109	Male	*Rattus norvegicus*	Sprague Dawley rat	20.9
SNP_mDIV_A10-657_102109	Male	*Rattus norvegicus*	Outbred Wistar rat	20.9
SNP_mDIV_B1-659_102109	Male	*Peromyscus aztecus*	Aztec Mouse	32.7
SNP_mDIV_B3-661_102109	Male	*Peromyscus californicus*	California Mouse	32.7
SNP_mDIV_B5-663_102109	Male	*Peromyscus maniculatus bairdii*	North American Deer Mouse	32.7
SNP_mDIV_B4-662_102109	Male	*Peromyscus maniculatus sonoriensis*	Sonoran Deer Mouse	32.7
SNP_mDIV_B2-660_102109	Male	*Peromyscus melanophrys*	Plateau Deer Mouse	32.7
SNP_mDIV_B6-664_102109	Male	*Peromyscus polionotus*	Oldfield Mouse	32.7
SNP_mDIV_B8-666_102109	Male	*Peromyscus leucopus*	White-Footed Mouse	32.7
SNP_mDIV_B9-667_102109	Male	*Sciuridae*[a]	Squirrel	71

[a] Only family level information available for CEL file SNP_mDIV_B9-667_102109; Genus and species of sample are unknown.

SNP_A2-GES11_4907_AGT-JLP-120115-24-35517	Male	*Diceros bicornis*	African Black Rhino	96
SNP_A1-GES11_4902_AGT-JLP-120115-24-35517	Male	*Tapirus pinchaque*	Mountain Tapir	96

Forty samples were genotyped by Maja Milojevic in Dr. Kathleen Hill's Laboratory using the protocol outlined by Locke et al. (2015). Affymetrix (Affy) Power Tools (Gao, Pirani, Webster 2013) was used to generate genotype calls of AA, AB, BB, or No Call (numerical representations 0, 1, 2, -1, respectively) using the BRLMM-P algorithm for 493,290 SNPs (Affymetrix (Affy) Power Tools (APT) Release 1.16.0). The SNP probes used in genotyping are a filtered list generated by previous members of the Hill laboratory (Eitutis, 2013, Thesis; Milojevic, 2019, Thesis; Locke et al. 2015). A training set of 114 classical laboratory mouse CEL files obtained from a set of 351 mice utilized by Didion et al. (2012) was used in conjunction with BRLMM-P to train the algorithm in accurate assignment of genotypes (Appendix A). The training set of 114 classical laboratory mice are recommended for use with training the genotyping algorithm for MDGA data (Eitutis, 2013, Thesis; Milojevic, 2019, Thesis; Locke et al. 2015).

A Fisher's exact test was utilized to assess the level of genetic differences between samples of genotyping sets. A nonparametric, unordered, Fisher-Freeman-Halton exact test (Monte Carlo simulation) was performed using the StatXact statistical analysis software package (CYTEL Software, Cambridge, MA). Pearson's r was used in tests of the significance of correlations between the genotyping results of test set samples using Graphpad Prism 8 software.

The estimated divergence time of each species within the forty CEL file sample set from the reference house mouse were obtained using an evolutionary timetree of life (Hedges et al., 2015) (http://www.timetree.org/) with a few exceptions. The estimated divergence time of the subspecies *M. m. castaneus* was determined through previous work by

Geraldes et al. (2012), and the evolutionary divergence time of the pygmy mouse species from the house mouse was determined by Kouassi et al. (2008).

Different combinations of samples based on divergence times from *M. musculus* comprise different test sets of study (Table 2.1.2). The test sets were organized according to differences in taxonomic classifications and maximum divergence times of samples from *M. musculus*, including inter-order (96 MYD), inter-family (73 MYD), inter-genus (32.7 MYD), and intra-genus (9.5 MYD) comparisons. The percentage of loci genotyped within test species using the MDGA and the percentage of loci with a heterozygous genotype were determined from the raw results generated by Affy Power Tools for the inter-order test set.

Pairwise comparison of SNP genotypes between species in the inter-order test set was utilized to create SNP-based distance matrices using R. The distance matrix values used to create phenograms (SNP trees) were generated using an in-house R script courtesy of Marjorie E. Osbourne Locke. The in-house script utilized the ‘bionj’ R package to create a tree of genetic relatedness using the neighbour-joining method (Gascuel, 1997). The resulting trees were modified using Figtree (v1.4.3) software. Pairwise genetic distances were computed by dividing the total number of genotypic differences between two samples by the total number of loci queried by the MDGA, where 493,290 total loci were used in this study (Locke et al., 2015). The values in the distance matrix are a numerical representation of the amount of genetic diversity between test species analyzed and the reference house mouse. A genetic distance value of zero indicates the species are genetically the same at the loci queried, and a value of one indicates the species compared are completely genetically dissimilar from one another at the loci queried. The

Table 2.1.2 Genotyping sets of study[a]

Genotyping Test Sets			Common Name	Scientific Name
Inter-Order	Inter-Genus	Intra-Genus & Inter-family*	House Mouse	*Mus musculus*
			South-Eastern House Mouse	*Mus musculus castaneus*
			Earth-Colored Mouse	*Mus dunni/Mus terricolor*
			Servant Mouse/Bonhote's Mouse	*Mus famulus*
			Sheath-Tailed Mouse	*Mus fragilicauda*
			Ryukyu Mouse	*Mus caroli*
			Fawn-Colored Mouse	*Mus cervicolor*
			Cook's Mouse	*Mus cookii*
			Flat-Haired Mouse	*Mus platythrix*
			Rock-Loving Mouse	*Mus saxicola*
			Gairdner's Shrewmouse	*Mus pahari*
			African Pygmy Mouse	*Mus (nannomys) minutoides*
			Orange Pygmy Mouse	*Mus (nannomys) orangiae*
			Matthey's Mouse	*Mus (nannomys) mattheyi*
			Wood Mouse	*Apodemus sylvaticus*
			Sprague Dawley Rat	*Rattus norvegicus*
			Wistar Rat	*Rattus norvegicus*
			Aztec Mouse	*Peromyscus aztecus*
			California Mouse	*Peromyscus californicus*
			North American Deer Mouse	*Peromyscus maniculatus*
			Sonoran Deer Mouse	*Peromyscus maniculatus*
			Plateau Deer Mouse	*Peromyscus melanophrys*
			Oldfield Mouse/Beach Mouse	*Peromyscus polionotus*
			White-Footed Mouse	*Peromyscus leucopus*
			Squirrel	*Sciuridae*
	Inter-family	*	Naked Mole Rat[b]	*Heterocephalus glaber*
			African Black Rhino	*Diceros bicornis*
			Mountain Tapir/Wooly Tapir	*Tapirus pinchaque*

[a] Genotyping sets organized in descending order according to bounds of taxonomic classification and differences in maximum genetic divergence of a test set from the reference C57BL/6J (*Mus musculus*) organism

[b] The naked mole rat samples are combined with the Mus samples to create the inter-family test set.

estimated evolutionary relationships seen in the SNP trees generated were compared to the divergence times of test samples from the reference house mouse provided in literature and the Timetree database (Geraldes et al., 2012; Hedges et al., 2015; Kouassi et al., 2008).

2.2 Naked mole rat case study of colony and sex differences in a eusocial mammal

Four CEL files were generated in-house from genomic DNA extracted from tail tissue samples of four *Heterocephalus glaber* individuals given to the Hill Laboratory by Dr. Melissa Holmes (Assistant Professor at the University of Toronto, Mississauga Campus; Table 2.2.1; Appendix A Online). DNA extractions were performed by Chloe Rose (2013, Thesis), and application of DNA to the MDGA was performed by the London Regional Genomics Center. The four samples were genotyped separately from publicly available test samples by Maja Milojevic (2019, Thesis).

The percentage of loci genotyped using the MDGA in the four naked mole rat samples as well as the percentage of loci with heterozygous genotypes were determined from the raw results generated by Affy Power Tools. Heterozygous loci represent potential polymorphic loci with utility cross-species in the naked mole rat. Pairwise distance measures were used in generation of SNP trees using the neighbour-joining method (Gascuel, 1997).

In silico validation of loci genotyped from MDGA data was performed using the program E-MEM (efficient computation of maximal exact matches for very large genomes) designed by Khiste and Ilie (2015). The publicly available genomes of rodents were

Table 2.2.1 Naked mole rat case study samples[a]

MDGA Data (CEL) File Name	Sex of Organism	Colony of Origin
DNA3340.CEL	Male	Colony Q
DNA3339.CEL	Male	Colony Q
DNA3338.CEL	Female	Colony Q
DNA3337.CEL	Female	Desperado Colony

[a] Four *Heterocephalus glab*er samples were donated by Dr. Melissa Holmes, Assistant Professor at the University of Toronto, Mississauga Campus.

searched for the unique presence of MDGA probe sequences. E-MEM was employed to search a publicly available genome of *H. glaber* available on NCBI (Table 2.2.2) for perfect 25 nt MDGA SNP probe target sequences that have only one genomic match (ftp.ncbi.nlm.nih.gov/genomes/). Unique MDGA matches discovered via E-MEM were filtered for SNP loci with an associated Ensembl (https://useast.ensembl.org/index.html) gene ID using Python and Microsoft Excel software.

Table 2.2.2 Study species with publicly available nuclear genome sequence information[a]

Sample Name	Scientific Name	Newest Assembly
House Mouse	*Mus musculus*	GRCm38.p6
Ryukyu Mouse	*Mus caroli*	CAROLI_EIJ_v1.1
Gairdner's Shrewmouse	*Mus pahari*	PAHARI_EIJ_v1.1
Sprague Dawley Rat	*Rattus norvegicus*	Rnor_6.0
North American Deer Mouse	*Peromyscus maniculatus*	Pman_1.0
Naked Mole Rat	*Heterocephalus glaber*	HetGla_female_1.0

[a] Genomes accessed through the NCBI Genomes FTP site of samples under study (ftp.ncbi.nlm.nih.gov/genomes/)

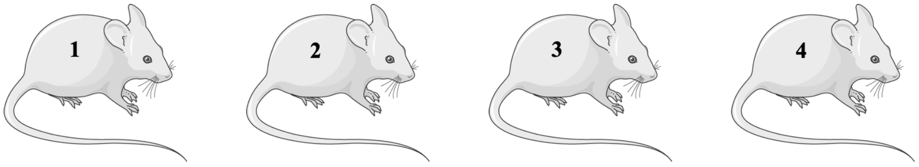
2.3 Deer mouse case study of non-Mus intra-genus genetic diversity

Seven Peromyscus (deer mouse) publicly available CEL files were genotyped together using Affy Power Tools. The percentage of queried loci that were genotyped and the percentage of genotyped loci with a heterozygous genotype were determined using the raw genotyping output. Heterozygous loci represented potential polymorphic loci that could have utility cross-species for Peromyscus. Pairwise distance measures and SNP trees were generated using the neighbour-joining method (Gascuel, 1997). The program E-MEM designed by Khiste and Ilie (2015) was utilized to search a publicly available genome of *Peromyscus maniculatus* available on NCBI for perfect 25 nt MDGA probe target sequences that map to the *P. maniculatus* genome only once (ftp://ftp.ncbi.nlm.nih.gov/genomes/). MDGA loci that gave a genotype in the mouse and were mapped to the *P. maniculatus* genome were assessed for functional associations using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) functional annotation tool (Huang et al., 2009a; b). The functional annotations used were mouse Ensembl gene IDs. The IDs were submitted to the DAVID functional annotation tool and pathways from The Kyoto Encyclopedia of Genes and Genomes (KEGG) that were significantly enriched (p<0.001) for genes associated with MDGA SNP loci were identified.

2.4 Assessing cross-species applicability of the MDGA across 27 DNA samples of wild Mus species

Twenty-seven publicly available CEL files of wild Mus species were genotyped together as the intra-genus test set using Affy Power Tools. Of the queried loci that could be genotyped, the number of loci with heterozygous genotypes were identified. Heterozygous loci represent potential polymorphic loci with utility in surveying diversity cross-species as both the A and B alleles can be identified in a non-model organism. Pairwise distance measures were utilized with the neighbour-joining method of generating SNP trees (Appendix B; Gascuel, 1997). *In silico* cross-validation of loci genotyped using MDGA data was performed using the program E-MEM. E-MEM was used to search publicly available genomes of *Mus pahari* and *Mus caroli* from NCBI for unique genomic matches of MDGA target sequences (Table 2.2.2). Genotyped SNP target sequences of *M. pahari* and *M. caroli* that were also mapped to the publicly available genomes using E-MEM were utilized as candidate conserved SNP loci. Candidate SNP loci with associated mouse Ensembl gene IDs were analyzed using the DAVID functional annotation tool (https://david.ncifcrf.gov/). KEGG pathways found to be significant (p<0.001) were assessed.

There are loci genotyped in Mus samples that share the same genotype for all samples of the study, and these were termed invariant genotype SNP loci (Figure 2.4.1). There can be invariant AA, AB, and BB loci that share the same genotype across all samples in a genotyping set. There are other loci that have different genotypes between samples in a genotyping set, and these loci were termed variable genotype SNP loci (Figure 2.4.1). A MDGA probe sequence that was attributed to only ‘No Calls’ or the inability to determine a genotype at a particular location in all samples of a genotyping set was termed an uninformative locus. There were no uninformative loci in any of the test sets of study.



Variable Locus	**AA**	**AB**	**BB**	**No Call**
Invariant AA Locus	**AA**	**AA**	**AA**	**AA**
Invariant AB Locus	**AB**	**AB**	**AB**	**AB**
Invariant BB Locus	**BB**	**BB**	**BB**	**BB**

Figure 2.4.1 Classification of loci as variable or invariant in genotype
A locus is termed variable if each of the four genotyping results (AA, AB, BB, and No Call) occur at least once across all samples in a genotyping set. Loci were termed invariant if all samples in the genotyping set had the same genotype call at that locus. There are three types of invariant loci: invariant AA, invariant AB, and invariant BB. No Calls indicate an inability to determine a genotype at that locus. MDGA loci queried that return No Calls for all samples are called uninformative loci. There were no uninformative loci for any of the test sets.

The relationship between the distribution of these invariant and variable loci across the mouse genome was examined using adapted rainfall plot visualizations (Nik-Zainal et al., 2012). SNP loci represented on the MDGA are associated with a genomic location in the genome of the house mouse (*M. musculus*). Variable and invariant loci were plotted with respect to the associated base-pair genomic position along the x-axis. The inter-locus distance is displayed on the y-axis. Trends in the spatial distribution of invariant and variable loci were examined.

SNPs genotyped using the MDGA were analyzed in the context of particular trinucleotide mutational signatures (Nik-Zainal et al., 2012). The R program 'deconstructSigs' (https://github.com/raerose01/deconstructSigs) was utilized to assess possible biases in transitions and transversions with respect to possible adjacent nucleotides that occur over evolutionary time. The mutational signatures of the sample mice were compared to the reference signature plot of the MDGA. The reference displays all nucleotide changes between the A and B allele for all SNP loci represented on the MDGA.

MDGA SNP loci queried in seventeen samples of wild Mus species that have 19 autosomes in a haploid genomic state were compared to the reference house mouse for SNP changes over evolutionary time (Table 2.4.1) (Britton-Davidian et al., 2012; Bryja et al., 2014; Harr, 2006; Ohno et al., 1957; Sharma et al., 1986; Yosida, 1981). An in-house script that plots genotype changes spatially across the genome and temporally over evolutionary time was created (Appendix C). These plots, termed SNP Spatial-Temporal Plots (SNPSTeP), plot the genome position of SNP loci queried by the MDGA on the x-axis, and Mus species are arranged ordinally in increasing divergence time along the y-axis. Each SNP locus queried is represented by a plotted point and each point is coloured

Table 2.4.1 Number of haploid autosomes of Mus study species

Scientific Name	**Common Name**	**Divergence Time[a] from *Mus musculus* (MYD)**	**Haploid Autosome Number**
Mus musculus	House Mouse Reference	0	19
Mus musculus castaneus	Southeastern Asian House Mouse	0.35	19
Mus dunni	Earth-Colored Mouse	6.4	19
Mus famulus	Servant Mouse	6.4	19
Mus fragilicauda	Sheath-Tailed Mouse	6.4	19
Mus caroli	Ryukyu Mouse	7.41	19
Mus cervicolor	Fawn-Coloured Mouse	7.41	19
Mus cookii	Cook's Mouse	7.41	19
Mus platythrix	Flat-Haired Mouse	8.1	12
Mus saxicola	Rock-Loving Mouse	8.1	10 or 11
Mus pahari	Shrew Mouse	8.29	23
Mus nannomys minutoides	African Pygmy Mouse	9.5	8 or 16
Mus nannomys orangiae	Orange Mouse	9.5	No data
Mus nannomys mattheyi	Matthey's Mouse	9.5	17

[a] Divergence times of all Mus species of this study are listed in millions of years from the reference house mouse (*M. musculus*).

according to genotype at the locus. SNPSTeP was created for each of the nineteen autosomes and the X chromosome. The visual representation of changes in genotype was used to identify patterns between chromosomes and along a single chromosome. Visualizing genotype changes allows one to identify if particular genotypes are clustered across the genome or if there is uniform spacing.

3 Results

3.1 The training set of 114 classical inbred mice utilized in training genotyping algorithms lacks the genetic diversity of the sample set

The training set of DNA samples from 114 classical, inbred laboratory mice used in training the genotyping algorithm employed by Affy Power Tools has a maximum genetic distance of approximately 0.225 with respect to the reference C57BL/6J house mouse (Table 3.1.1, Figure 3.1.1). A genetic distance value of approximately 0.225 is over four times smaller in comparison to the maximum genetic distance value of 0.926 from the inter-order genotyping set. The inter-family genotyping set has a maximum genetic distance of 0.930 and the inter-genus genotyping set had a SNP-based genetic distance maximum of 0.924, which are both four times larger than the range of genetic distance covered by the reference set. The maximum genetic distance value of 0.836 for the intra-genus genotyping set of all Mus species is over three times larger than the maximum genetic distance of the reference set of 114 classical inbred mice (Figure 3.1.2).

A Fisher's Exact test revealed that the samples of all test sets are significantly different in genotypic composition and allele frequency ($P<0.0001$) (Appendix A). Two *R. norvegicus* samples were compared to one another and the genotypic composition is not significantly different ($p = 0.0934$). Differences in allelic composition between *R. norvegicus* samples are also not significant ($p = 0.2232$). The four *H. glaber* (naked mole rat) samples genotyped together are significantly different in the genotype composition ($p<0.001$), but not allelic composition ($p<0.0038$).

Table 3.1.1 Summary of maximum, mean, and minimum genetic distances[a] from the house mouse reference sequence for the training and test sets

Training Set and Test Sets	Maximum Divergence Time[b] of Set (MYD[c])	Minimum Genetic Distance from House Mouse	Mean Genetic Distance from House Mouse	Maximum Genetic Distance from House Mouse
114 Classical Inbred Training Set	0	0.004	0.156	0.225
Intra-Genus Genotyping Set	9.5	0.537	0.720	0.836
Inter-Genus Genotyping Set	32.7	0.553	0.780	0.924
Inter-Family Genotyping Set	73	0.542	0.750	0.930
Inter-Order Genotyping Set	96	0.556	0.793	0.926

[a] Genetic distance values were determined by dividing the total number of loci with a genotype difference between two test samples by the total number of loci queried

[b] Divergence times from reference house mouse (*M. musculus*) estimated using TimeTree public knowledge base (www.timetree.org)

[c] MYD = millions of years divergence

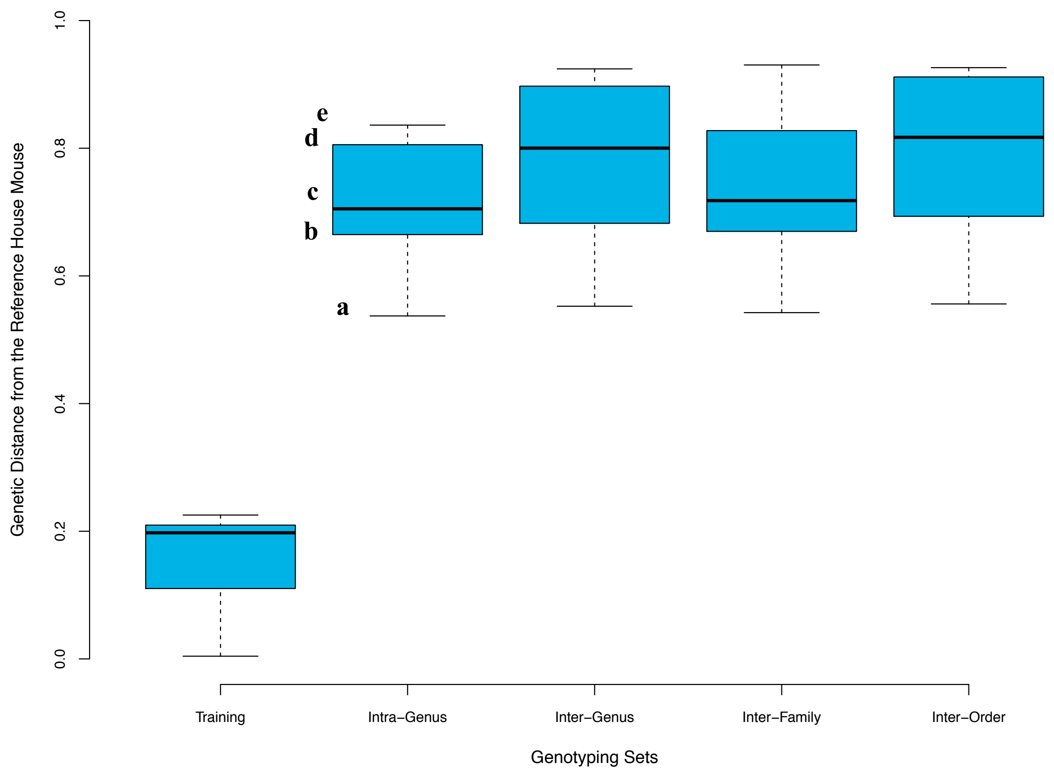


Figure 3.1.1 The distribution of genetic distances from the house mouse for samples in the training and test sets
(a) Minimum genetic distance value of set of samples genotyped. (b) First quartile of genetic distance data of set of samples genotyped. (c) Median genetic distance value of set of samples genotyped. (d) Third quartile of genetic distance data of set of samples genotyped. (e) Maximum genetic distance value of set of samples genotyped. Training set (n = 114) was used to train the genotyping algorithm utilized by Affymetrix Power Tools software. Genetic distance values based on SNP genotypes are plotted for the training set, and four test sets Intra-Genus (n = 27), Inter-Genus (n = 37), Inter-Family (n = 31), and Inter-Order (n = 40).

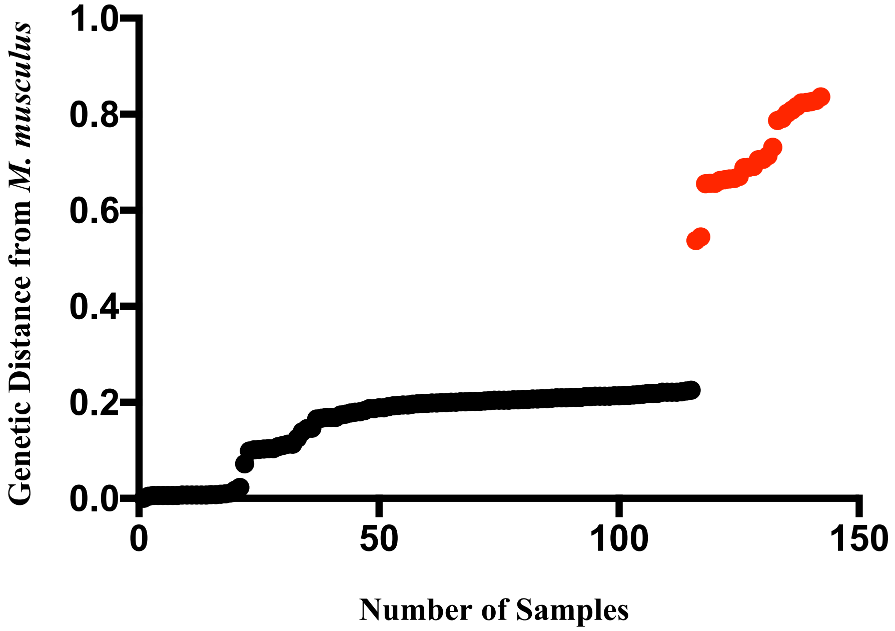


Figure 3.1.2 Genetic distance of the intra-genus test set exceeds the maximum genetic distance of the training set

Each sample in the training set (black dot) is a classical, inbred mouse used to teach the genotyping algorithm what typical genotype results should look like when using the Mouse Diversity Genotyping Array. Each sample in the intra-genus test set (red dot; n=27) is a wild Mus species that is a non-model organism. Genetic distances of the samples in the training set and the intra-genus test set from the reference house mouse *Mus musculus* are displayed. The minimum genetic distance of the intra-genus test set exceeds the maximum genetic distance of the training set.

3.2 Percentage of loci genotyped differs for the same sample depending on the composition of the test set

Genotyping results reveal changes in the percentage of loci that were genotyped (AA, AB, or BB) depending on the number and nature (or composition) of samples included in the test set (Table 2.1.2, Appendix A, Tables A2 & A3). The percentage of loci genotyped for *Diceros bicornis* (African rhino) and *Tapirus pinchaque* (Mountain Tapir; 96 MYD), and Sciuridae (71 MYD) is high (>89% of loci genotyped) when genotyped collectively with Mus samples. Comparatively, the percentage of loci genotyped of four naked mole rats (*Heterocephalus glaber*, 73 MYD) that were analyzed separately in a case study are approximately 44% (Appendix A Table A2, Table 3.2.1). Given that the naked mole rat has an approximate 44% of loci that can be genotyped using the MDGA, the 89% of loci genotyped for a rhino and tapir indicates an issue in test set composition. There are nine MDGA raw data (CEL) files of samples in the genus Mus that have a technical replicate or "redo" file included in the test set. The technical replicate files have an average of 2,130 fewer No Call genotype assignments than the original CEL file.

Table 3.2.1 Percentage of loci genotyped and the percentage of genotyped loci with a heterozygous genotype for samples of the naked mole rat case study

MDGA Data (CEL) File	Sample Scientific Name	Colony Origin of Sample	Sex of Organism	Loci Genotyped (%)	Heterozygosity (%)
DNA3337.CEL	*H. glaber*	Desperado	Male	43.6	27.0
DNA3339.CEL	*H. glaber*	Colony Q	Male	43.9	27.5
DNA3338.CEL	*H. glaber*	Colony Q	Female	44.2	27.7
DNA3340.CEL	*H. glaber*	Colony Q	Female	44.3	27.4

3.3 There is underestimation of genetic diversity in non-Mus samples

For samples in the inter-order test set, a general decrease is observed in the percentage of loci genotyped as divergence time increases from *M. musculus* (r = -0.66; p-value<0.0001) (Figure 3.3.1A). As divergence time increases from *M. musculus*, the number of 'no calls' increases. The percent homozygosity decreases as divergence time from *M. musculus* increases (Figure 3.3.2). There is an approximate 2% difference in the percentage of loci genotyped between the *M. m. castaneus* sample and the technical replicate file. There is a linear negative correlation between percentage of loci with an AA genotype and known divergence times from the house mouse (r = -0.64; p-value<0.0001) (Figure 3.3.2A). A linear negative correlation is observed between the percentage of loci with a BB genotype and known divergence times from the house mouse (r = 0.64; p-value<0.0001) (Figure 3.3.2B). A general decrease in percent homozygosity is followed by a plateau in percent homozygosity for species beyond the genus Mus, beginning between 10-15 million years of divergence (MYD) from the house mouse (Figure 3.3.2). As percent homozygosity decreases in the inter-order test set, percent heterozygosity increases (Figure 3.3.3A). There is a positive correlation between increasing percent heterozygosity and the known divergence times from the house mouse (r = 0.63; p-value<0.0001). A plateau in percent heterozygosity is observed between 10-15 million years divergence from *M. musculus* (Figure 3.3.3A). SNP-based genetic distance increases as divergence time increases, followed by a plateau in SNP-based genetic distance for non-Mus species between 10-15 MYD from the house mouse (r = 0.64; P-value <0.0001) (Figure 3.3.3B). When the maximum divergence time of the test set is reduced from 96 to 73 MYD, a plateau in SNP-based genetic distance is observed

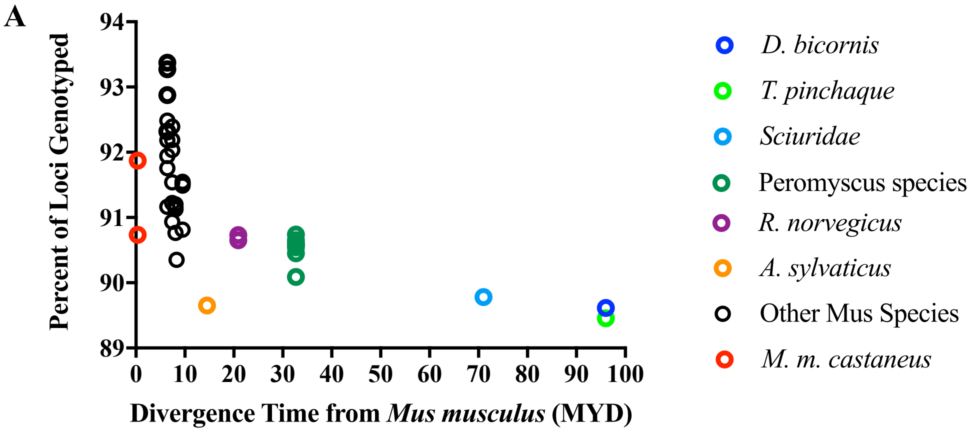


Figure 3.3.1 Percentage of loci genotyped in inter-order test samples with respect to divergence time from the house mouse

Divergence time is listed in millions of years divergence (MYD) from the reference house mouse, *M. musculus* for n = 40 samples of the inter-order test set.

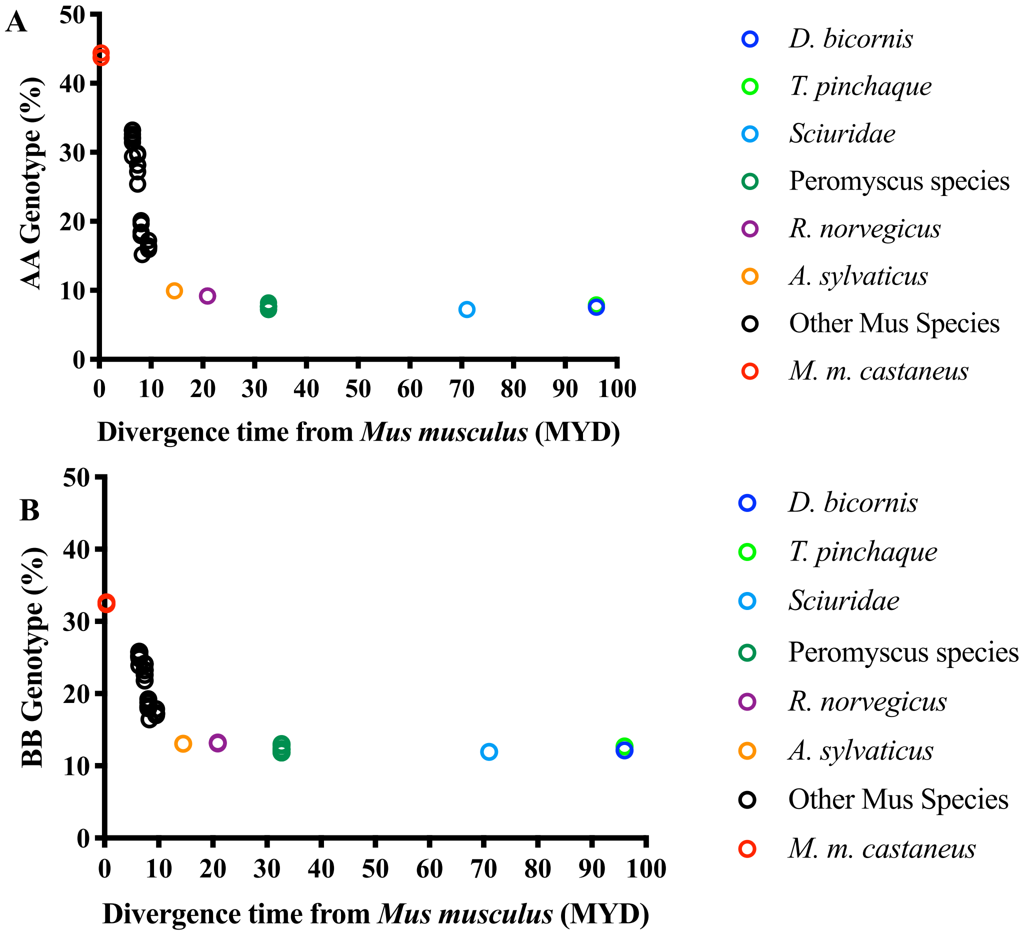


Figure 3.3.2 Percentage of loci with homozygous genotypes for inter-order test samples

(A) The percentage of loci genotyped with a homozygous AA genotype. (B) The percentage of loci genotyped with a homozygous BB genotype. Divergence time is displayed in millions of years (MYD) from the reference house mouse, *M. musculus* for n = 40 samples of the inter-order test set.

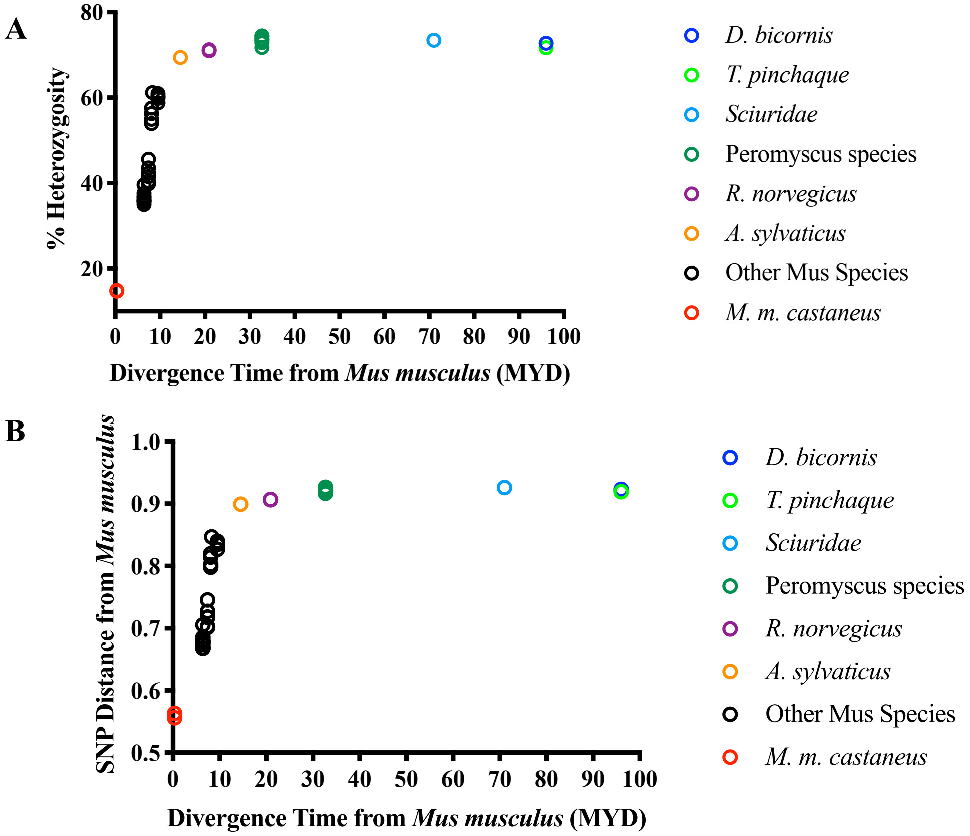


Figure 3.3.3 Percentage of loci with heterozygous genotypes and SNP-based genetic distance values from the reference house mouse for inter-order test samples
(A) Of the loci genotyped for the inter-order test set, the percent of loci with a heterozygous genotype is displayed on the y-axis. (B) The SNP-based genetic distances for samples of the inter-order set with respect to the reference house mouse are displayed. Divergence time is displayed in millions of years (MYD) from the reference house mouse, *M. musculus* for n = 40 samples of the inter-order test set.

for the inter-family test set (n = 31; r = 0.71; p-value<0.0001; Figure 3.3.4A). A plateau in SNP-based genetic distance is observed between 10-15 MYD for non-Mus inter-genus test set samples (n = 37; r = 0.82; p-value<0.0001; Figure 3.3.4B).

The SNP tree of genetic relatedness of samples of the inter-family (n = 31) test set does not distinguish between naked mole rat samples based on their source colony or sex of the organisms (Figure 3.4.1). Naked mole rat samples of the inter-family test set differentiate as pairs, rather than differentiation of samples by the 3:1 ratio of Colony Q to Colony Desperado, respectively (Table 2.2.1). The addition of *H. glaber* samples in the inter-family test set generally increased the genetic distances of the Mus samples in the inter-family set when compared to the genetic distances of the same samples included in the intra-genus all Mus test set (Appendix B, Table A5).

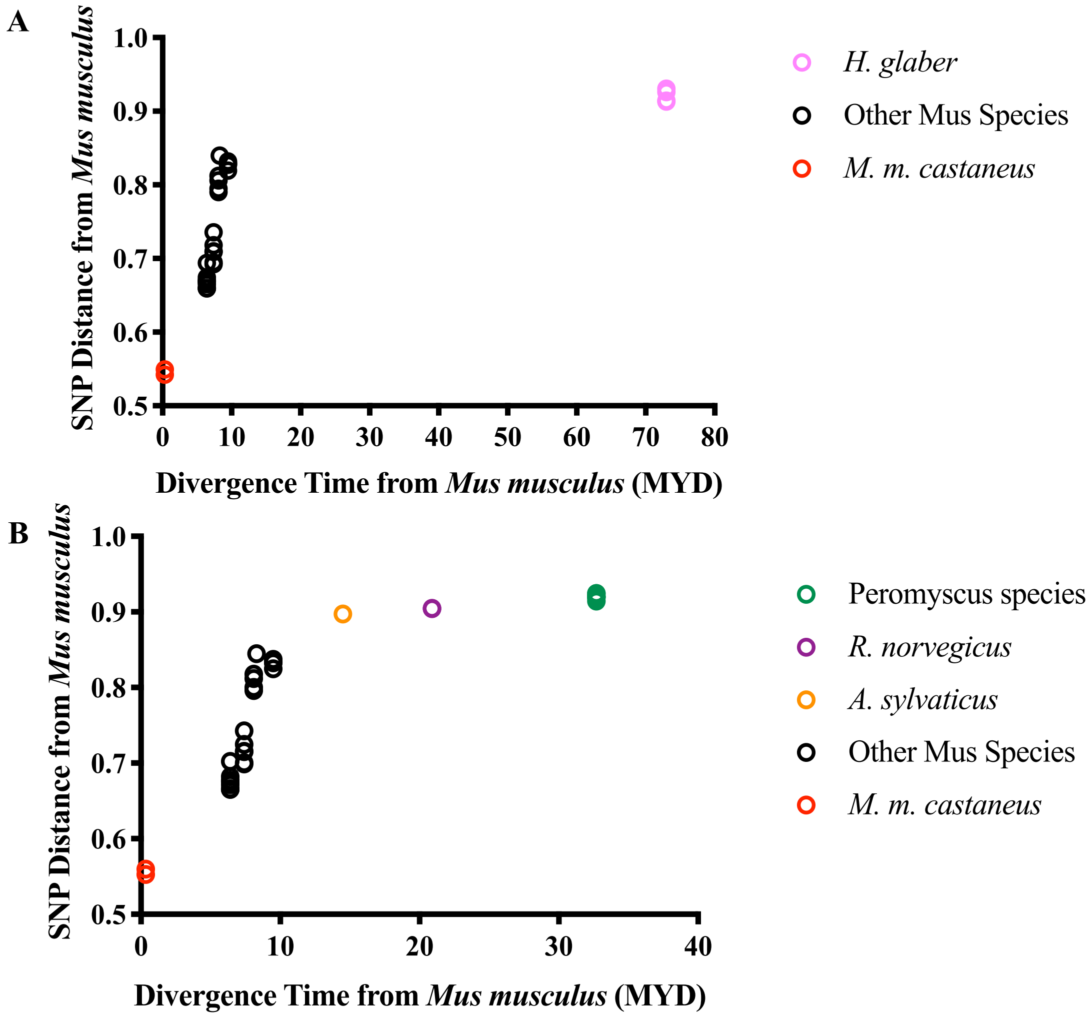


Figure 3.3.4 Increases in SNP-based genetic distance values from the reference house mouse for A) inter-family and B) inter-genus test samples

Divergence time is displayed in millions of years (MYD) from the reference house mouse, *M. musculus* for n = 31 samples of the inter-family test set, and B) n = 37 non-Mus samples of the inter-genus test set.

3.4 Greater divergence time creates challenges when applying the MDGA to naked mole rat samples

In the case study of a test set comprised of only four naked mole rat DNA samples, there are approximately 217,048 loci (~44%) that could be genotyped per sample (Table 3.2.1). Of the 217,048 loci that could be genotyped per sample, there are 91,324 loci that were genotyped as heterozygous in all four naked mole rat samples (Appendix A, Table A7). Of the 91,324 heterozygous loci genotyped using the MDGA, there are 52 loci that were cross-validated as being present in the genome sequence available for the naked mole rat. The 52 loci that were genotyped as heterozygous and mapped to the naked mole rat incomplete genome sequence are potential polymorphic loci that may have utility cross-species. An *in silico* search of the *H. glaber* partial genome available (Table 2.2.2), using the computational program E-MEM identifies only 1,179 MDGA probe sequences that are a perfect and unique match to the partial naked mole rat genome sequence (Appendix E, Online). SNP-based genetic distances of the four *H. glaber* samples do not reflect differentiation by source colonies but are consistent with the sex of the organisms (Figure 3.4.2).

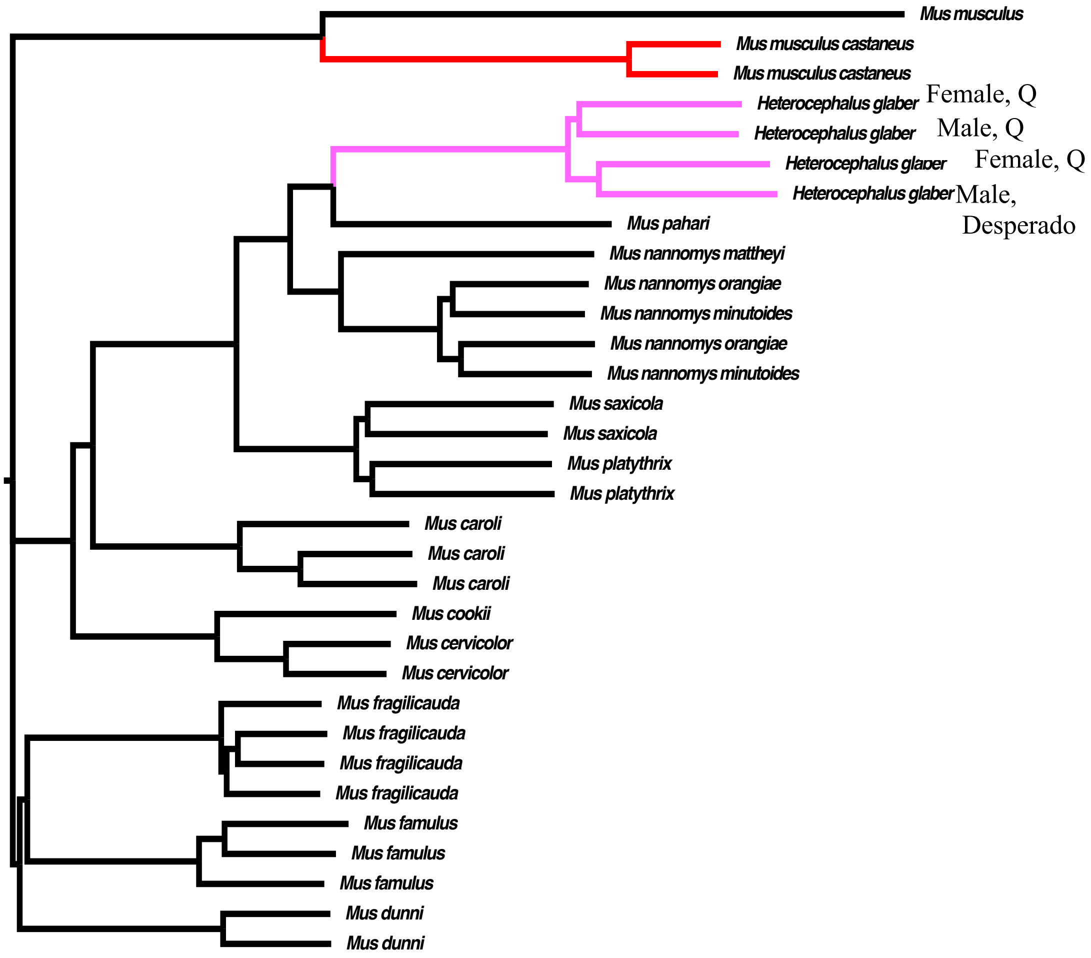


Figure 3.4.1 Inability to differentiate *H. glaber* inter-family test set samples by source colony or sex of the organism

SNP tree of genetic relatedness derived from SNP genotypes of inter-family test set samples (n = 31). This test set contains naked mole rat and Mus samples that were genotyped together. *M. m. castaneus* nodes are coloured red to emphasize the species as being the most closely related to the reference house mouse, *M. musculus*. Naked mole rat samples are coloured pink, and the sex and source colony of the organisms are noted on the tree.

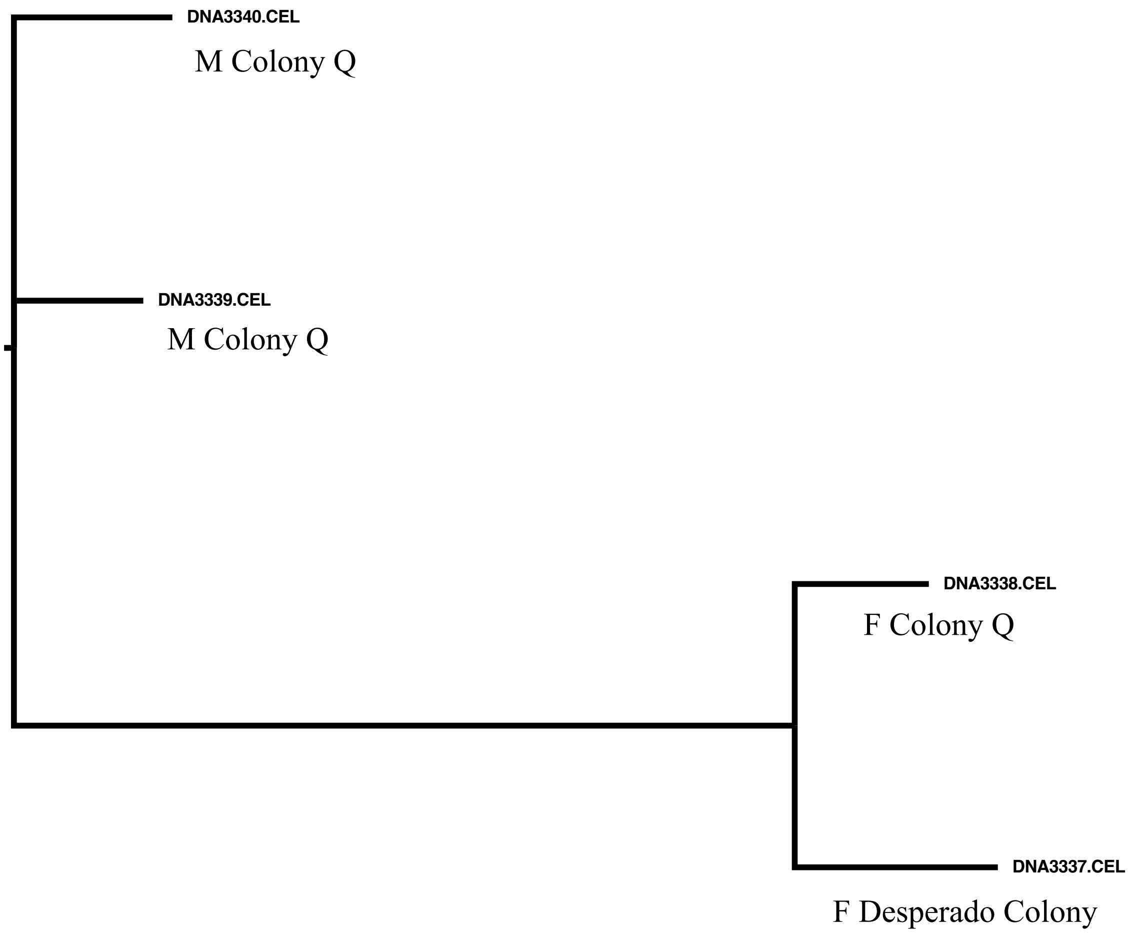


Figure 3.4.2 Differentiation of four naked mole rat samples by sex of the organism in case study using genetic distance values derived from SNP genotypes
Intra-specific genetic differentiation of four naked mole rat samples derived from SNP loci genotypes. Sex and source colony are indicated where F denotes female naked mole rat samples and M denotes male naked mole rat samples. Naked mole rat samples were genotyped together as a separate test set. Genetic distance measures of the four naked mole rat samples differ from the test set where they were genotyped in isolation in comparison to the inter-family test set, where the four naked mole rat samples were genotyped with 27 Mus samples.

3.5 SNP-based genetic distances reflect known taxonomy for Mus species but not for the subspecies analyzed

The Mus samples of the intra-genus test set (n = 27) have a 3.44% average decrease in the percentage of loci genotyped compared to the same Mus samples when they are included in the inter-order test set (Appendix A, Table A4). There is a decrease in loci genotyped between the same Mus samples when included in the two different test sets. There is an increase in loci genotyped for the two *M. m. castaneus* samples included in the inter-order and intra-genus test sets (Appendix A, Table A4). The number of loci genotyped decreases an average of 1.46% for loci genotyped in Mus samples in the inter-family test set compared to those of the intra-genus test set (Appendix A, Table A5). There is an increase in loci genotyped for the two *M. m. castaneus* samples when included in the inter-family test set compared to the same two samples in the intra-genus test set (Appendix A, Table A5).

In the intra-genus test set, homozygosity decreases as divergence time increases (Figure 3.5.1). There is a strong linear negative correlation between the decrease in homozygous AA genotyped loci with divergence time from the house mouse (r = -0.90; p-value<0.0001) (Figure 3.5.1A). The decrease in homozygous BB loci for Mus samples is negatively correlated with divergence time from *M. musculus* (r = -0.91; p-value<0.0001) (Figure 3.5.1B). The increase in percent heterozygosity of Mus samples is positively correlated with an increase in divergence times (r = 0.93; p-value<0.0001) (Figure 3.5.2A). There is a strong positive correlation between calculated SNP-based genetic distances from the house mouse and divergence time from *M. musculus* (r = 0.90; p-value<0.0001) (Figure 3.5.2B). A tree of relatedness derived from SNP-based genetic

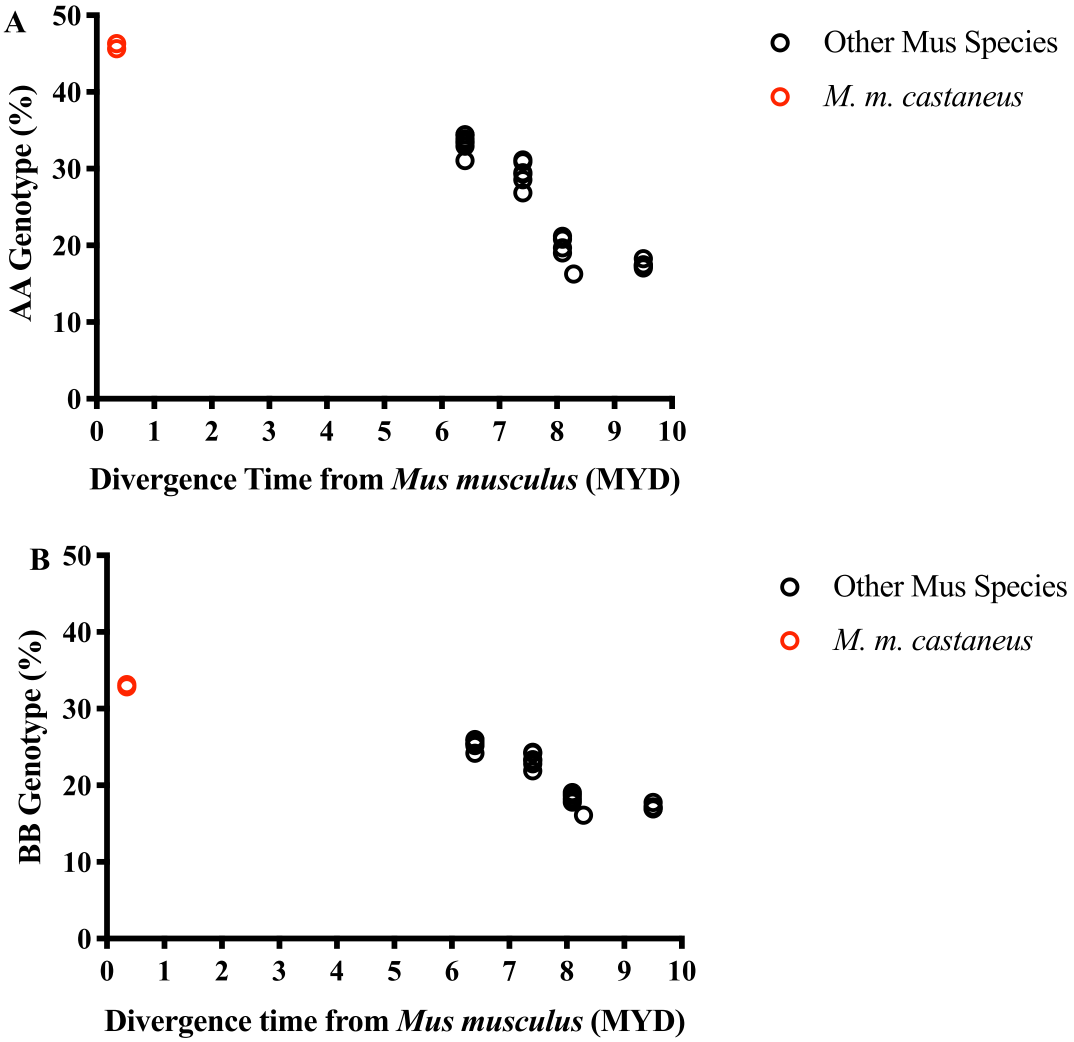


Figure 3.5.1 Decrease in homozygosity as divergence time increases for Mus samples of the intra-genus test set

(A) For samples of the intra-genus test set (n=27), of the loci that could be genotyped, the percentage of loci with a homozygous AA genotype for each sample is displayed. (B) of the loci that could be genotyped, the percentage of loci with a homozygous BB genotype for each sample is displayed. For Mus samples of the intra-genus genotyping set (n = 27), divergence time is displayed in millions of years (MYD) from the reference house mouse, *M. musculus*.

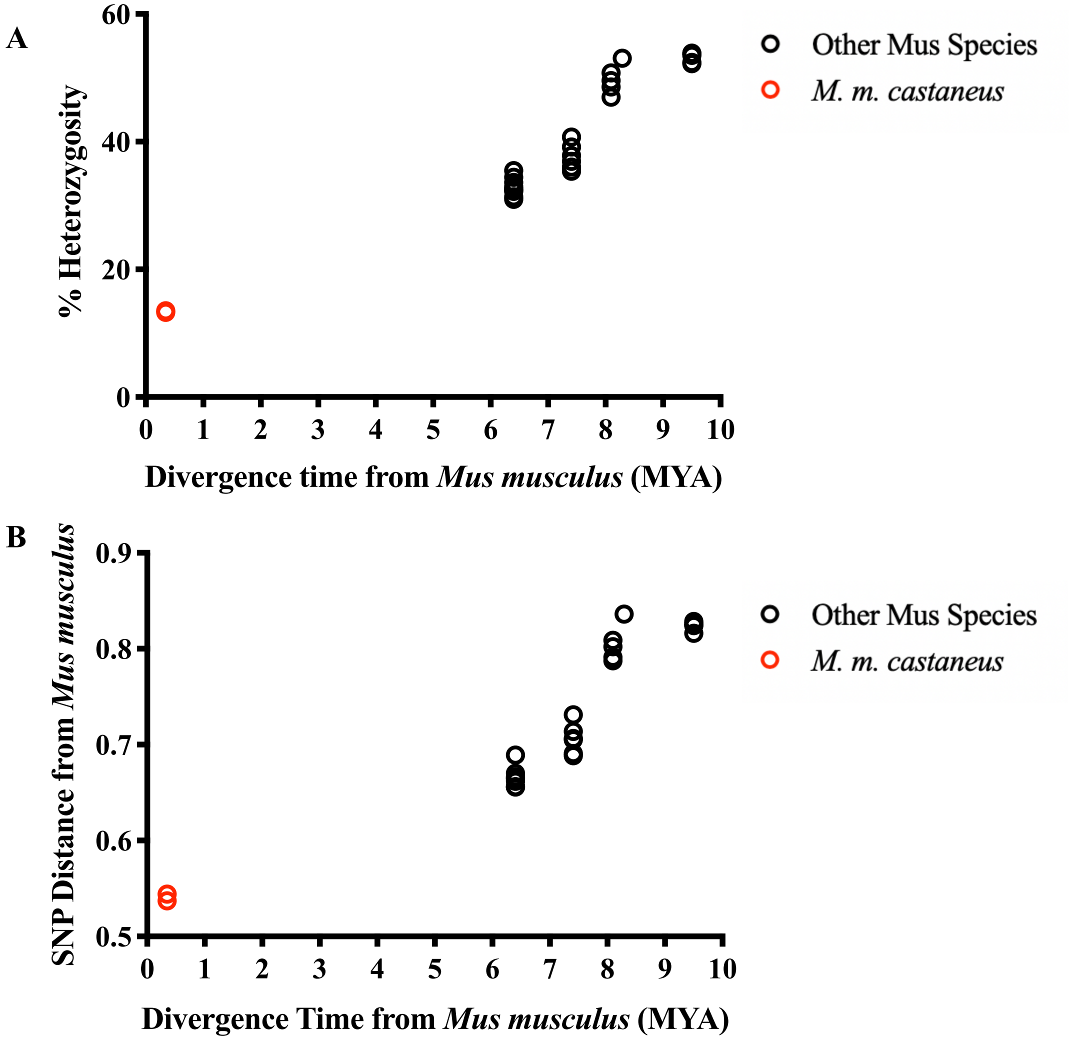
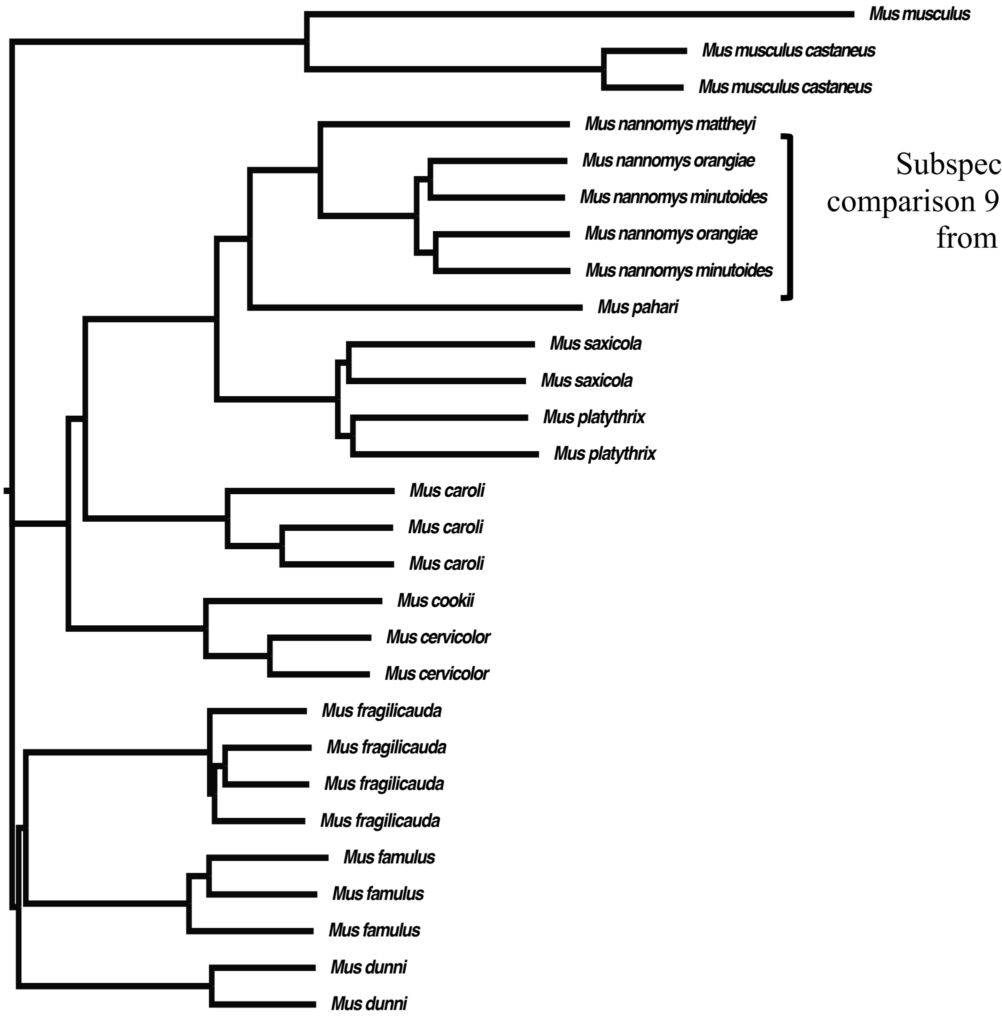


Figure 3.5.2 Heterozygosity and SNP-based genetic distance increase with divergence time for Mus samples in the intra-genus test set
(A) Of the loci that could be genotyped for the 27 sample intra-genus test set, the percentage of loci with a heterozygous genotype is displayed. (B) SNP distances of the non-model Mus samples from the reference house mouse are displayed on the y-axis. For samples of the intra-genus genotyping set (n = 27), divergence time is displayed in millions of years (MYD) from the reference house mouse, *M. musculus*.

distance values differentiates Mus samples of the intra-genus test set from one another at a species level (Figure 3.5.3). At 9.5 MYD, the pygmy mouse subspecies *M. n. minutoides* is grouped with the subspecies *M. n. orangiae* and not the "redo" data file of the same species (Figure 3.5.3).



Subspecies
comparison 9.5 MYD
from

Figure 3.5.3 Mus species, but not subspecies, in the intra-genus test set are differentiated according to known genetic relatedness by genetic distance values obtained from MDGA genotyped loci

SNP-based genetic distance tree of relatedness of samples from the intra-genus test set (n = 27). At 9.5 MYD a pygmy mouse subspecies *M. n. orangiae* has SNP-based genetic distances that reflect greater genetic similarity to another pygmy mouse subspecies *M. n. minutoides* than the redo MDGA data file of the same *M. n. orangiae* sample.

3.6 Variable and invariant genotype loci are clustered in specific regions of the mouse genome

Of the 493,290 loci queried by the MDGA, there are 24,331 loci considered variable, and the corresponding genomic positions are located more densely across autosomes (1-19) of the mouse genome (Figure 3.6.1). Diploid genotypes on X chromosome apply only to female samples (n=2) in 27 Mus samples. Hemizygous genotypes identified on the X chromosome of male mice are assigned AA or BB by the genotyping algorithm, despite being haploid. Only 256 of the 18,578 loci located on the X chromosome of the mouse reference genome are variable genotype loci, corresponding to approximately 1.4% of all loci located on the X chromosome. MDGA genotyped loci with a heterozygous genotype in all 27 samples are deemed invariant heterozygous loci. There are 1,307 loci that share a heterozygous genotype across 27 Mus samples. Invariant heterozygous loci are scattered across the 19 autosomes of the reference *M. musculus* genome (Figure 3.6.2). No invariant heterozygous loci are found on the X chromosome. Of the 493,290 MDGA loci queried, there are 2,412 loci that are considered invariant homozygous AA. Approximately 20%, or 485 loci termed invariant AA loci are located on the X chromosome, with the remaining 80% spread across the 19 autosomes (Figure 3.6.3). There are 1,736 loci genotyped of the 493,290 total loci queried that were termed invariant homozygous BB loci, with 284 of these loci, or 16%, located on the X chromosome (Figure 3.6.4).

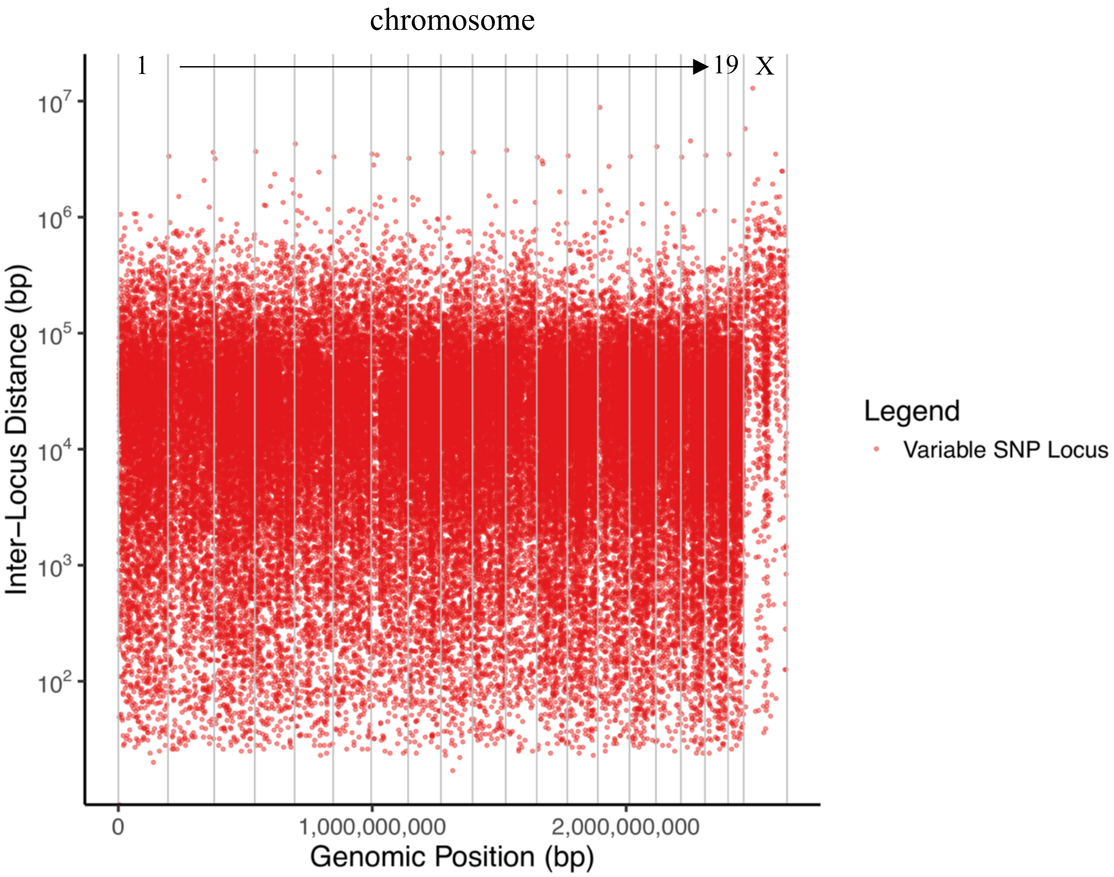


Figure 3.6.1 Rainfall plot of the inter-locus spacing for *M. musculus* loci that are variable in genotype across 27 Mus samples
Chromosome number is indicated above the plot. True diploid genotypes on X chromosome apply only to female samples (n=2) in 27 Mus samples. Hemizygous genotypes determined on the X chromosome of male mice are considered AA or BB, despite not being diploid.

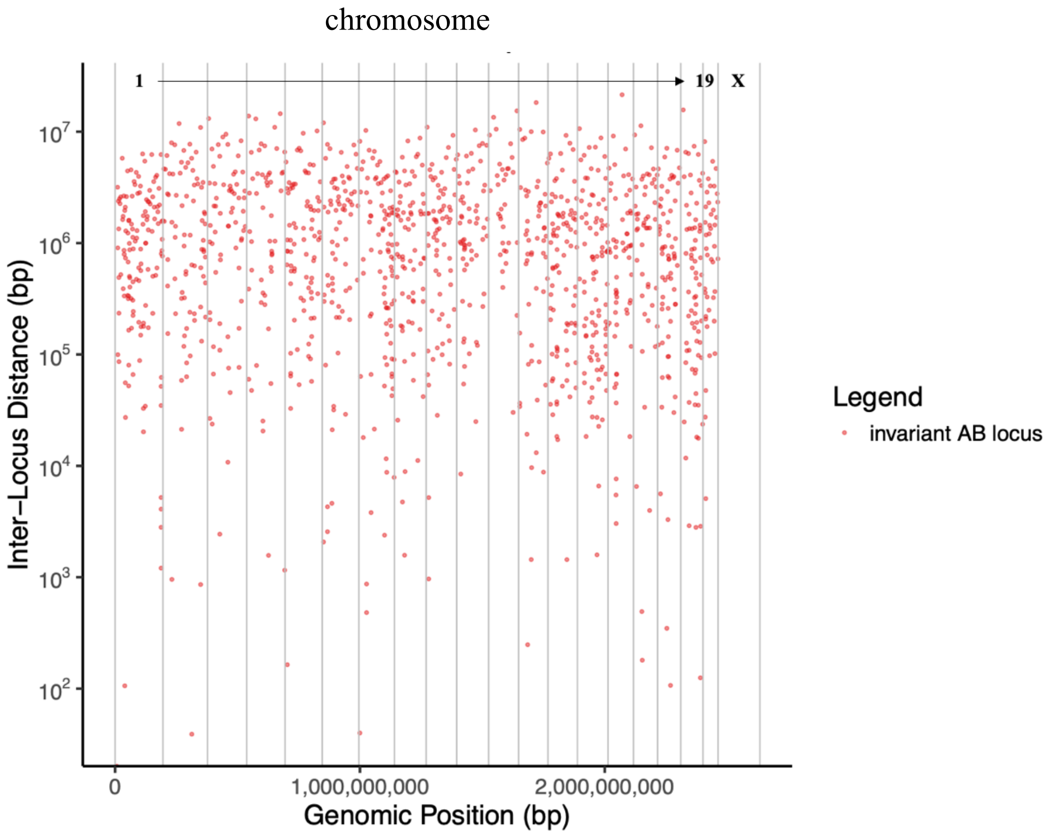


Figure 3.6.2 Rainfall plot of the inter-locus spacing for *M. musculus* loci that are invariant AB across 27 Mus samples
Chromosome number is indicated above the plot. True invariant AB genotypes on X chromosome apply only to female samples (n=2) in 27 Mus samples.

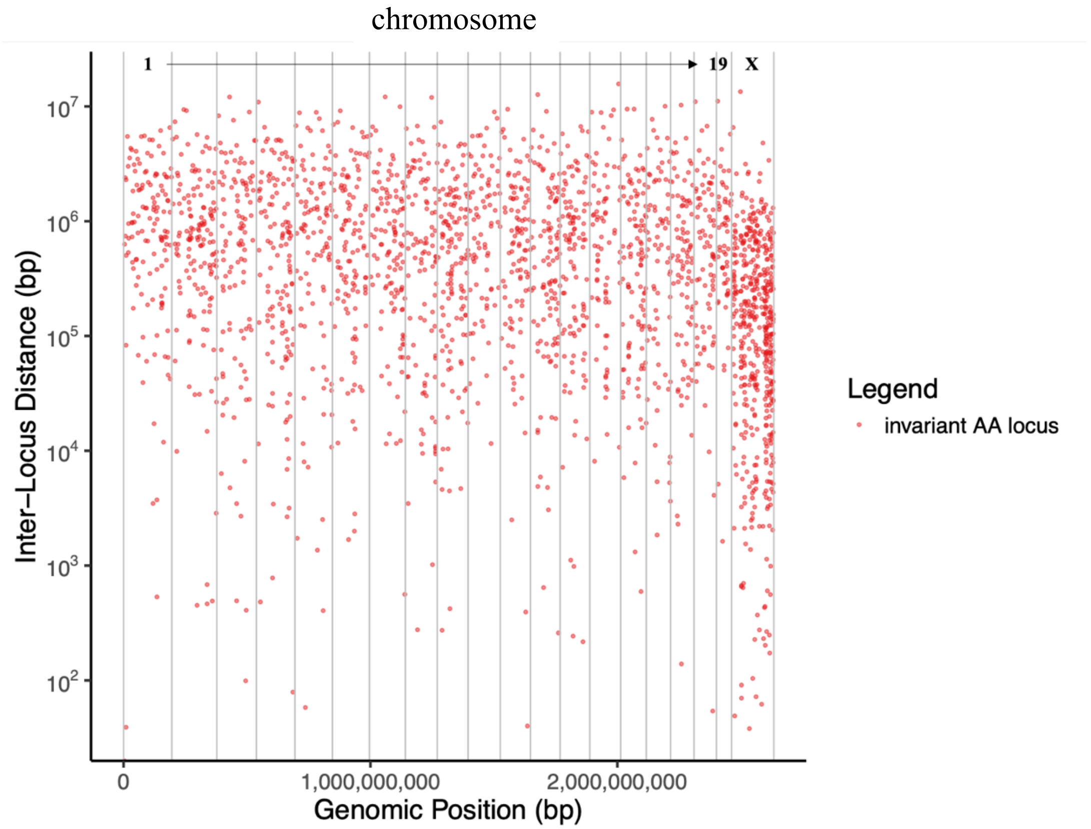


Figure 3.6.3 Rainfall plot of the inter-locus spacing for *M. musculus* loci that are invariant AA across 27 Mus samples
Chromosome number is indicated above the plot. True invariant AA genotypes on X chromosome apply only to female samples (n=2) in 27 Mus samples. Hemizygous genotypes determined on the X chromosome of male mice are considered AA or BB, despite not being diploid.

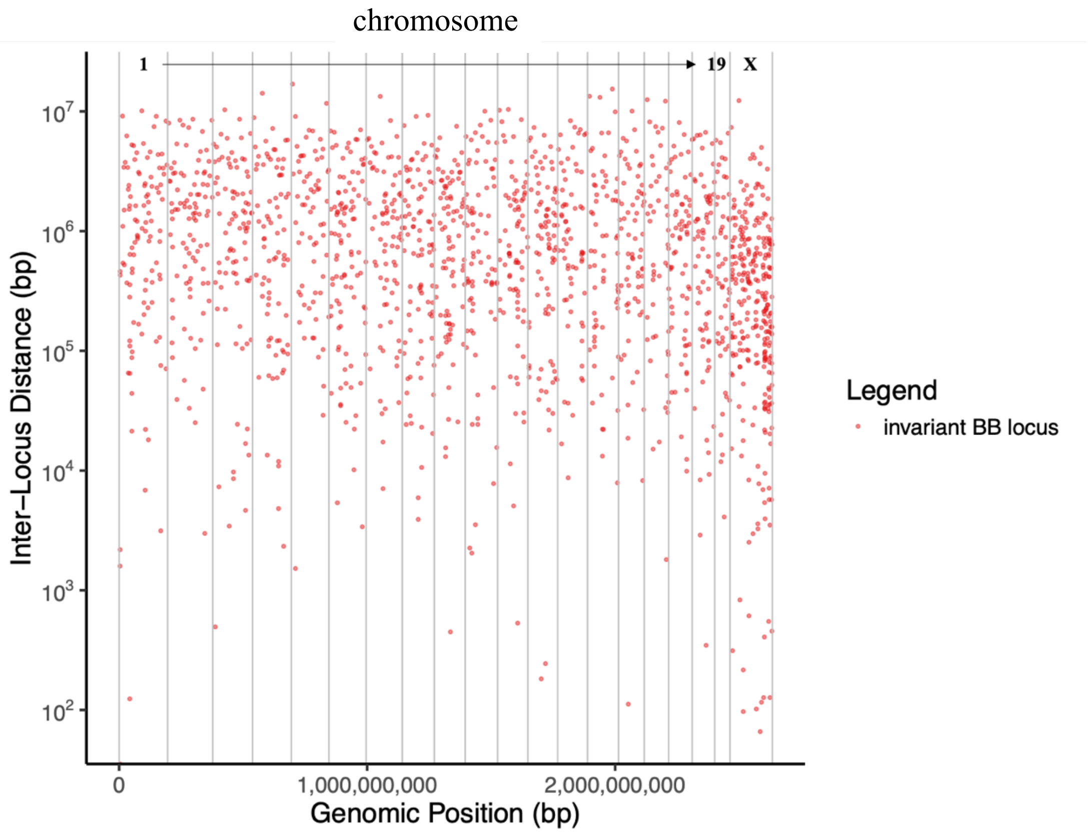


Figure 3.6.4 Rainfall plot of the inter-locus spacing for *M. musculus* loci that are invariant BB across 27 Mus samples

Chromosome number is indicated above the plot. True invariant BB genotypes on X chromosome apply only to female samples (n=2) in 27 Mus samples. Hemizygous genotypes determined on the X chromosome of male mice are considered AA or BB, despite not being diploid.

3.7 Mutational signature qualitative analysis reveals bias for transitions in mice

The Mus samples of the intra-genus genotyping set (n = 27) were analyzed for qualitative trends or changes in the 96 possible trinucleotide mutational signatures (Figure 3.7.1). SNPs targeted by the MDGA are biased for T>C and C>T transitions. Across all Mus samples there is a similar bias for T>C and C>T transitions.

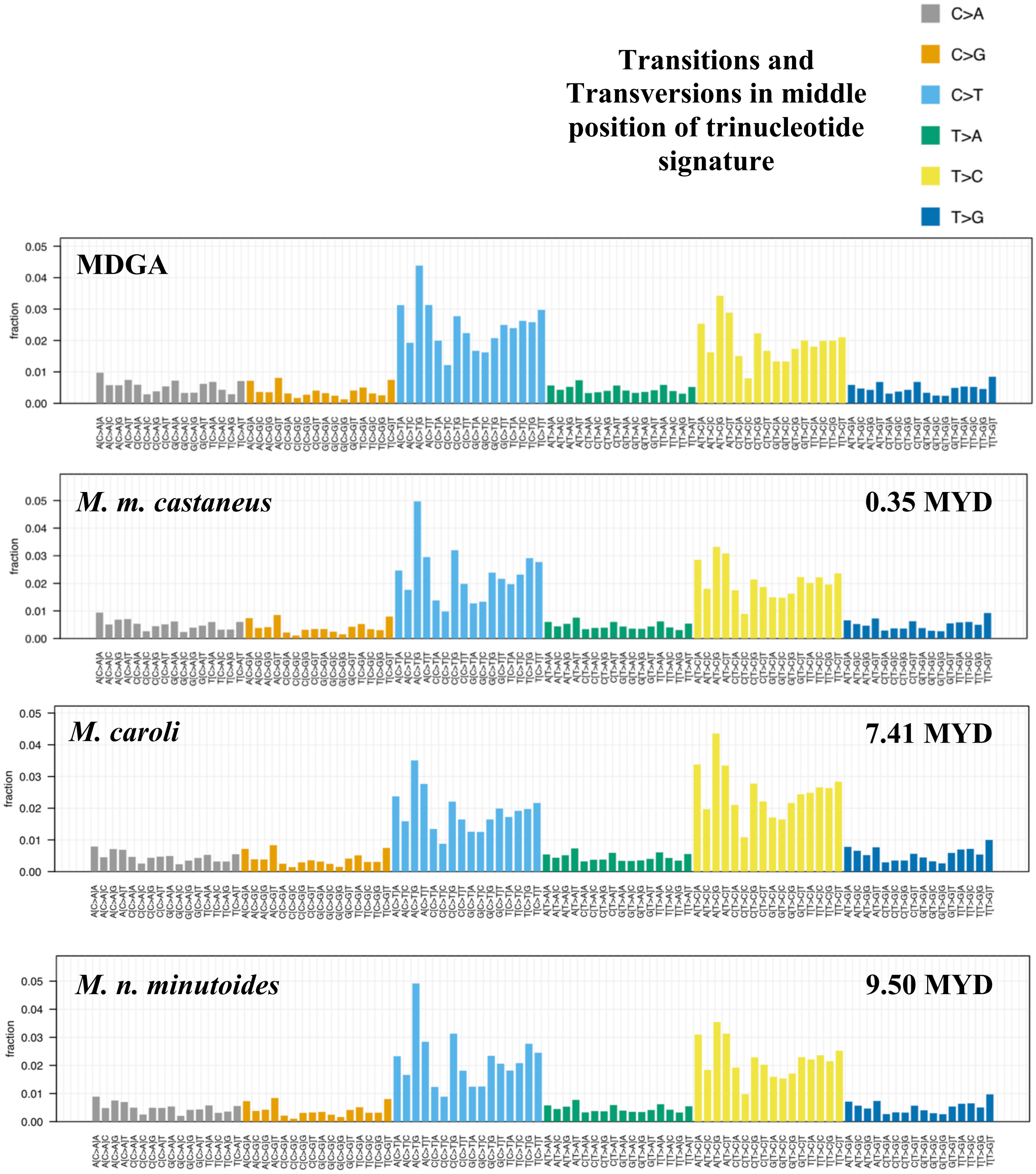


Figure 3.7.1 The mutation signatures of three Mus species, *M. m. castaneus M. caroli*, and *M. n. minutoides* of the intra-genus test set

(Top) This mutation signature plot shows the relative proportions of the base substitutions detectable by the probe sequences on the Mouse Diversity Genotyping Array. Divergence time listed in millions of years from the reference house mouse, *M. musculus*. As an example, the notation A[C>G]A indicates a transversion of a C to a G in the context of A nucleotides at the 5' and 3' locations (5' NNN 3').

3.8 Visualization of SNP genotype changes cross-species

Of the 27 Mus samples analyzed from the intra-genus test set, 17 samples have 19 autosomal chromosomes like the reference *M. musculus* genome on which the MDGA was designed. The 17 samples that contain 19 chromosomes were analyzed both spatially across each of the 19 chromosomes, and temporally across the maximum 7.2 million years divergence from the reference *M. musculus*. This analysis design is named the SNP Spatial-Temporal Plot (SNPSTeP). In viewing the X chromosome as an example, at a chromosomal view of single nucleotide variation along this chromosome, a large expanse of heterozygosity can be seen in the central region of the chromosome within both *M. m. castaneus* samples when compared to the reference house mouse and the wild Mus species (Figure 3.8.1). There is a change in SNP genotypes moving from the reference house mouse with a majority of homozygous AA genotypes to wild Mus species with a greater number of homozygous BB genotypes and heterozygous loci. In contrast to the X chromosome, genotypes across chromosome 19 are much more variable (Figure 3.8.1). Greater resolution of the X chromosome and chromosome 19 through a visual window of 5421 loci enables the identification of more subtle patterns of genotype changes between different Mus species (Figure 3.8.2).

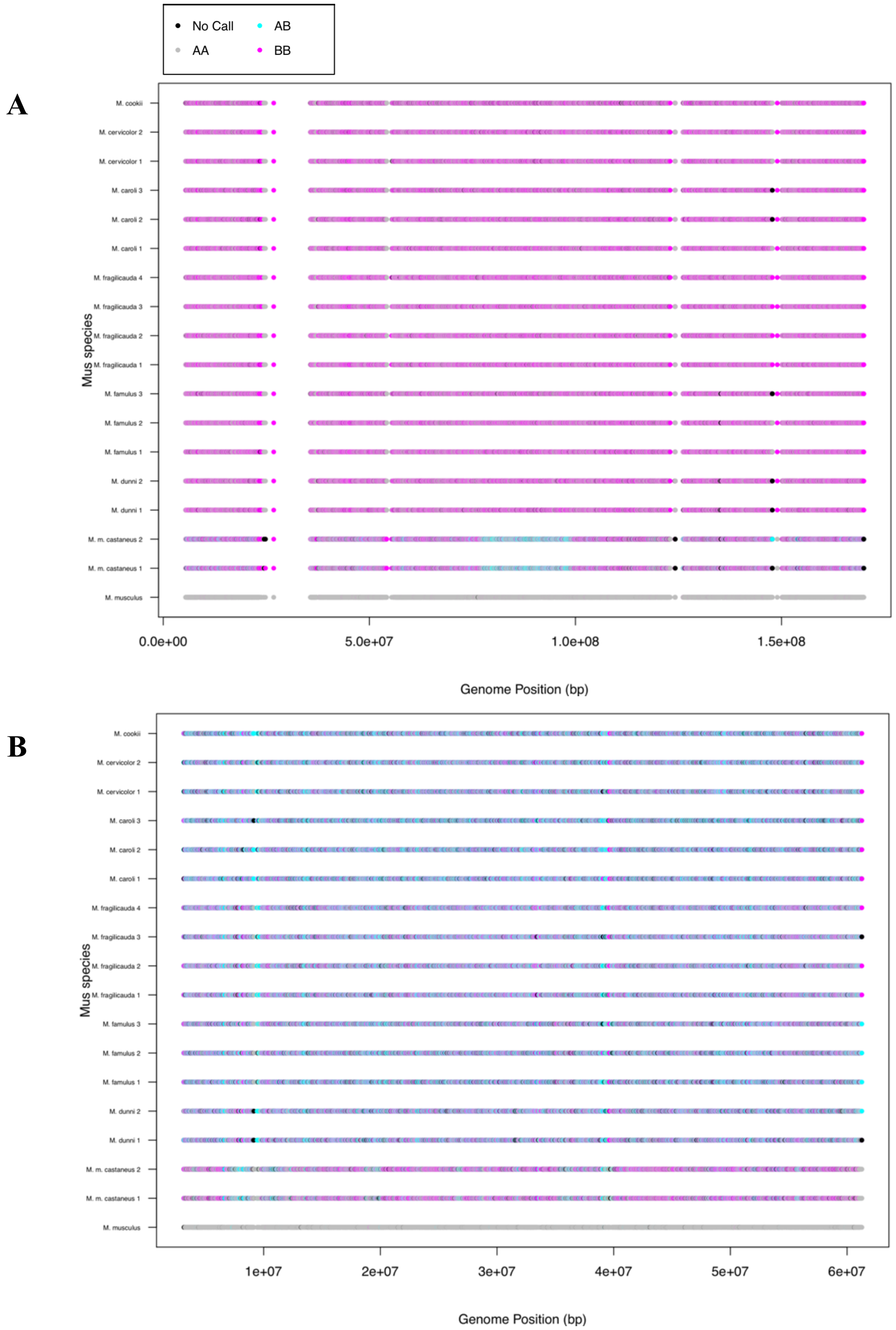


Figure 3.8.1 Visualization of genotype changes cross-species in the context of *M. musculus* chromosomes X and 19

SNP genotype changes across 17 wild Mus samples compared to the reference *M. musculus* on (A) the X chromosome and (B) chromosome 19. *M. m. castaneus* only female samples (n = 2) in X chromosome analysis.

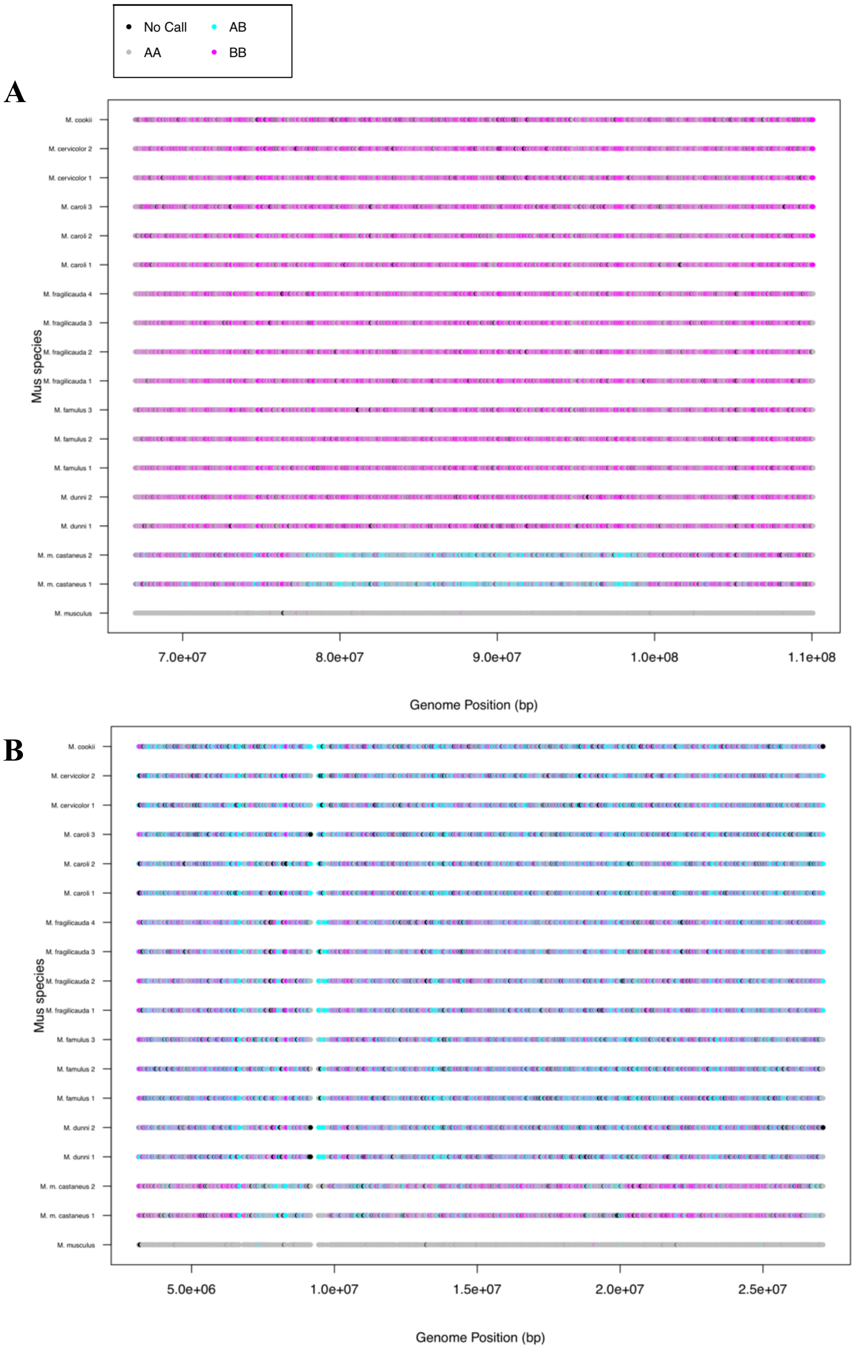


Figure 3.8.2 Localized visualization of genotype changes cross-species in the context of *M. musculus* chromosomes X and 19

SNP genotype changes across 17 wild Mus samples compared to the reference *M. musculus* at 5421 loci on (A) the X chromosome and (B) chromosome 19. *M. m. castaneus* only female samples (n = 2) in X chromosome analysis.

3.9 Successful differentiation of deer mouse samples based on known divergence times

Seven Peromyscus species were genotyped in isolation from Mus species, with 159,797 loci genotyped across all seven samples (32% of loci queried by the array) despite a 32.7 million year divergence time from *M. musculus* (Table 3.9.1). *P. maniculatus* was examined as there is a partial genome sequence available online for *in silico* search of unique and perfect 25 nt MDGA probe target sequence matches. There are 226,265 loci on the MDGA genotyped (~52%) for both *P. maniculatus bairdii* and *P. maniculatus sonoriensis* within this study (Table 3.9.1). Of these loci that were genotyped, there are 143,971 loci that were genotyped as heterozygous in both *P. maniculatus* samples (Appendix A, Table A7). There are 6,076 MDGA probe sequences that perfectly match a unique position within the *P. maniculatus* genome (Appendix D Online), and 481 of the *in silico* sequence matches are associated with heterozygous loci (Appendix A, Table A7). When comparing *in silico* and experimental results, 3,195 sequences were found to be both empirically genotyped using the MDGA and theoretically present in the genome (Appendix D Online). There are 1,909 mouse Ensembl gene ID matches associated with the list of 3,195 consensus sequences present theoretically and empirically for the subspecies of *P. maniculatus* (Appendix D Online). Among the top functional associations found utilizing DAVID are neurological signaling pathways and circadian entrainment (p-value<0.001) (Table 3.9.2). Despite 32.7 million years divergence from the house mouse, SNP-based genetic distances of Peromyscus species could be utilized to build trees of genetic relatedness that reflect the known divergence times of these species (Figure 3.9.1). The ability to accurately differentiate and group Peromyscus species

Table 3.9.1 Percentage of loci genotyped and percentage of genotyped loci with a heterozygous genotype for the case study of deer mouse species

MDGA Data (CEL) File	Sample Scientific Name	Loci Genotyped (%)	Heterozygosity (%)
SNP_mDIV_B2-660_102109.CEL	*P. melanophrys*	51.31	34.83
SNP_mDIV_B1-659_102109.CEL	*P. aztecus*	52.03	36.02
SNP_mDIV_B3-661_102109.CEL	*P. californicus*	52.13	36.27
SNP_mDIV_B4-662_102109.CEL	*P. m. sonoriensis*	52.26	35.95
SNP_mDIV_B5-663_102109.CEL	*P. m. bairdii*	52.27	36.71
SNP_mDIV_B6-664_102109.CEL	*P. polionotus*	52.57	37.02
SNP_mDIV_B8-666_102109.CEL	*P. leucopus*	52.62	36.55

Table 3.9.2 Top associated house mouse pathways with MDGA probe matches to the *P. maniculatus* genome

KEGG Pathway[a]	**p-value**
Glutamatergic synapse	1.24E-08
Circadian entrainment	5.11E-08
Axon guidance	2.03E-06
Retrograde endocannabinoid signaling	9.50E-06
Dopaminergic synapse	1.36E-05
Morphine addiction	1.15E-04
Long-term depression	2.24E-04
Hippo signaling pathway	2.51E-04
cAMP signaling pathway	2.90E-04
Cholinergic synapse	3.80E-04
Rap1 signaling pathway	4.39E-04
Long-term potentiation	4.82E-04
GABAergic synapse	6.16E-04

[a] Top KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways determined using the DAVID functional annotation tool

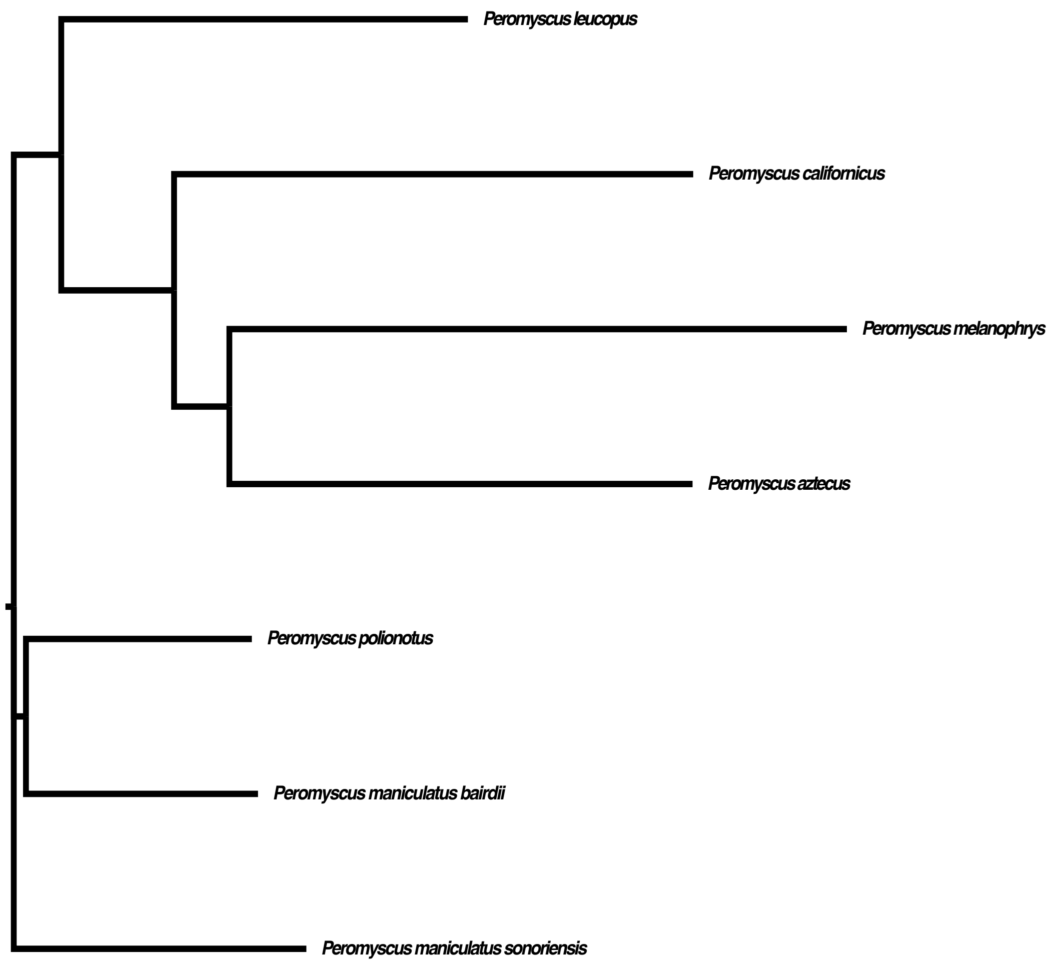


Figure 3.9.1 Intra-genus SNP tree of relatedness based on SNP-based genetic distance values between seven Peromyscus samples

Peromyscus samples were genotyped separately from other publicly available data.

using genotyping results from the MDGA is possible with the addition of non-Peromyscus samples (Figure 3.9.2). Pairwise genetic distances between Peromyscus samples in the intra-genus set genotyped separately are in the approximate range of 0.025 and smaller (Appendix B, Online). The genetic distance values of Peromyscus samples are 0.16 and higher when genotyped with other non-Peromyscus samples in the inter-genus test set (Appendix B, Online).

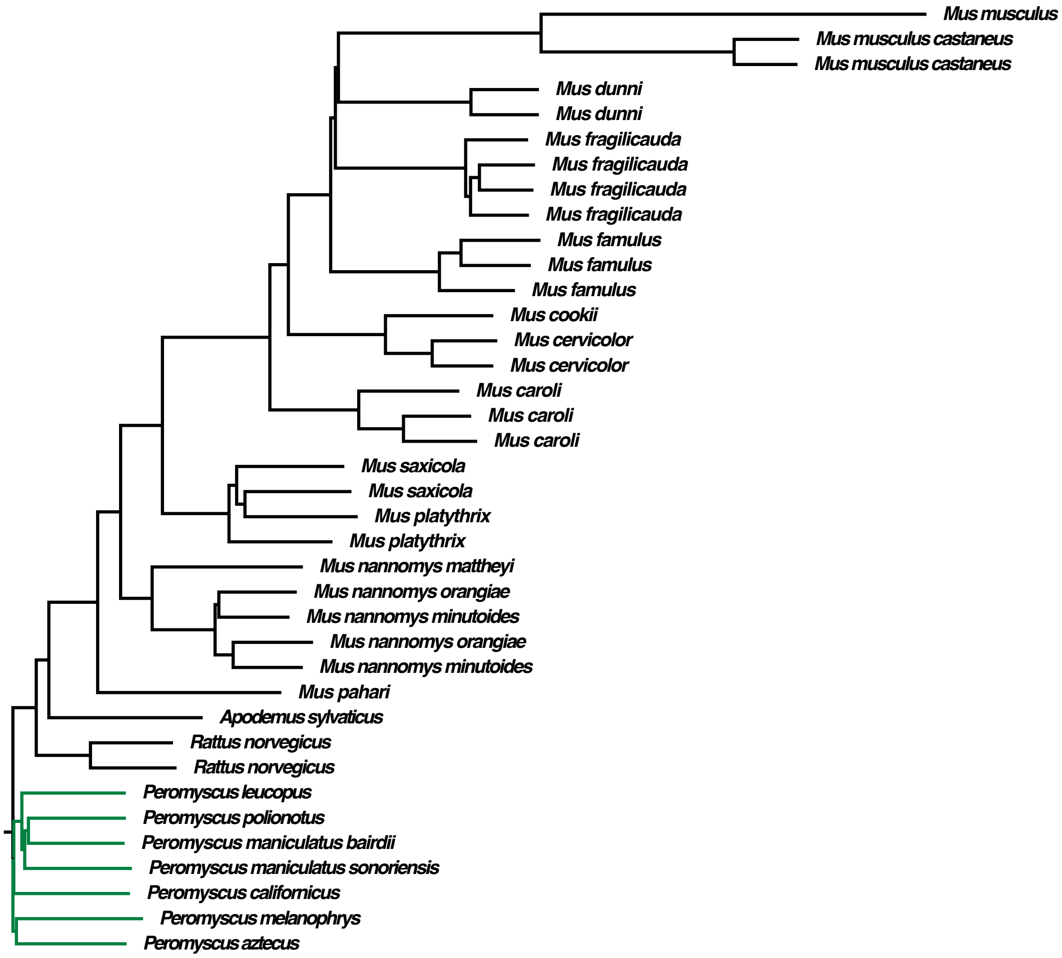


Figure 3.9.2 Tree of relatedness created from SNP-based genetic distance values for samples in the inter-genus test set differentiates Peromyscus species from Mus species

SNP-based genetic distance tree of relatedness for samples genotyped in the inter-genus test set (n = 37).

3.10 Theoretical matches to publicly available wild rodent genomes reveal fewer unique matches when compared experimental genotype results

An average of 382,968 loci that were genotyped between three available *M. caroli* CEL files using the MDGA, and there are 303,680 unique theoretical matches to the *M. caroli* genome determined through an *in silico* search using E-MEM. Of the possible theoretical and experimentally determined matches, there are 161,149 loci on the MDGA that are determined to be present in all three *M. caroli* samples using the MDGA and were determined to be theoretically present in the genome. A shrew mouse (*M. pahari*) applied to the array has 411,514 loci that were genotyped experimentally using the MDGA. Theoretically, there are 152,970 unique sequences from the MDGA that are present in the shrew mouse only once (Appendix D Online). There are 67,820 loci that are genotyped experimentally using the MDGA and were found to be theoretically present within the online *M. pahari* genome resource using the E-MEM program (Appendix D Online). The Sprague Dawley rat (*R. norvegicus*) has a fully sequenced and annotated genome available online. There are 170,156 loci that were genotyped experimentally in both *R. norvegicus* samples using the MDGA. Using the E-MEM *in silico* program, 61,372 sequences were determined to be theoretically present within the genome (Appendix D Online). There are 11,582 sequences that match theoretically to the rat genome and were genotyped using the array (Appendix D Online).

Special attention was given to potential polymorphic loci that were genotyped as heterozygous in samples using the MDGA and could be cross-validated as being present in the genome using an *in-silico* search of publicly available genomes. There is a trend of there being more heterozygous loci genotyped using the MDGA than the number of those loci that can be cross-validated as being present in the publicly available genomes (Appendix A, Table A7). There are 147,452 heterozygous loci genotyped in all three *M. caroli* samples, and 9,413 of these loci were validated as being present in the publicly available genome. There are 9,341 of the 147,452 heterozygous loci genotyped in a *M. pahari* sample that were cross-validated as potential polymorphic SNP loci. In two *R. norvegicus* samples, there are 85,926 loci that were genotyped empirically using the MDGA, and 1,019 loci that were cross-validated using an *in-silico* genome search.

3.11 Functional associations for SNP loci genotyped in wild rodent samples that are also present in available genome assemblies

MDGA loci that were genotyped using the MDGA in wild rodent samples and have had associated probe sequences confirmed as being present within publicly available genomes are candidate loci that may represent conserved SNP loci between *M. musculus* and Mus samples. Candidate loci were analyzed for associated Ensembl mouse reference gene identifiers (IDs). MDGA candidate SNP loci with an associated gene ID were placed as a gene list within the Database for Annotation, Visualization, and Integrated Discovery (DAVID 6.8). The MDGA has a total of 116,217 loci with an associated Ensembl mouse gene ID. The KEGG pathways enriched for all 493,290 SNP loci on the MDGA include olfactory transduction, neuroactive ligand-receptor interaction, and Mucin type O-glycan biosynthesis ($p<0.001$) (Appendix E Online). *M. caroli* and *M. pahari* test samples have a publicly available genome and of the top KEGG pathways ($p<0.001$) associated with these samples, there are 15 pathways that are shared between the current build 38 of the house mouse genome, and the two wild Mus species (Table 3.11.1, Appendix E Online).

Table 3.11.1 Top KEGG pathways enriched for house mouse gene annotations with genotype assignments across wild Mus and Rattus species

KEGG pathways[a] significant (p<0.001) in reference house mouse (build 38) and wild Mus test samples[b]	KEGG Pathways significant (p<0.001) in Mus and Rattus test samples[c]
Focal adhesion	Focal adhesion
Rap1 signaling pathway	cAMP signaling pathway
Adherens junction	ErbB signaling pathway
cAMP signaling pathway	Neuroactive ligand-receptor interaction
ErbB signaling pathway	Calcium signaling pathway
cGMP-PKG signaling pathway	Oxytocin signaling pathway
Neuroactive ligand-receptor interaction	
Platelet activation	
Calcium signaling pathway	
Purine metabolism	
Phosphatidylinositol signaling system	
Amoebiasis	
Regulation of actin cytoskeleton	
PI3K-Akt signaling pathway	
Oxytocin signaling pathway	

[a] Enriched KEGG pathways determined using DAVID functional annotation tool

[b] Mus test samples are *M. pahari* and *M. caroli* species

[c] KEGG pathways are shared between the reference *M. musculus*, *M. pahari*, *M. caroli*, and *R. norvegicus* species

The pathways shared between these three species are primarily signaling pathways and pathways involved in maintaining the structural integrity of a cell, such as focal adhesion and adherens junction. There are six pathways that are shared between the reference *M. musculus*, *M. pahari*, *M. caroli*, and *R. norvegicus* test samples (Table 3.11.1, Appendix E Online). The pathways shared between the four species include focal adhesion and signalling pathways ($p<0.001$).

4 Discussion

4.1 Array-based MDGA genetic distances between samples reflect known taxonomic relationships

MDGA-based genetic distances between wild species reflect relationships based on published times of divergence. The number of loci genotyped decreased as divergence time from the house mouse increased. Cross-species findings for Bovine, Ovine, and Equine SNP50 Beadchip array genotyping data with respect to the reference organism reflected the cross-species findings using the MDGA (Miller et al., 2012). There was a loss of resolution at a subspecies level of examination for Mus samples of the intra-genus test set in the MDGA study. Incorrect differentiation between *M. n. minutoides* and *M. n. orangiae* may be attributed to the controversy surrounding the classification of *M. n. orangiae*. While considered a separate species, there is a paucity of molecular data for *M. n. orangiae*, and in fact, *M. n. orangiae* may be a cryptotype, or phenotypic variant of *M. n. minutoides* (Britton-Davidian et al., 2012; Chevret et al., 2014). The SNP data for the two species of African pygmy mice may indicate that these species are not phylogenetically separate but are the same species, but the sample size was very small and requires further testing. This is an interesting future direction to test at a population level using array-based genotyping and sequencing technologies with *M. n. minutoides* and *M. n. orangiae*, as these pygmy species are an understudied avenue of research. Further testing of the MDGA is required with large populations of subspecies to determine if there are enough informative SNP loci conserved in wild Mus subspecies to identify differences between the population structures of specific subspecies.

4.2 Array-based SNP genotyping cross-species requires attention to the test set composition

There are three considerations to take into account regarding the samples of a test set for a cross-species array genotyping study. The first consideration is that the criteria of which samples to include in a test set for a cross-species study is different from the criteria for a study that utilizes the model organism on which the array is designed. The MDGA is designed to capture the SNP diversity in strains of mice commonly used in research, and having greater than 97% of all loci genotyped in test samples is a benchmark for genotyping results to be included in a research analyses (Yang et al., 2009). Test samples of inbred mice that do not meet the inclusion criteria of having at least 97% of loci genotyped are removed from the test set and are considered poor quality DNA samples. Following the same inclusion criteria and standards when using the MDGA cross-species is not possible as the risk for off-target mutation increases with divergence time from the model organism.

The second consideration is that DNA hybridization and preparation conditions can alter hybridization of DNA to array probe sequences. The technical replicate files had fewer no call genotype assignments than the original CEL data files, which may be attributed to differences in hybridization conditions of sample DNA to array probes. Optimal DNA preparation temperatures will be affected if the DNA GC content is significantly different between model mouse species and non-model species (Lesnik and Freier, 1995). Differences in composition of test DNA due to off-target mutations can indirectly result in variation of hybridization intensity that affects genotyping (Didion et al., 2012).

The third consideration is that the genotyping algorithm used is optimized to work most effectively when particular conditions are met. Previous research has demonstrated that the genotyping algorithm recommended by Affymetrix, BRLMM-P, is sensitive to the composition of the samples included in a genotyping set (Hong et al., 2008; Miclaus et al., 2010). Samples in a genotyping set that are more similar to one another genetically will produce fewer false genotyping results (Hong et al., 2008). Upon closer examination of the Mus samples of the inter-order test set, the number of loci genotyped for the majority of wild species was much higher than would be expected in comparison to the results of *M. m. castaneus* samples that are 0.35 MYD from the reference. The increased number of loci genotyped is thought to be caused by effects of including very genetically dissimilar samples the of the test set (Miclaus et al., 2010). The number of loci genotyped can become inflated if the samples in the genotyping test set are too genetically different. The greater genetic homogeneity of only Mus samples in the intra-genus test set produced genotyping results that matched what was expected of the species based on divergence times. An underestimate of the genetic diversity of Mus samples in the intra-genus test set was not observed. The linear decrease in loci genotyped in Mus samples as divergence time increased reflected previous cross-species findings (Miller et al., 2012).

Recommendations for the construction of a test set of samples for an experiment utilizing the MDGA cross-species would be dependent upon the hypothesis tested. A large number of samples would be required to establish whether SNPs are present in populations of non-model species since the minor allele of a polymorphism may be present in as little as 1% of the population (Akey, 2003; Wang et al., 1998). Technical replicates should also be included to assess the quality of DNA hybridization to array probes for a particular species. Optimization of hybridization conditions should be made to reduce differences in array hybridization intensities and the resulting differences in genotype assignments between technical replicates.

4.3 Array-based SNP genotyping cross-species requires attention to the composition of the training set

A training set of samples genotyped using the MDGA that are not a part of the study set is employed to teach the genotyping algorithm how to assign genotypes to experimental test samples (Huang et al., 2011; Yang et al., 2009; Zhang et al., 2013). A sample can differ in percentage of loci genotyped depending on the nature of the other samples included in a genotyping set. The use of a training set that has sufficient genetic diversity to encompass that of the experimental test sets can assist in producing accurate genotyping of samples (Huang et al., 2011; Zhang et al., 2013). A training set optimized for cross-species genotyping would be composed of members of the same species as the test set. The MDGA genotype assignments of the training set would be validated to ensure accurate training of the genotyping algorithm. Inclusion of male and female samples would ensure more accurate genotype assignments on the X chromosome (Zhao et al., 2018). Males are hemizygous for SNP genotypes on the X chromosome, and a challenge of the genotyping algorithm is that a hemizygous allele is assigned a diploid homozygous genotype (Zhao et al., 2018). Analyzing SNPs on the X chromosome separately from autosomal SNPs and separating male and female samples would aid in fewer false genotype assignments.

The reference set of 114 classical inbred strains of mice utilized does not encompass the high relative genetic diversity of the sample sets of this cross-species study. To increase the accuracy of genotyping using the BRLMM-P algorithm, creating a training set with greater genetic diversity could decrease the possibility of falsely genotyped loci. A future experiment examining wild Mus species at a population level could establish a number of wild Mus samples as a training set for the BRLMM-P algorithm, but the genotypes of the training set samples must be validated using a different method than the MDGA, such as sequencing. Using wild Mus genotyping data to train the BRLMM-P algorithm would allow for a fewer number of false genotype assignments by the genotyping algorithm.

4.4 Limits and challenges in genotyping cross-species using the MDGA

There is a general decrease in the number of loci genotyped using the MDGA and an increase in the number of heterozygous loci genotyped as divergence time increases from the reference house mouse. The decrease in the number of loci genotyped cross-species with increasing divergence time reaches a plateau in the number of loci that can be genotyped between 10-15 MYD from the house mouse. The increase in percent heterozygosity within samples observed as divergence time increases also reaches a plateau in the amount of heterozygosity observed for non-Mus samples. Outside of the genus Mus, the plateau in SNP loci genotyped is attributed to off-target mutations that hinder DNA hybridization to array probe sequences. The plateau in percent heterozygosity represents an increase in the number of off-target hybridization of sample DNA to array probes. When DNA hybridizes to a probe on the MDGA, the hybridization does not have to be a perfect 25 nt match, where incomplete hybridization of the sample

DNA to the probe is enough to result in a genotype assignment (Binder and Preibisch, 2005). The nonspecific binding of DNA to MDGA probes and loss of allele specificity results in an inflation in the number of false heterozygous genotype assignments. Determination of the divergence time at which underestimates of genetic diversity begin is limited by the samples available for use in this study. A greater number of species genotyped using the MDGA that have a divergence time between 10-15 MYD from the house mouse would be beneficial in identifying when underestimations of genetic diversity begin. Researchers Miller et al. (2012), found previously that applying the Bovine, Ovine, and Equine SNP50 Beadchip arrays cross-species resulted in a linear decrease in genotyped loci as the millions of years of divergence from the model species increased (Miller et al., 2012). Along with a decrease in genotyped loci, there is an increase in heterozygous genotypes. Another aspect that reveals the challenges of applying the MDGA cross-species can be seen in changes of SNP-based genetic distances for the same samples depending on the composition of other samples in the test set. The interpretation of the relatedness through SNP-based genetic distances can be affected by the diversity of samples across the test set.

4.5 Difficulties in differentiating naked mole rat samples

The 73 million-year divergence time of the naked mole rat from the reference house mouse proved to be a challenge in genotyping samples. Only an approximate 44% of SNP loci were genotyped in naked mole rat samples, and a lack of genomic sequencing and annotation information makes *in silico* forms of genotype validation difficult. The naked mole rat genome that was available for use in the *in silico* sequence match analysis is a collection of unplaced, unannotated genome scaffolding (ftp.ncbi.nlm.nih.gov/genomes/Heterocephalus_glaber/). With only an approximate 1000 matches to the naked mole rat available genome sequence, it is difficult to determine cases of conserved variation between *M. musculus* and *H. glaber* without a more informative naked mole rat reference sequence. As more naked mole rat genomic sequence information and annotation becomes available, it will be easier to determine conserved variation between these two rodents (Keane et al., 2014).

The genotyped four naked mole rat samples in the case study were primarily differentiated based on the sex of the samples and not by the colony population structure. Naked mole rats are eusocial organisms with extremely genetically similar populations due to the high inbreeding coefficient of the species brought about through consanguineous mating (Reeve et al., 1990). Not much is known regarding the population structure of the two colonies (Desperado and Q) from which the donated naked mole rat samples are from. It is possible that by being donated by the same source, the two colonies have been interbred, which would interfere with the ability to differentiate the naked mole rat samples based on population structure alone. Given that there are over 18,000 probes on the MDGA that query the mouse X chromosome, the greatest difference between the samples would be differences in the sex chromosomes. The small sample size of this case study is a major limitation and a much larger sample size is needed to determine if naked mole rat samples can be differentiated from one another based on MDGA SNP loci. At a divergence time of 73 million years from the house mouse, the naked mole rat is too genetically distant and has populations with too little genetic diversity for cross-species application of the MDGA to be feasible for this species.

4.6 Deer mice are interesting candidate species for further analysis using the MDGA

The genotyping results of Peromyscus species were used to create SNP trees of genetic relatedness that reflect the known patterns of divergence for the seven Peromyscus samples studied (Bedford and Hoekstra, 2015; Bradley et al., 2007; Natarajan et al., 2015). The consensus of relative relatedness between Peromyscus samples determined using SNP genotypes and other molecular resources indicate that the MDGA may be a useful resource for learning more about conservation of variation between Mus and Peromyscus. The recapitulation of known divergence times for highly diverged species like the deer mouse (32.7 MYD) is possible if the test set of samples are from the same genus. Identifying polymorphic loci that are conserved between the model house mouse and non-model species is key to assessing population structure in the non-model species (Hoffman et al., 2013). In the two subspecies of *P. maniculatus*, there are over 140,000 loci that were assigned a heterozygous genotype for both samples. The SNP loci with a heterozygous genotype represent potential polymorphic loci in *P. maniculatus*.

Online genome sequence is available for one species of deer mouse, *P. maniculatus*. The 3,195 MDGA unique probe matches to the *P. maniculatus* genome determined using E-MEM that cross-validate loci genotyped using the MDGA represent a panel of candidate genome variation that may be conserved evolutionarily from the MRCA between Peromyscus and *M. musculus*. There are 481SNP loci with a heterozygous genotype in *P.*

maniculatus samples that were cross-validated to be present in the available genome assembly. The 481 loci may be informative polymorphic loci within *P. maniculatus*, but further validation of the SNPs in populations of deer mice are required in the future. Peromyscus species live in a variety of environments all across North America and as they are exposed to different environmental pressures, it would be interesting to learn if the panel of candidate conserved MDGA sequences in Peromyscus can reveal population specific genetic variation. The genic associations of population specific genetic variation discovered in Peromyscus may reveal information about genes undergoing directional selection as a response to a changing environment (Harris et al., 2013). The major KEGG pathways found to be significant for the mouse gene Ensembl IDs associated with the 3,195 cross-validated SNP loci in *P. maniculatus* are primarily neurological signaling pathways that would be expected to be conserved between the house mouse and deer mouse. For example, the top pathway associated with SNP loci genotyped in *P. maniculatus* is the glutamatergic synapse pathway. Glutamate is an important neurotransmitter in mammalian species and identifying SNP loci that are associated with this pathway is not unexpected (Parmentier et al., 2000).

4.7 Mutation signatures of wild Mus species

Patterns of transitions and transversions within the genome have been used to identify markers of evolutionary change in humans (Harris and Pritchard, 2017). Understanding signatures of mutational change can aid in identifying genomic mechanisms that cause adaptive evolutionary traits and episodes of rapid evolution in Mus (Harris et al., 2013;

Linnen et al., 2013). The trinucleotide mutational signature visualization demonstrates a sampling bias for C>T and T>C transitions in MDGA genotyped loci. The C>T and T>C bias is reflected in all wild Mus species analyzed. It is known that there is a mutational bias for transitions in rodents (Collins and Jukes, 1994), but the bias for C>T and T>C transitions found in Mus samples may be a reflection of the bias in MDGA design. There is a need for a quantitative method to normalize the results for wild Mus species against the array bias and then analyze for significant differences in mutational signatures.

4.8 Spatial visualization of variable and invariant loci with respect to the *Mus musculus* genome

Rainfall plots of SNP loci genotyped in Mus samples demonstrated known expectations of clustering of SNP variation. Loci variable in genotype across the test set were primarily located on autosomes and invariant loci were located in high frequency on the X chromosome. Fewer loci variable in genotype on the X chromosome reflects the high genetic conservation of the X chromosome between mammals (Raudsepp et al., 2004). The challenge of genotyping the X chromosome in test samples must be considered. The intra-genus test set of Mus samples is composed of primarily male mice with only two female samples, affecting analysis of the X chromosome. Males are hemizygous for the X chromosome, and thus it is not possible for male samples to be heterozygous for SNP loci on the X chromosome. Zero SNP loci were genotyped as heterozygous in all samples as expected, but issues arise in genotyping homozygous SNPs on the X chromosome. The hybridization intensity of X chromosome DNA binding to MDGA probes is interpreted as a diploid genotype of AA or BB for male samples, leading to false genotype assignments (Zhao et al., 2018). Rainfall plots of invariant AA and BB SNP loci on the X chromosome only reflect true AA and BB genotypes for the two *M. m. castaneus* samples. The X chromosome should be analyzed separately for male and female samples in future studies of cross-species hybridization. As more genomic information becomes available for wild species, it will be possible to quantitatively analyze clustering of SNP loci for populations of non-model organisms.

4.9 Comparisons of Mus cross-species array utility to other mammalian cross-species SNP-genotyping studies

Previous studies that have examined the utility of the cross-species application of commercially available genotyping array technology have identified trends of decreasing ability to genotype loci as divergence time from the model organism increases (Hoffman et al., 2013; Miller et al., 2012; Ogden et al., 2012). The MDGA study is unique as it tests the array technology on a wide range of species spanning multiple millions of years divergence from the reference house mouse. Previous studies such as by Ogden *et al.* (2012) focused on testing the commercial array technology on a few wild species rather than experimenting to determine the limits of cross-species utility of the technology. Previous research has determined potentially conserved sequences between model organisms and the wild species of interest through application commercial arrays to test samples (More et al., 2019; Ogden et al., 2012). The study of the MDGA cross-validates genotyped loci in rodent samples with an *in silico* analysis of available genomic sequences for wild species. The *in silico* search for the presence of a unique match of the 25 nt MDGA probe sequences within publicly available genomes of *M. caroli* and *M.*

pahari cross-validated 161,149 and 67,820 potentially conserved SNP variation shared respectively between these wild Mus samples and the reference *M. musculus.* The SNP variation of the MDGA study that was genotyped in rodent samples and cross-validated through *in silico* analyses are candidate SNPs that can be tested for conservation in populations of wild rodents. To be truly considered a SNP cross-species, the variation must be validated in wild populations with the alternate, or minor allele present in at least 1% of the population.

The study by Hoffman *et al.* (2013) that examined the cross-species utility of a canine genotyping array with Antarctic fur seals discovered 173 polymorphic SNPs that could be used to assay fur seal population structure. Heterozygous loci represent potential polymorphic loci in the MDGA study. After cross-validation of heterozygous SNPs genotyped in wild rodents, 481 potential polymorphic loci were found in *P. maniculatus* samples at a divergence time of 32.7 MYD from the house mouse. Given that there are fewer million years of divergence between the model house mouse and deer mice than the 44 MYD between the dog and seal, it was expected that a greater number of potential polymorphic SNP loci were discovered. For the two rat samples, over 1000 polymorphic loci appear to be conserved between *R. norvegicus* and the house mouse. The most closely-related samples with a genome assembly available *M. caroli* and *M. pahari* both had over 9000 potential polymorphic loci cross-validated with the *in silico* analysis. The presence of the potential polymorphic loci identified in the MDGA study should be investigated in wild populations in order to validate a set of SNPs that will be informative for non-model organisms.

The study by Hoffman *et al.* (2013) also identified pathways involved in energy metabolism as being conserved over the 44 MYD between the dog and seal. The study of the MDGA identified several signaling pathways and pathways associated in cellular integrity/functioning that are conserved between the house mouse and wild mouse species. The identification of a greater number of significant pathways in the MDGA study can be attributed to the shorter maximum divergence time of 8.29 MYD between the wild Mus species and the reference house mouse compared to the 44 MYD between the dog and the seal. The MDGA also surveys over 300,000 more loci than the canine array, contributing to an increased amount of genomic information to study. SNP variation associated with pathways that are significant in wild Mus samples and the reference house mouse may represent conserved SNPs between the reference and test species. Confirmation of the enrichment of the identified pathways in populations of the wild species must be made before variation shared between the samples can truly be considered conserved. The pathways that are significant in the reference and test samples are large pathways that include genes that are involved in multiple gene networks. Variation in genes key to multiple functional pathways may be involved in important biological functions that are less likely to rapidly evolve or tolerate mutations (Gussow et al., 2016; Wolf et al., 2009). The main caveats of the functional study are that all functional gene annotations are with respect to the reference house mouse. Due to the genome shuffling and rearrangements that occur during evolution, it is possible that the candidate conserved variants are associated with different regions of the genome and the functional associations are not the same between the house mouse and wild species (Zhao et al., 2004).

4.10 Future cross-species applications of the MDGA

The proposed SNPSTeP method of visualizing SNP genotype changes across the genome can be used to identify regions characterized by specific SNP changes. The *M. m. castaneus* sample and technical replicate of the MDGA study comprise the only female samples of the dataset, and from the SNPSTeP visualization of the X chromosome a central region of high heterozygosity was found. The general low genetic diversity seen on the X chromosome can be attributed to highly conserved coding regions and the region of variability may represent variation associated with adaptive genes in the Mus sample (Chen et al., 2018; Mácha et al., 2012). SNPSTeP visualizations could inform researchers about key genomic regions of Mus species involved in adaptive variation and polymorphisms involved in rapid evolution (Harris et al., 2013). An example of adaptive variation in rodents is the introduction of a polymorphism into wild populations of mice that conferred resistance to harmful rodenticides (Song et al., 2011).

The MDGA may be used in conjunction with current technologies like restriction-site associated (RAD) sequencing, which is based on fragmenting DNA with a restriction enzyme digest, and filtering fragments by size to reduce the DNA sequencing library. Fragments of a specific length are than sequenced to identify SNP variation in populations of model and non-model organisms (Peterson et al., 2012). Using array-based genotyping technologies cross-species will be useful in identifying known SNP variation conserved between model and non-model species, while technologies like RADseq can be used to identify novel SNPs in non-model species. As next-generation sequencing costs continue to decrease, the possibility of the creation of fully sequenced and annotated genomes for wild species becomes a greater possibility. Cross-species utility of the MDGA is a first step in identifying SNP variation in the genomes of non-model organisms. The new generation of genotyping arrays such as the Axiom array (ThermoFisher Scientific) was designed for SNP genotype identification in the house mouse and was based on the design of the MDGA. The mouse Axiom array shares 488,945 of the same SNP loci that are targeted by the MDGA. The new Axiom array also identifies genotypes at over 100,000 additional SNP loci compared to the MDGA, opening new cross-species research opportunities for the future.

5 Conclusions

Due to the decreasing amount of genetic relatedness as divergence time increases between species intra-genus, to inter-genera, to inter-family, to inter-order, the cross-species utility of the MDGA is best suited for species of the genus Mus. Within Mus, the number of loci genotyped decreases with increasing divergence time from the reference house mouse, but SNP-based genetic distances obtained from cross-species application of the MDGA reflect the known taxonomic relationships between Mus samples. The validation of the presence SNP loci with heterozygous genotypes in a population is necessary to identify informative polymorphic SNPs that can be used cross-species. Despite the 32.7 MYD between the house mouse and deer mouse, there is evidence for cross-species utility of the MDGA beyond the genus Mus, but special consideration must be made regarding the composition of the training and test sets of samples. For very highly diverged species from the house mouse like the naked mole rat (73 MYD) that also have populations with little genetic variation between them, the utility of the array is very limited.

In silico analyses provided a cross-validation for the MDGA genotyped loci within the genomes of wild rodent species. A panel of SNPs was identified for *M. caroli*, *M. pahari*, *R. norvegicus*, and *P. maniculatus* that represent potentially conserved SNP variation between the reference house mouse and wild rodent samples. The cross-validated SNP loci identified as being potentially polymorphic are key loci to be targeted in tests of SNP conservation in wild populations. Learning the functional annotations of conserved variation will be a key step in discovering the interplay between genotype and phenotype in non-model species. New genotyping array technologies that are more cost and time efficient than the MDGA are valuable tools that can be used to identify SNPs cross-species in conjunction with other current technologies RADseq that do not rely on a reference genome.

6 Bibliography

Abegglen, L. M., Caulin, A. F., Chan, A., Lee, K., Robinson, R., Campbell, M. S., Kiso, W. K., Schmitt, D. L., Waddell, P. J., Bhaskara, S., et al. (2015). Potential Mechanisms for Cancer Resistance in Elephants and Comparative Cellular Response to DNA Damage in Humans. *JAMA* **314**, 1850.

Aditi, K., Shakarad, M. N. and Agrawal, N. (2016). Altered lipid metabolism in Drosophila model of Huntington's disease. *Sci. Rep.* **6**, 31411.

Akey, J. M. (2003). The Effect of Single Nucleotide Polymorphism Identification Strategies on Estimates of Linkage Disequilibrium. *Mol. Biol. Evol.* **20**, 232–242.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V, Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al. (2013). Signatures of mutational processes in human cancer. *Nature* **500**, 415–21.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Bedford, N. L. and Hoekstra, H. E. (2015). Peromyscus mice as a model for studying natural variation. *Elife* **4**, 1–13.

Bennett, E. A., Champlot, S., Peters, J., Arbuckle, B. S., Guimaraes, S., Pruvost, M., Bar-David, S., Davis, S. J. M., Gautier, M., Kaczensky, P., et al. (2017). Taming the late Quaternary phylogeography of the Eurasiatic wild ass through ancient and modern DNA. *PLoS One* **12**, e0174216.

Bickham, J. W., Sandhu, S., Hebert, P. D. ., Chikhi, L. and Athwal, R. (2000). Effects of chemical contaminants on genetic diversity in natural populations: implications for biomonitoring and ecotoxicology. *Mutat. Res. Mutat. Res.* **463**, 33–51.

Binder, H. and Preibisch, S. (2005). Specific and nonspecific hybridization of oligonucleotide probes on microarrays. *Biophys. J.* **89**, 337–52.

Bradley, R. D., Durish, N. D., Rogers, D. S., Miller, J. R., Engstrom, M. D. and Kilpatrick, C. W. (2007). TOWARD A MOLECULAR PHYLOGENY FOR PEROMYSCUS: EVIDENCE FROM MITOCHONDRIAL CYTOCHROME-b SEQUENCES. *J Mammal* **88**, 1146–1159.

Britton-Davidian, J., Robinson, T. J. and Veyrunes, F. (2012). Systematics and evolution of the African pygmy mice, subgenus Nannomys: A review. *Acta Oecologica* **42**, 41–49.

Bryja, J., Mikula, O., Šumbera, R., Meheretu, Y., Aghová, T., Lavrenchenko, L. A., Mazoch, V., Oguge, N., Mbau, J. S., Welegerima, K., et al. (2014). Pan-African phylogeny of Mus (subgenus Nannomys) reveals one of the most successful mammal radiations in Africa. *BMC Evol. Biol.* **14**, 256.

Bumgarner, R. (2013). Overview of DNA microarrays: types, applications, and their future. *Curr. Protoc. Mol. Biol.* **Chapter 22**, Unit 22.1.

Chen, Z.-H., Zhang, M., Lv, F.-H., Ren, X., Li, W.-R., Liu, M.-J., Nam, K., Bruford, M. W. and Li, M.-H. (2018). Contrasting Patterns of Genomic Diversity Reveal Accelerated Genetic Drift but Reduced Directional Selection on X-Chromosome in Wild and Domestic Sheep Species. *Genome Biol. Evol.* **10**, 1282–1297.

Chevret, P., Robinson, T. J., Perez, J., Veyrunes, F. and Britton-Davidian, J. (2014). A Phylogeographic Survey of the Pygmy Mouse Mus minutoides in South Africa: Taxonomic and Karyotypic Inference from Cytochrome b Sequences of Museum Specimens. *PLoS One* **9**, 98499.

Coll, F., McNerney, R., Guerra-Assunção, J. A., Glynn, J. R., Perdigão, J., Viveiros, M., Portugal, I., Pain, A., Martin, N. and Clark, T. G. (2014). A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nat. Commun.* **5**, 4812.

Collins, D. W. and Jukes, T. H. (1994). Rates of Transition and Transversion in Coding Sequences since the Human-Rodent Divergence. *Genomics* **20**, 386–396.

Coop, G., Pickrell, J. K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R. M., Cavalli-Sforza, L. L., Feldman, M. W. and Pritchard, J. K. (2009). The Role of Geography in Human Adaptation. *PLoS Genet.* **5**, e1000500.

Csiszar, A., Labinskyy, N., Orosz, Z., Xiangmin, Z., Buffenstein, R. and Ungvari, Z. (2007). Vascular aging in the longest-living rodent, the naked mole rat. *Am. J. Physiol. Heart Circ. Physiol.* **293**, H919-27.

da Silva, J., de Freitas, T. R. O., Heuser, V., Marinho, J. R., Bittencourt, F., Cerski, C. T. S., Kliemann, L. M. and Erdtmann, B. (2000). Effects of chronic exposure to coal in wild rodents (Ctenomys torquatus) evaluated by multiple methods and tissues. *Mutat. Res. Toxicol. Environ. Mutagen.* **470**, 39–51.

DeMay, S. M., Becker, P. A., Rachlow, J. L. and Waits, L. P. (2017). Genetic monitoring of an endangered species recovery: demographic and genetic trends for reintroduced pygmy rabbits (Brachylagus idahoensis). *J. Mammal.* **98**, 350–364.

Didion, J. P., Yang, H., Sheppard, K., Fu, C.-P., McMillan, L., de Villena, F. and Churchill, G. A. (2012). Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. *BMC Genomics* **13**, 34.

Domanska, D., Vodák, D., Lund-Andersen, C., Salvatore, S., Hovig, E. and Sandve, G. K. (2017). The rainfall plot: its motivation, characteristics and pitfalls. *BMC Bioinformatics* **18**, 264.

Draghici, S., Khatri, P., Eklund, A. C. and Szallasi, Z. (2006). Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet.* **22**, 101–9.

Eppig, J. T., Richardson, J. E., Kadin, J. A., Smith, C. L., Blake, J. A., Bult, C. J. and MGD Team (2015). Mouse Genome Database: From sequence to phenotypes and disease models. *genesis* **53**, 458–473.

Eusebi, P. G., Cortés, O., Dunner, S. and Cañón, J. (2017). Genetic diversity of the Mexican Lidia bovine breed and its divergence from the Spanish population. *J. Anim. Breed. Genet.* **134**, 332–339.

Gao, Hong, Pirani Ali, Webster, Teresa, S. M.-M. (2013). Systems and Methods for SNP Characterization and Identifying off Target Variants.

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14**, 685–695.

Geraldes, A., Basset, P., Smith, K. L. and Nachman, M. W. (2012). Higher differentiation among subspecies of the house mouse (Mus musculus) in genomic regions with low recombination. *Mol. Ecol.* **20**, 4722–4736.

Grant, P. R. and Grant, B. R. (2002). Unpredictable evolution in a 30-year study of Darwin's finches. *Science* **296**, 707–11.

Gunderson, K. L., Steemers, F. J., Lee, G., Mendoza, L. G. and Chee, M. S. (2005). A genome-wide scalable SNP genotyping assay using microarray technology. *Nat. Genet.* **37**, 549–554.

Gussow, A. B., Petrovski, S., Wang, Q., Allen, A. S. and Goldstein, D. B. (2016). The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biol.* **17**, 9.

Guterres, D. C., Galvão-Elias, S., de Souza, B. C. P., Pinho, D. B., dos Santos, M. do D. M., Miller, R. N. G. and Dianese, J. C. (2018). Taxonomy, phylogeny, and divergence time estimation for *Apiosphaeria guaranitica* , a Neotropical parasite on bignoniaceous hosts. *Mycologia* **110**, 526–545.

Hannigan, G. D., Zheng, Q., Meisel, J. S., Minot, S. S., Bushman, F. D. and Grice, E. A. (2017). Evolutionary and functional implications of hypervariable loci within the skin virome. *PeerJ* **5**, e2959.

Harr, B. (2006). Genomic islands of differentiation between house mouse subspecies. *Genome Res.* **16**, 730–7.

Harris, K. and Pritchard, J. K. (2017). Rapid evolution of the human mutation spectrum. *Elife* **6**,.

Harris, S. E., Munshi-South, J., Obergfell, C. and Neill, O. (2013). Signatures of Rapid Evolution in Urban and Rural Transcriptomes of White-Footed Mice (Peromyscus leucopus) in the New York Metropolitan Area. *PLoS One* **8**, 74938.

Hedges, S. B., Marin, J., Suleski, M., Paymer, M. and Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. *Mol. Biol. Evol.* **32**, 835–845.

Hoffman, J. I., Thorne, M. A. S., McEwing, R., Forcada, J. and Ogden, R. (2013). Cross-Amplification and Validation of SNPs Conserved over 44 Million Years between Seals and Dogs. *PLoS One* **8**, 1–10.

Hong, H., Su, Z., Ge, W., Shi, L., Perkins, R., Fang, H., Xu, J., Chen, J. J., Han, T., Kaput, J., et al. (2008). Assessing batch effects of genotype calling algorithm BRLMM for the Affymetrix GeneChip Human Mapping 500 K array set using 270 HapMap samples.

Hu, H. and Larson, R. G. (2006). Marangoni Effect Reverses Coffee-Ring Depositions.

Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13.

Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57.

Huang, L., Jakobsson, M., Pemberton, T. J., Ibrahim, M., Nyambo, T., Omar, S., Pritchard, J. K., Tishkoff, S. A. and Rosenberg, N. A. (2011). Haplotype variation and genotype imputation in African populations. *Genet. Epidemiol.* **35**, 766–80.

Hulme-Beaman, A., Dobney, K., Cucchi, T. and Searle, J. B. (2016). An Ecological and Evolutionary Framework for Commensalism in Anthropogenic Environments. *Trends Ecol. Evol.* **31**, 633–645.

Jarvis, J. U. (1981). Eusociality in a mammal: cooperative breeding in naked mole-rat colonies. *Science* **212**, 571–3.

Johnson, W. E., Culver, M., Iriarte, J., Eizirik, E., Seymour, K. and O'Brien, S. (1998). Tracking the evolution of the elusive Andean mountain cat (Oreailurus jacobita from mitochondrial DNA. *J. Hered.* **89**, 227–232.

Jose M. Moran-Mirabal, †, Christine P. Tan, ‡, Reid N. Orth, ‡, Eric O. Williams, §, Harold G. Craighead, † and and David M. Lin*, § (2006). Controlling Microarray Spot Morphology with Polymer Liftoff Arrays.

Keane, T. M., Goodstadt, L., Danecek, P., White, M. A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., et al. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294.

Keane, M., Craig, T., Alföldi, J., Berlin, A. M., Johnson, J., Seluanov, A., Gorbunova, V., Di Palma, F., Lindblad-Toh, K., Church, G. M., et al. (2014). The Naked Mole Rat Genome Resource: Facilitating analyses of cancer and longevity-related adaptations. *Bioinformatics*.

Keating, B. J., Tischfield, S., Murray, S. S., Bhangale, T., Price, T. S., Glessner, J. T., Galver, L., Barrett, J. C., Grant, S. F. A., Farlow, D. N., et al. (2008). Concept, Design and Implementation of a Cardiovascular Gene-Centric 50 K SNP Array for Large-Scale Genomic Association Studies. *PLoS One* **3**, e3583.

Kharzinova, V. R., Sermyagin, A. A., Gladyr, E. A., Okhlopkov, I. M., Brem, G. and Zinovieva, N. A. (2015). A study of applicability of SNP chips developed for bovine and ovine species to whole-genome analysis of reindeer rangifer tarandus. *J. Hered.* **106**, 758–761.

Khiste, N. and Ilie, L. (2015). E-MEM: efficient computation of maximal exact matches for very large genomes. *Bioinformatics* **31**, 509–514.

Kouassi, S. K., Nicolas, V., Aniskine, V., Lalis, A., Cruaud, C., Couloux, A., Colyn, M., Dosso, M., Koivogui, L., Verheyen, E., et al. (2008). Taxonomy and biogeography of the African Pygmy mice, Subgenus Nannomys (Rodentia, Murinae, Mus) in Ivory Coast and Guinea (West Africa). *Mammalia* **72**, 237–252.

Krubitzer, L., Campi, K. L. and Cooke, D. F. (2011). All Rodents Are Not the Same: A Modern Synthesis of Cortical Organization. *Brain. Behav. Evol.* **78**, 51.

Kucab, J. E., Zou, X., Morganella, S., Joel, M., Nanda, A. S., Nagy, E., Gomez, C., Degasperi, A., Harris, R., Jackson, S. P., et al. (2019). A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821-836.e16.

Kuperwasser, C., Dessain, S., Bierbaum, B. E., Garnet, D., Sperandio, K., Gauvin, G. P., Naber, S. P., Weinberg, R. A. and Rosenblatt, M. (2005). A Mouse Model of Human Breast Cancer Metastasis to Human Bone. *Cancer Res.* **65**, 6130–6138.

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res.* **37**, 4181–93.

Lah, L., Trense, D., Benke, H., Berggren, P., Gunnlaugsson, Þ., Lockyer, C., Öztürk, A., Öztürk, B., Pawliczka, I., Roos, A., et al. (2016). Spatially Explicit Analysis of Genome-Wide SNPs Detects Subtle Population Structure in a Mobile Marine Mammal, the Harbor Porpoise. *PLoS One* **11**, e0162792.

Lamy, P., Grove, J. and Wiuf, C. (2011). A review of software for microarray genotyping. *Hum. Genomics* **5**, 304–9.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Lesnik, E. A. and Freier, S. M. (1995). Relative Thermodynamic Stability of DNA, RNA, and DNA:RNA Hybrid Duplexes: Relationship with Base Composition and Structure. *Biochemistry* **34**, 10807–10815.

Li, W.-H., Gouy, M., Sharp, P. M., O'huigin, C. and Yang, Y.-W. (1990). *Molecular phylogeny of Rodentia, Lagomorpha, Primates, Artiodactyla, and Carnivora and molecular clocks (mammalian phylogeny/DNA sequence trees/branching dates)*.

Libiger, O., Nievergelt, C. M. and Schork, N. J. (2009). Comparison of genetic distance measures using human SNP genotype data. *Hum. Biol.* **81**, 389–407.

Linnen, C. R., Poh, Y.-P., Peterson, B. K., Barrett, R. D. H., Larson, J. G., Jensen, J. D. and Hoekstra, H. E. (2013). Adaptive Evolution of Multiple Traits Through Multiple Mutations at a Single Gene. *Science (80-.).* **339**, 1312–1316.

Locke, M. E. O., Milojevic, M., Eitutis, S. T., Patel, N., Wishart, A. E., Daley, M. and Hill, K. A. (2015). Genomic copy number variation in Mus musculus. Additional file 1. *BMC Genomics* **16**, 497.

Luis Villanueva-Cañas, J., Ruiz-Orera, J., Agea, M. I., Gallo, M., Andreu, D. and Albà, M. M. (2017). New Genes and Functional Innovation in Mammals. *Genome Biol. Evol.* **9**, 1886–1900.

M. Raafat El-Gewely *Biotechnology Annual Review - M. Raafat El-Gewely - Google Books*.

Mácha, J., Teichmanová, R., Sater, A. K., Wells, D. E., Tlapáková, T., Zimmerman, L. B. and Krylov, V. (2012). Deep ancestry of mammalian X chromosome revealed by comparison with the basal tetrapod Xenopus tropicalis. *BMC Genomics* **13**, 315.

Marth, G. T., Korf, I., Yandell, M. D., Yeh, R. T., Gu, Z., Zakeri, H., Stitziel, N. O., Hillier, L., Kwok, P.-Y. and Gish, W. R. (1999). A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* **23**, 452–456.

McCue, M. E., Bannasch, D. L., Petersen, J. L., Gurr, J., Bailey, E., Binns, M. M., Distl, O., Guérin, G., Hasegawa, T., Hill, E. W., et al. (2012). A high density SNP array for the domestic horse and extant Perissodactyla: Utility for association mapping, genetic diversity, and phylogeny studies. *PLoS Genet.* **8**,.

Miclaus, K., Wolfinger, R., Vega, S., Chierici, M., Furlanello, C., Lambert, C., Hong, H., Zhang, L., Yin, S. and Goodsaid, F. (2010). Batch effects in the BRLMM genotype calling algorithm influence GWAS results for the Affymetrix 500K array. *Pharmacogenomics J.* **10**, 336–46.

Millburn, G. H., Crosby, M. A., Gramates, L. S., Tweedie, S. and FlyBase Consortium (2016). FlyBase portals to human disease research using *Drosophila* models. *Dis. Model. Mech.* **9**, 245–252.

Miller, J. M., Kijas, J. W., Heaton, M. P., McEwan, J. C. and Coltman, D. W. (2012). Consistent divergence times and allele sharing measured from cross-species application of SNP chips developed for three domestic species. *Mol. Ecol. Resour.* **12**, 1145–1150.

Miller, J. M., Moore, S. S., Stothard, P., Liao, X. and Coltman, D. W. (2015). Harnessing cross-species alignment to discover SNPs and generate a draft genome sequence of a bighorn sheep (Ovis canadensis). *BMC Genomics* **16**, 397.

Miller, J. M., Festa-Bianchet, M. and Coltman, D. W. (2018). Genomic analysis of morphometric traits in bighorn sheep using the Ovine Infinium ® HD SNP BeadChip. *PeerJ* **6**, e4364.

Montana, L., Caniglia, R., Galaverni, M., Fabbri, E., Ahmed, A., Bolfíková, B. Č., Czarnomska, S. D., Galov, A., Godinho, R., Hindrikson, M., et al. (2017). Combining phylogenetic and demographic inferences to assess the origin of the genetic diversity in an isolated wolf population. *PLoS One* **12**, e0176560.

Moravcikova, N., Kirchner, R., Sidlova, V., Kasarda, R. and Trakovicka, A. (2015). Estimation of genomic variation in cervids using cross-species application of SNP arrays. *Poljoprivreda/Agriculture* **21**, 33–36.

More, M., Gutiérrez, G., Rothschild, M., Bertolini, F. and Ponce de León, F. A. (2019). Evaluation of SNP Genotyping in Alpacas Using the Bovine HD Genotyping Beadchip. *Front. Genet.* **10**, 361.

Morgan, A. P., Fu, C.-P., Kao, C.-Y., Welsh, C. E., Didion, J. P., Yadgary, L.,

Hyacinth, L., Ferris, M. T., Bell, T. A., Miller, D. R., et al. (2016). The Mouse Universal Genotyping Array: From Substrains to Subspecies. *G3&amp;#58; Genes|Genomes|Genetics* **6**, 263–279.

Morin, P. A., Luikart, G., Wayne, R. K. and the SNP workshop group (2004). SNPs in ecology, evolution and conservation. *Trends Ecol. Evol.* **19**, 208–216.

Mungall, C. J., Washington, N. L., Nguyen-Xuan, J., Condit, C., Smedley, D., Köhler, S., Groza, T., Shefchek, K., Hochheiser, H., Robinson, P. N., et al. (2015). Use of model organism and disease databases to support matchmaking for human disease gene discovery. *Hum. Mutat.* **36**, 979–84.

Natarajan, C., Hoffmann, F. G., Lanier, H. C., Wolf, C. J., Cheviron, Z. A., Spangler, M. L., Weber, R. E., Fago, A. and Storz, J. F. (2015). Intraspecific polymorphism, interspecific divergence, and the origins of function-altering mutations in deer mouse hemoglobin. *Mol. Biol. Evol.* **32**, 978–997.

Nicolas, V., Schaeffer, B., Missoup, A. D., Kennis, J., Colyn, M., Denys, C., Tatard, C., Cruaud, C. and Laredo, C. (2012). Assessment of three mitochondrial genes (16S, Cytb, CO1) for identifying species in the Praomyini tribe (Rodentia: Muridae). *PLoS One* **7**, e36586.

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L. A., et al. (2012). Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* **149**, 979–993.

Nik-Zainal, S., Kucab, J. E., Morganella, S., Glodzik, D., Alexandrov, L. B., Arlt, V. M., Weninger, A., Hollstein, M., Stratton, M. R. and Phillips, D. H. (2015). The genome as a record of environmental exposure. *Mutagenesis* **30**, 763–770.

O'Brien, S. J., Menotti-Raymond, M., Murphy, W. J., Nash, W. G., Wienberg, J., Stanyon, R., Copeland, N. G., Jenkins, N. A., Womack, J. E. and Marshall Graves, J. A. (1999). The Promise of Comparative Genomics in Mammals. *Science (80-.).* **286**, 458–481.

Ogden, R., Baird, J., Senn, H. and McEwing, R. (2012). The use of cross-species genome-wide arrays to discover SNP markers for conservation genetics: A case study from Arabian and scimitar-horned oryx. *Conserv. Genet. Resour.* **4**, 471–473.

Ohno, S., Kaplan, W. D. and Kinosita, R. (1957). Heterochromatic regions and nucleolus organizers in chromosomes of the mouse, Mus musculus. *Exp. Cell Res.* **13**, 358–364.

Parmentier, M.-L., Galvez, T., Acher, F., Peyre, B., Pellicciari, R., Grau, Y.,

Bockaert, J. and Pin, J.-P. (2000). Conservation of the ligand recognition site of metabotropic glutamate receptors during evolution. *Neuropharmacology* **39**, 1119–1131.

Pertoldi, C., Wójcik, J. M., Tokarska, M., Kawałko, A., Kristensen, T. N., Loeschcke, V., Gregersen, V. R., Coltman, D., Wilson, G. A., Randi, E., et al. (2010). Genome variability in European and American bison detected using the BovineSNP50 BeadChip. *Conserv. Genet.* **11**, 627–634.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S. and Hoekstra, H. E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS One* **7**, e37135.

Pisciottano, F., Cinalli, A. R., Stopiello, J. M., Castagna, V. C., Elgoyhen, A. B., Rubinstein, M., Gómez-Casati, M. E. and Franchini, L. F. (2019). Inner Ear Genes Underwent Positive Selection and Adaptation in the Mammalian Lineage. *Mol. Biol. Evol.*

Pounds, S., Cheng, C., Mullighan, C., Raimondi, S. C., Shurtleff, S. and Downing, J. R. (2009). Reference alignment of SNP microarray signals for copy number analysis of tumors. *Bioinformatics* **25**, 315–21.

Rabbee, N. and Speed, T. P. (2006). A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics* **22**, 7–12.

Ramsdell, C. M., Lewandowski, A. A., Glenn, J., Vrana, P. B., O'Neill, R. J. and Dewey, M. J. (2008). Comparative genome mapping of the deer mouse (Peromyscus maniculatus) reveals greater similarity to rat (Rattus norvegicus) than to the lab mouse (Mus musculus). *BMC Evol. Biol.* **8**, 65.

Raudsepp, T., Lee, E.-J., Kata, S. R., Brinkmeyer, C., Mickelson, J. R., Skow, L. C., Womack, J. E. and Chowdhary, B. P. (2004). *Exceptional conservation of horse-human gene order on X chromosome revealed by high-resolution radiation hybrid mapping*.

Razgour, O., Forester, B., Taggart, J. B., Bekaert, M., Juste, J., Ibáñez, C., Puechmaille, S. J., Novella-Fernandez, R., Alberdi, A. and Manel, S. (2019). Considering adaptive genetic variation in climate change vulnerability assessment reduces species range loss projections. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 10418–10423.

Reeve, H. K., Westneatt, D. F., Noont, W. A., Sherman, P. W. and Aquadrot, C. F. (1990). *DNA &quot;fingerprinting&quot; reveals high levels of inbreeding in colonies of the eusocial naked mole-rat (cooperative breeding/eusoality/kin selection/hypervariable minisateilite DNA)*.

Rodríguez-Estival, J. and Smits, J. E. G. (2016). Small mammals as sentinels of oil sands related contaminants and health effects in northeastern Alberta, Canada. *Ecotoxicol. Environ. Saf.* **124**, 285–295.

Rudra, M., Chatterjee, B. and Bahadur, M. (2016). Phylogenetic relationship and time of divergence of Mus terricolor with reference to other Mus species. *J. Genet.* **95**, 399–409.

Sahm, A., Bens, M., Szafranski, K., Holtze, S., Groth, M., Gö Rlach, M., Calkhoven, C., Mü Ller, C., Schwab, M., Kraus, J., et al. (2018). Long-lived rodents reveal signatures of positive selection in genes associated with lifespan.

Saifullah and Tsukahara, T. (2018). Genotyping of single nucleotide polymorphisms using the SNP-RFLP method. *Biosci. Trends* **12**, 240–246.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–25.

Seluanov, A., Gladyshev, V. N., Vijg, J. and Gorbunova, V. (2018). Mechanisms of cancer resistance in long-lived mammals. *Nat. Rev. Cancer* **18**, 433–441.

Sharma, T., Cheong, N., Sen, P. and Sen, S. (1986). Constitutive Heterochromatin and Evolutionary Divergence of Mus dunni, M. booduga and M. musculus.pp. 35–44. Springer, Berlin, Heidelberg.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M. and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311.

Shimoyama, M., Laulederkind, S. J. F., De Pons, J., Nigam, R., Smith, J. R., Tutaj, M., Petri, V., Hayman, G. T., Wang, S.-J., Ghiasvand, O., et al. (2016). Exploring human disease using the Rat Genome Database. *Dis. Model. Mech.* **9**, 1089–1095.

Silva, J. da, Freitas, T. R. O. de, Marinho, J. R., Speit, G. and Erdtmann, B. (2000). An alkaline single-cell gel electrophoresis (comet) assay for environmental biomonitoring with native rodents. *Genet. Mol. Biol.* **23**, 241–245.

Sivashankari, S. and Shanmughavel, P. (2007). Comparative genomics - a perspective. *Bioinformation* **1**, 376–8.

Song, Y., Endepols, S., Klemann, N., Richter, D., Matuschka, F.-R., Shih, C.-H., Nachman, M. W. and Kohn, M. H. (2011). Adaptive Introgression of Anticoagulant Rodent Poison Resistance by Hybridization between Old World Mice. *Curr. Biol.* **21**, 1296–1301.

Štambuk, A., Šrut, M., Šatović, Z., Tkalec, M. and Klobučar, G. I. V. (2013). Gene flow vs. pollution pressure: Genetic diversity of Mytilus galloprovincialis in eastern Adriatic. *Aquat. Toxicol.* **136–137**, 22–31.

Styczyńska-Soczka, K., Zechini, L. and Zografos, L. (2017). Validating the Predicted Effect of Astemizole and Ketoconazole Using a *Drosophila* Model of Parkinson's Disease. *Assay Drug Dev. Technol.* **15**, 106–112.

Sun, J. X., Mullikin, J. C., Patterson, N. and Reich, D. E. (2009). Microsatellites are molecular clocks that support accurate inferences about history. *Mol. Biol. Evol.* **26**, 1017–27.

Tavaré, S., Marshall, C. R., Will, O., Soligo, C. and Martin, R. D. (2002). Using the fossil record to estimate the age of the last common ancestor of extant primates. *Nature* **416**, 726–729.

Tian, X., Azpurua, J., Hine, C., Vaidya, A., Myakishev-Rempel, M., Ablaeva, J., Mao, Z., Nevo, E., Gorbunova, V. and Seluanov, A. (2013). High-molecular-mass hyaluronan mediates the cancer resistance of the naked mole rat. *Nature* **499**, 346–349.

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., Burtt, N. P., Fuchsberger, C., Li, Y., Erdmann, J., et al. (2012). The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. *PLoS Genet.* **8**, e1002793.

vonHoldt, B. M., Pollinger, J. P., Lohmueller, K. E., Han, E., Parker, H. G., Quignon, P., Degenhardt, J. D., Boyko, A. R., Earl, D. A., Auton, A., et al. (2010). Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* **464**, 898–902.

Vonholdt, B. M., Pollinger, J. P., Lohmueller, K. E., Han, E., Parker, H. G., Quignon, P., Degenhardt, J. D., Boyko, A. R., Earl, D. A., Auton, A., et al. (2010). Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* **464**, 898–902.

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., et al. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–82.

Wang, W., Zheng, H., Yang, S., Yu, H., Li, J., Jiang, H., Su, J., Yang, L., Zhang, J., McDermott, J., et al. (2005). Origin and evolution of new exons in rodents. *Genome Res.* **15**, 1258–1264.

Wang, W., Gan, J., Fang, D., Tang, H., Wang, H., Yi, J. and Fu, M. (2018). Genome-wide SNP discovery and evaluation of genetic diversity among six Chinese indigenous cattle breeds in Sichuan. *PLoS One* **13**, e0201534.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562.

Wetterstrand Kris A DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).

Williams, L. M., Ma, X., Boyko, A. R., Bustamante, C. D. and Oleksiak, M. F. (2010). SNP identification, verification, and utility for population genetics in a non-model genus. *BMC Genet.* **11**, 32.

Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V. and Lipman, D. J. (2009). The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc. Natl. Acad. Sci.* **106**, 7273–7280.

Wray, N. R., Goddard, M. E. and Visscher, P. M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17**, 1520–8.

Yang, H., Ding, Y., Hutchins, L. N., Szatkiewicz, J., Bell, T. A., Paigen, B. J., Graber, J. H., de Villena, F. P.-M. and Churchill, G. A. (2009). A customized and versatile high-density genotyping array for the mouse. *Nat. Methods* **6**, 663–666.

Ye, S., Dhillon, S., Ke, X., Collins, A. R. and Day, I. N. (2001). An efficient procedure for genotyping single nucleotide polymorphisms. *Nucleic Acids Res.* **29**, E88-8.

Yosida, T. H. (1981). Chromosome alteration and the development of tumors. XXIII. Banding karyotype analyses of methylcholanthrene-induced tumors in the Indian spiny mouse, Mus platythrix, with special regard to the anomalies of chromosomes with nucleolar organizer regions. *Cancer Genet. Cytogenet.* **3**, 211–220.

Zeef, D. H., van Goethem, N. P., Vlamings, R., Schaper, F., Jahanshahi, A., Hescham, S., von Hörsten, S., Prickaerts, J. and Temel, Y. (2012). Memory deficits in the transgenic rat model of Huntington's disease. *Behav. Brain Res.* **227**, 194–198.

Zhang, P., Zhan, X., Rosenberg, N. A. and Zöllner, S. (2013). Genotype Imputation Reference Panel Selection Using Maximal Phylogenetic Diversity.

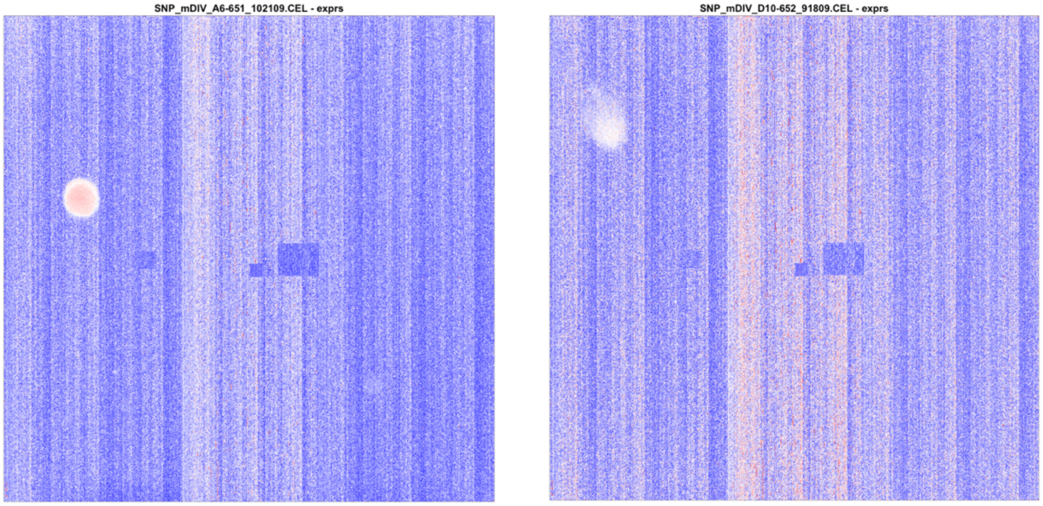
Zhao, S., Shetty, J., Hou, L., Delcher, A., Zhu, B., Osoegawa, K., de Jong, P., Nierman, W. C., Strausberg, R. L. and Fraser, C. M. (2004). Human, mouse, and rat genome large-scale rearrangements: stability versus speciation. *Genome Res.* **14**, 1851–60.

Zhao, S., Jing, W., Samuels, D. C., Sheng, Q., Shyr, Y. and Guo, Y. (2018). Strategies for processing and quality control of Illumina genotyping arrays. *Brief. Bioinform.* **19**, 765–775.

Zhu, Y., Richardson, J. E., Hale, P., Baldarelli, R. M., Reed, D. J., Recla, J. M., Sinclair, R., Reddy, T. B. K. and Bult, C. J. (2015). A unified gene catalog for the laboratory mouse reference genome. *Mamm. Genome* **26**, 295–304.

Appendix A. Supplementary Figures

Figure A.1 Abnormalities in two MDGA raw intensity CEL file images


SNP_mDIV_A6-651_102109.CEL - exprs
SNP_mDIV_D10-652_91809.CEL - exprs

CEL file raw array intensity images were analyzed for quality control purposes and abnormalities in array images were noted for two CEL files. The two samples were not removed from analysis.

Table A1 Reference set of 114 classical inbred laboratory mouse strains used to train the genotyping algorithm employed by Affymetrix Power Tools Software

114 Reference Set MDGA data (CEL) Files	Mouse Sample Strain Name	Sex of Organism	SNP Genetic Distance[a] from C57BL/6J Reference Mouse
SNP_mDIV_A7-7_081308.CEL	C57BL/6J	Male	0.004
SNP_mDIV_A4-SNP08_002_103008.CEL	C57BL/6J	Male	0.006
SNP_mDIV_B1-385_012709.CEL	C57BL/6NCI	Male	0.006
SNP_mDIV_A9-382_012709.CEL	C57BL/6NCI	Male	0.006
SNP_mDIV_B1-SNP08_004_103008_4.CEL	C57BL/6NJ	Male	0.006
SNP_mDIV_A11-384_012709.CEL	C57BL/6Tc	Male	0.006
SNP_mDIV_A9-SNP08_003_103008.CEL	C57BL/6NJ	Female	0.007
SNP_mDIV_A8-381_012709.CEL	C57BL/6Crl	Male	0.007
SNP_mDIV_A10-SNP08_004_103008.CEL	C57BL/6NJ	Male	0.007
SNP_mDIV_A1-SNP08_001_103008.CEL	C57BL/6J	Female	0.007
SNP_mDIV_A11-SNP08_004_103008.CEL	C57BL/6NJ	Male	0.007

[a] SNP genetic distance values calculated through pairwise comparison of SNP genotypes at 493,290 genomic loci queried by the MDGA

SNP_mDIV_A6-SNP08_002_103008.CEL	C57BL/6J	Male	0.007
SNP_mDIV_A3-SNP08_001_103008.CEL	C57BL/6J	Female	0.007
SNP_mDIV_A5-SNP08_002_103008.CEL	C57BL/6J	Male	0.008
SNP_mDIV_A7-SNP08_003_103008.CEL	C57BL/6NJ	Female	0.008
SNP_mDIV_A8-SNP08_003_103008.CEL	C57BL/6NJ	Female	0.009
SNP_mDIV_A2-SNP08_001_103008.CEL	C57BL/6J	Female	0.010
SNP_mDIV_A10-383_012709.CEL	C57BL/6Tc	Male	0.011
SNP_mDIV_A5-378_121608.CEL	C57BL/6J	Male	0.017
SNP_mDIV_B8-85_090908.CEL	C57BL/10J	Male	0.023
SNP_mDIV_D2-SNP09_024_022709.CEL	C57BLKS/J	Male	0.072
SNP_mDIV_A4-150_111308_2.CEL	SSL/LeJ	Male	0.099
SNP_mDIV_B11-88_090908.CEL	C57L/J	Male	0.101
SNP_mDIV_B9-86_090908.CEL	C57L/J	Male	0.102
SNP_mDIV_B4-118_091708.CEL	AEJ/GnLeJ	Male	0.103
SNP_mDIV_D3-129_090908.CEL	CHMU/LeJ	Male	0.104

SNP_mDIV_B8-392_012709.CEL	AEJ/GnRk	Male	0.104
SNP_mDIV_B10-87_090908.CEL	C57BR/cdJ	Male	0.108
SNP_mDIV_C2-91_090908.CEL	JE/LeJ	Male	0.110
SNP_mDIV_B10-394_012709.CEL	BXSB/MpJ	Male	0.113
SNP_mDIV_C1-89_090908.CEL	C58/J	Male	0.113
SNP_mDIV_D6-412_012709.CEL	STX/Le	Male	0.125
SNP_mDIV_A7-153_111308.CEL	TKDU/DnJ	Male	0.139
SNP_mDIV_C6-401_012709.CEL	LT/SvEiJ	Male	0.145
SNP_mDIV_A9-155_111308.CEL	ZRDCT Rax<ey1>/ChUmdJ	Male	0.146
SNP_mDIV_D11-146_103008_3.CEL	SH1/LeJ	Male	0.165
SNP_mDIV_B1-432_022709.CEL	ISS/IbgTejJ	Male	0.167
SNP_mDIV_A1-147_111308.CEL	SI/Col Tyrp1 Dnahc11/J	Male	0.168
SNP_mDIV_B5-123_091708.CEL	BPH/2J	Male	0.169
SNP_mDIV_D4-130_090908.CEL	DLS/LeJ	Male	0.169
SNP_mDIV_A8-427_022709.CEL	COLD2	Male	0.174
SNP_mDIV_A7-425_022709.CEL	HOT2	Male	0.175
SNP_mDIV_B9-142_103008_3.CEL	RHJ/LeJ	Male	0.178

SNP_mDIV_B6-124_091708.CEL	BPN/3J	Male	0.180
SNP_mDIV_B10-143_103008_3.CEL	RSV/LeJ	Male	0.180
SNP_mDIV_A6-424_022709.CEL	HOT1	Male	0.183
SNP_mDIV_A9-429_022709.CEL	WSR2	Male	0.187
SNP_mDIV_A5-5_081308.CEL	BTBR T<+> Itpr3<tf>-Fbxl3<Ovtm>/J	Male	0.187
SNP_mDIV_D7-413_012709.CEL	YBR/EiJ	Male	0.188
SNP_mDIV_A11-431_022709.CEL	WSP2	Female	0.189
SNP_mDIV_A3-49_082108.CEL	NOR/LtJ	Male	0.191
SNP_mDIV_B2-433_022709.CEL	ILS/IbgTejJ	Male	0.193
SNP_mDIV_A8-154_111308.CEL	TSJ/LeJ	Male	0.194
SNP_mDIV_A4-4_081308.CEL	BALB/cByJ	Male	0.194
SNP_mDIV_D5-253_111308.CEL	BALB/cJ	Male	0.195
SNP_mDIV_D11-139_090908.CEL	PN/nBSwUmabJ	Male	0.197
SNP_mDIV_B3-316_120908.CEL	BPL/1J	Male	0.197
SNP_mDIV_C5-94_090908.CEL	NZL/LtJ	Male	0.198

SNP_mDIV_D10-145_103008_3.CEL	SEA/GnJ	Male	0.198
SNP_mDIV_D3-409_012709.CEL	SEC/1ReJ	Male	0.199
SNP_mDIV_D2-408_012709.CEL	SEC/1GnLeJ	Male	0.199
SNP_mDIV_D9-136_090908.CEL	NU/J	Male	0.200
SNP_mDIV_B9-393_012709.CEL	BDP/J	Male	0.200
SNP_mDIV_A2-48_082108.CEL	NON/ShiLtJ	Male	0.200
SNP_mDIV_D9-144_103008_3.CEL	SB/LeJ	Male	0.201
SNP_mDIV_D5-411_012709.CEL	ST/bJ	Male	0.201
SNP_mDIV_B9-138_091708.CEL	P/J	Male	0.201
SNP_mDIV_A6-152_111308.CEL	TALLYHO/JngJ	Male	0.202
SNP_mDIV_C11-406_012709.CEL	NONcNZO5/LtJ	Male	0.202
SNP_mDIV_C7-31_081308.CEL	NZO/HlLtJ	Male	0.202
SNP_mDIV_D4-410_012709.CEL	SJL/Bm	Male	0.203
SNP_mDIV_D1-36_081308.CEL	SJL/J	Male	0.204
SNP_mDIV_B10-21_081308.CEL	FVB/NJ	Male	0.204
SNP_mDIV_A3-3_081308.CEL	AKR/J	Male	0.204
SNP_mDIV_C9-120_090908.CEL	ALS/LtJ	Male	0.205

SNP_mDIV_C3-92_090908.CEL	LG/J	Male	0.205
SNP_mDIV_C8-97_090908.CEL	RIIIS/J	Male	0.205
SNP_mDIV_A5-151_111308.CEL	SWR/J	Male	0.206
SNP_mDIV_A6-119_090908.CEL	ALR/LtJ	Male	0.206
SNP_mDIV_C11-125_090908.CEL	BUB/BnJ	Male	0.207
SNP_mDIV_B9-20_081308.CEL	DDY/JclSidSeyFrkJ	Male	0.207
SNP_mDIV_A8-56_082108.CEL	DDK/Pas	Female	0.207
SNP_mDIV_A1-50_091708.CEL	NZB/BlNJ	Male	0.208
SNP_mDIV_B11-141_091708.CEL	RF/J	Male	0.208
SNP_mDIV_C9-404_012709.CEL	NOD/ShiLtJ	Male	0.209
SNP_mDIV_B4-15_081308.CEL	CBA/CaJ	Male	0.210
SNP_mDIV_A2-148_111308.CEL	SM/J	Male	0.210
SNP_mDIV_C6-30_081308.CEL	NOD/ShiLtJ	Male	0.210
SNP_mDIV_D5-131_090908.CEL	EL/SuzSeyFrkJ	Male	0.210
SNP_mDIV_C3-398_012709.CEL	DBA/1LacJ	Male	0.210
SNP_mDIV_D7-134_090908.CEL	MRL/MpJ	Male	0.211
SNP_mDIV_B7-18_081308.CEL	DBA/1J	Male	0.212
SNP_mDIV_B8-132_091708.CEL	HPG/BmJ	Male	0.212

SNP_mDIV_C6-95_090908.CEL	PL/J	Male	0.213
SNP_mDIV_D1-126_090908.CEL	C3HeB/FeJ	Male	0.213
SNP_mDIV_B8-19_081308.CEL	DBA/2J	Male	0.213
SNP_mDIV_A6-6_081308.CEL	C3H/HeJ	Male	0.213
SNP_mDIV_D10-137_090908.CEL	NZM2410/J	Male	0.214
SNP_mDIV_B7-391_012709.CEL	A/WySnJ	Male	0.214
SNP_mDIV_D8-256_111308.CEL	CBA/J	Male	0.215
SNP_mDIV_C5-400_012709.CEL	DBA/2HaSmnJ	Male	0.215
SNP_mDIV_A2-2_081308.CEL	A/J	Male	0.215
SNP_mDIV_D6-254_111308.CEL	129X1/SvJ	Male	0.216
SNP_mDIV_C8-32_081308.CEL	NZW/LacJ	Male	0.217
SNP_mDIV_A1-1_081308.CEL	129S1/SvImJ	Male	0.218
SNP_mDIV_B5-389_012709.CEL	129T2/SvEmsJ	Male	0.219
SNP_mDIV_B2-90_091708.CEL	I/LnJ	Male	0.219
SNP_mDIV_C4-93_090908.CEL	LP/J	Male	0.221
SNP_mDIV_A8-199_091708.CEL	129S6	Male	0.221
SNP_mDIV_D2-128_090908.CEL	CE/J	Male	0.222
SNP_mDIV_B3-387_022709.CEL	129P1/ReJ	Male	0.222

SNP_mDIV_B11-22_081308.CEL	KK/HlJ	Male	0.222
SNP_mDIV_C4-399_012709.CEL	DBA/2DeJ	Female	0.224
SNP_mDIV_B4-388_012709.CEL	129P3/J	Male	0.225

Table A2 Percentage of loci genotyped and the percentage of genotyped loci with a heterozygous genotype for samples of the inter-order genotyping set (n = 40)[a]

MDGA Data (CEL) File	Sample Scientific Name	Loci Genotyped (%)	Heterozygosity (%)
SNP_A1-GES11_4902_AGT-JLP-120115-24-35517	*T. pinchaque*	89.5	71.7
SNP_A2-GES11_4907_AGT-JLP-120115-24-35517	*D. bicornis*	89.6	72.8
SNP_mDIV_B8-1190_082410	*A. sylvaticus*	89.7	69.4
SNP_mDIV_B9-667_102109	Sciuridae[b]	89.8	73.5
SNP_mDIV_B2-660_102109	*P. melanophrys*	90.1	71.7
SNP_mDIV_D7-473_012209	*M. pahari*	90.4	61.2
SNP_mDIV_B3-661_102109	*P. californicus*	90.5	73.4
SNP_mDIV_B4-662_102109	*P. m. sonoriensis*	90.5	72.8
SNP_mDIV_B5-663_102109	*P. m. bairdii*	90.6	74.4

[a] Samples organized according to increasing percentage of loci genotyped

[b] Only family level classification information available for this sample

SNP_mDIV_B6-664_102109	*P. polionotus*	90.6	74.4
SNP_mDIV_A9-656_102109	*R. norvegicus*	90.6	71.0
SNP_mDIV_B1-659_102109	*P. aztecus*	90.7	73.6
SNP_mDIV_A10-657_102109	*R. norvegicus*	90.7	71.3
SNP_mDIV_D3-639_91809	*M. m. castaneus*	90.7	14.7
SNP_mDIV_B8-666_102109	*P. leucopus*	90.7	73.7
SNP_mDIV_A6-651_102109	*M. saxicola*	90.8	56.3
SNP_mDIV_A7-654_102109	*M. n. mattheyi*	90.8	59.9
SNP_mDIV_D6-472_012209	*M. caroli*	90.9	45.6
SNP_mDIV_A4-649_102109	*M. platythrix*	91.1	54.0
SNP_mDIV_A3-648_102109	*M. platythrix*	91.2	57.6
SNP_mDIV_D4-640_91809	*M. famulus*	91.2	39.6
SNP_mDIV_A5-650_102109	*M. saxicola*	91.2	55.0

SNP_mDIV_A2-645_102109	*M. cookii*	91.2	41.5
SNP_mDIV_D11-653_91809	*M. n. minutoides*	91.5	61.0
SNP_mDIV_D10-652_101509-redo	*M. n. orangiae*	91.5	58.8
SNP_mDIV_D11-653_101509-redo	*M. n. minutoides*	91.5	60.2
SNP_mDIV_D7-644_91809	*M. caroli*	91.5	43.6
SNP_mDIV_D10-652_91809	*M. n. orangiae*	91.6	60.4
SNP_mDIV_D9-647_101509-redo	*M. dunni*	91.8	36.2
SNP_mDIV_D3-639_101509-redo	*M. m. castaneus*	91.9	14.9
SNP_mDIV_D9-647_91809	*M. dunni*	91.9	35.6
SNP_mDIV_D7-644_101509-redo	*M. caroli*	92.0	42.4
SNP_mDIV_D4-640_101509-redo	*M. famulus*	92.2	36.7

SNP_mDIV_D8-646_91809	*M. cervicolor*	92.2	39.8
SNP_mDIV_D6-643_91809	*M. fragilicauda*	92.3	37.6
SNP_mDIV_D8-646_101509-redo	*M. cervicolor*	92.4	40.4
SNP_mDIV_D8-474_012209	*M. famulus*	92.5	35.0
SNP_mDIV_D5-642_91809	*M. fragilicauda*	92.9	36.9
SNP_mDIV_D6-643_101509-redo	*M. fragilicauda*	93.3	35.8
SNP_mDIV_D5-642_101509-redo	*M. fragilicauda*	93.4	35.9

Table A3 Percentage of loci genotyped and the percentage of genotyped loci with a heterozygous genotype for samples of the intra-genus genotyping set (n = 27)[a]

MDGA Data (CEL) File	Sample Scientific Name	Loci Genotyped (%)	Heterozygosity (%)
SNP_mDIV_D7-473_012209.CEL	*M. pahari*	83.1	53.1
SNP_mDIV_A7-654_102109.CEL	*M. n. mattheyi*	84.3	52.5
SNP_mDIV_A4-649_102109.CEL	*M. platythrix*	85.1	47.0
SNP_mDIV_A6-651_102109.CEL	*M. saxicola*	85.3	49.6
SNP_mDIV_A3-648_102109.CEL	*M. platythrix*	85.4	50.8
SNP_mDIV_D11-653_91809.CEL	*M. n. minutoides*	85.5	53.9
SNP_mDIV_D10-652_91809.CEL	*M. n. orangiae*	85.8	53.5
SNP_mDIV_A5-650_102109.CEL	*M. saxicola*	86.0	48.6
SNP_mDIV_D10-652_101509-redo.CEL	*M. n. orangiae*	86.0	52.3
SNP_mDIV_D11-653_101509-redo.CEL	*M. n. minutoides*	86.1	53.8
SNP_mDIV_D6-472_012209.CEL	*M. caroli*	87.7	40.7

[a] Samples organized according to increasing percentage of loci genotyped

SNP_mDIV_A2-645_102109.CEL	*M. cookii*	88.0	37.0
SNP_mDIV_D7-644_91809.CEL	*M. caroli*	88.8	39.2
SNP_mDIV_D7-644_101509-redo.CEL	*M. caroli*	89.0	37.9
SNP_mDIV_D4-640_91809.CEL	*M. famulus*	89.1	35.5
SNP_mDIV_D8-646_91809.CEL	*M. cervicolor*	89.2	35.4
SNP_mDIV_D9-647_91809.CEL	*M. dunni*	89.3	31.4
SNP_mDIV_D9-647_101509-redo.CEL	*M. dunni*	89.5	32.2
SNP_mDIV_D8-646_101509-redo.CEL	*M. cervicolor*	89.5	36.0
SNP_mDIV_D8-474_012209.CEL	*M. famulus*	89.9	31.0
SNP_mDIV_D4-640_101509-redo.CEL	*M. famulus*	90.3	33.0
SNP_mDIV_D6-643_91809.CEL	*M. fragilicauda*	91.0	34.4
SNP_mDIV_D5-642_91809.CEL	*M. fragilicauda*	91.3	33.6
SNP_mDIV_D6-643_101509-redo.CEL	*M. fragilicauda*	91.4	32.4
SNP_mDIV_D5-642_101509-redo.CEL	*M. fragilicauda*	91.6	32.6
SNP_mDIV_D3-639_91809.CEL	*M. m. castaneus*	91.7	13.3

SNP_mDIV_D3-639_101509-redo.CEL	*M. m. castaneus*	93.0	13.5

Table A4 Differences in percentage of loci genotyped in Mus samples included in the inter-order genotyping set and the intra-genus genotyping set[a]

MDGA Data (CEL) File Name	Scientific Name of Species	Loci Genotyped % (Inter-Order Set)	Loci Genotyped % (Intra-Genus Set)	Difference Between Inter-Order Set & Intra-Genus Set
SNP_mDIV_D3-639_101509-redo	*M. castaneus*	91.9	93.0	-1.1[b]
SNP_mDIV_D3-639_91809	*M. castaneus*	90.7	91.7	-1.0
SNP_mDIV_D6-643_91809	*M. fragilicauda*	92.3	91.0	1.3
SNP_mDIV_D5-642_91809	*M. fragilicauda*	92.9	91.3	1.6
SNP_mDIV_D5-642_101509-redo	*M. fragilicauda*	93.4	91.6	1.8
SNP_mDIV_D6-643_101509-redo	*M. fragilicauda*	93.3	91.4	1.9
SNP_mDIV_D4-640_101509-redo	*M. famulus*	92.2	90.3	1.9
SNP_mDIV_D4-640_91809	*M. famulus*	91.2	89.9	2.0

[a] Samples organized by increasing difference between percentage of loci genotyped in the inter-order test set vs intra-genus test set.

[b] Negative difference values indicate an increase in percentage of loci genotyped for a sample in the intra-genus set compared to the same sample in the inter-order set.

SNP_mDIV_D9-647_101509-redo	*M. dunni*	91.8	89.5	2.3
SNP_mDIV_D8-474_012209	*M. famulus*	92.5	89.9	2.6
SNP_mDIV_D9-647_91809	*M. dunni*	91.9	89.3	2.6
SNP_mDIV_D7-644_91809	*M. caroli*	91.5	88.8	2.7
SNP_mDIV_D8-646_101509-redo	*M. cervicolor*	92.4	89.5	2.9
SNP_mDIV_D8-646_91809	*M. cervicolor*	92.2	89.2	3.0
SNP_mDIV_D7-644_101509-redo	*M. caroli*	92.0	89.0	3.0
SNP_mDIV_D6-472_012209	*M. caroli*	90.9	87.7	3.2
SNP_mDIV_A2-645_102109	*M. cookii*	91.2	88.0	3.2
SNP_mDIV_A5-650_102109	*M. saxicola*	91.2	86.0	5.2
SNP_mDIV_D11-653_101509-redo	*M. n. minutoides*	91.5	86.1	5.4
SNP_mDIV_A6-651_102109	*M. saxicola*	90.8	85.3	5.5

SNP_mDIV_D10-652_101509-redo	*M. n. orangiae*	91.5	86.0	5.5
SNP_mDIV_D10-652_91809	*M. n. orangiae*	91.6	85.8	5.8
SNP_mDIV_A3-648_102109	*M. platythrix*	91.2	85.4	5.8
SNP_mDIV_A4-649_102109	*M. platythrix*	91.1	85.1	6.0
SNP_mDIV_D11-653_91809	*M. n. minutoides*	91.5	85.5	6.0
SNP_mDIV_A7-654_102109	*M. n. mattheyi*	90.8	84.3	6.5
SNP_mDIV_D7-473_012209	*M. pahari*	90.4	83.1	7.3

Table A5 Differences in percentage of loci genotyped in Mus samples included in the inter-family genotyping set and the intra-genus genotyping set[a]

MDGA Data (CEL) File Name	Species Scientific Name	Loci Genotyped (%) Inter-Family Set	Loci Genotyped (%) Intra-Genus Set	Difference Between Inter-Family & Intra-Genus Sets
SNP_mDIV_D3-639_101509-redo	*M. castaneus*	92.7	93.0	-0.3[b]
SNP_mDIV_D3-639_91809	*M. castaneus*	91.5	91.7	-0.2
SNP_mDIV_D6-643_91809	*M. fragilicauda*	91.8	91.0	0.8
SNP_mDIV_D5-642_91809	*M. fragilicauda*	92.1	91.3	0.8
SNP_mDIV_D6-643_101509-redo	*M. fragilicauda*	92.3	91.4	0.9
SNP_mDIV_D5-642_101509-redo	*M. fragilicauda*	92.4	91.5	0.9
SNP_mDIV_D4-640_101509-redo	*M. famulus*	91.2	90.3	0.9
SNP_mDIV_D4-640_91809	*M. famulus*	90.1	89.1	1.0
SNP_mDIV_D9-647_101509-redo	*M. dunni*	90.5	89.5	1.0
SNP_mDIV_D9-647_91809	*M. dunni*	90.4	89.3	1.1
SNP_mDIV_D8-474_012209	*M. famulus*	91.0	89.9	1.1

[a] Samples organized by increasing difference between percentage of loci genotyped in the inter-family test set vs intra-genus test set.

[b] Negative difference values indicate an increase in percentage of loci genotyped for a sample in the intra-genus set compared to the same sample in the inter-family set.

SNP_mDIV_D8-646_101509-redo	*M. cervicolor*	90.9	89.5	1.4
SNP_mDIV_D8-646_91809	*M. cervicolor*	90.6	89.2	1.4
SNP_mDIV_D7-644_91809	*M. caroli*	90.2	88.8	1.4
SNP_mDIV_A2-645_102109	*M. cookii*	89.4	88.0	1.4
SNP_mDIV_D7-644_101509-redo	*M. caroli*	90.4	89.0	1.4
SNP_mDIV_D6-472_012209	*M. caroli*	89.1	87.7	1.4
SNP_mDIV_A5-650_102109	*M. saxicola*	88.3	86.0	2.3
SNP_mDIV_A6-651_102109	*M. saxicola*	87.7	85.3	2.4
SNP_mDIV_D10-652_101509-redo	*M. n. orangiae*	88.5	86.0	2.5
SNP_mDIV_A3-648_102109	*M. platythrix*	87.9	85.4	2.5
SNP_mDIV_A4-649_102109	*M. platythrix*	87.6	85.1	2.5
SNP_mDIV_D10-652_91809	*M. n. orangiae*	88.3	85.8	2.5
SNP_mDIV_D11-653_101509-redo	*M. n. minutoides*	88.7	86.1	2.6
SNP_mDIV_D11-653_91809	*M. n. minutoides*	88.1	85.5	2.6
SNP_mDIV_A7-654_102109	*M. n. mattheyi*	87.2	84.3	2.9
SNP_mDIV_D7-473_012209	*M. pahari*	86.2	83.1	3.1

Table A6 Fisher's Exact[a] test of significance of genotypic composition and allelic frequencies across genotyping sets

Genotyping Sets	Genotyping p-value	Allelic Frequency p-value
Intra-Genus (Mus)	<0.0001	<0.0001
Inter-Genus (Mus + Apodemus + Peromyscus + Rattus)	<0.0001	<0.0001
R. norvegicus[b]	0.09336	0.2232
Inter-Family (Mus + *H. glaber*)	<0.0001	<0.0001
Inter-Order	<0.0001	<0.0001
H. glaber[c]	<0.0001	<0.0038

[a] Nonparametric, unordered Fisher-Freeman-Halton exact test (Monte Carlo Simulation) using Statexact (Cytel Studio)

[b] Results for *Rattus* samples (n = 2) were obtained from the inter-genus genotyping set

[c] *Heterocephalus glaber* samples (n = 4) were genotyped separately from other samples

Table A7 MDGA SNP loci with heterozygous genotypes and with perfect probe sequence matches in publicly available genome[a] sequences

Number of	*M. caroli*	*M. pahari*	*R. norvegicus*	*P. maniculatus*	*H. glaber*
Number of samples genotyped	3	1	2	2	4
Loci with heterozygous genotypes[b]	147,4 52	251,90 2	85,926	143,971	91,324
Loci with probe sequences in the publicly available genome sequence with a heterozygous genotype	9,413	9,341	1,019	481	52

[a] Genomes accessed through the NCBI Genomes FTP site of samples under study (ftp.ncbi.nlm.nih.gov/genomes/)

[b] If more than one sample was genotyped per species, the loci must have heterozygous genotypes in all samples

Appendix B. Online Distance Matrices

Please see Appendix B online for large distance matrix data: (https://www.dropbox.com/sh/keuszhh8a0ornob/AABk5a0aMM4HEDqSyFnP2R8Oa?dl=0)

Appendix C. R scripts

SNP Spatial-Temporal Plot (SNPSTeP) Code for R

This code will visualize Single Nucleotide Polymorphism (SNP) genotype changes across the genome as well as changes to genotypes at particular positions as evolutionary time increases from the model species (house mouse in this study).

```
#Set the working directory. I set it to my desktop
setwd("/Users/Your_Directory_Here")

# Read in the csv file with data.
# There is a header line in data, so header = TRUE
# I assigned my csv data to the name musstackSNPs
musstackSNPs <- read.csv('/Users/Your_Directory_Here/File_Name.csv', he
ader  =  TRUE)

# Assign the SNP state column from my musstackSNPs dataframe as a facto
r. Stored the four possible genotype results (-1 or No Call, 0 or AA, 1
 or AB, 2 or BB) as levels
#SNPstate <- factor(musstackSNPs$SNP_State, levels = c("-1", "0", "1",
"2"))
#change colours of SNP state by assigning new numbers corresponding wit
h colour

SNPstate <- musstackSNPs$SNP_State
SNPstate[SNPstate  =  =  1] <- 5 #blue
SNPstate[SNPstate  =  =  -1] <- 1 #black
SNPstate[SNPstate  =  =  0] <- 8 #grey
SNPstate[SNPstate  =  =  2] <- 6 #pink



# Assign the data from the Name column from my musstackSNPs dataframe a
s a factor. Stored the eight Mus species I examined as levels
musstackSNPs$Name <- factor(musstackSNPs$Name, levels  =  c("M. musculu
s", "M. m. castaneus 1", "M. m. castaneus 2", "M. dunni 1", "M. dunni 2
", "M. famulus 1", "M. famulus 2", "M. famulus 3", "M. fragilicauda 1",
 "M. fragilicauda 2", "M. fragilicauda 3", "M. fragilicauda 4", "M. car
oli 1", "M. caroli 2", "M. caroli 3", "M. cervicolor 1", "M. cervicolor
 2", "M. cookii"))

# Adjusted plot parameters. Added space to the left margin by increasin
g second value in mar vector to 7.
# Adujsted the axis label locations (mgp) (first value in vector (origi
nal 3 changed to 4)) to move them further away from the inner axis labe
l
# Set xpd = NA to allow for adding a legend outside of the plot area
```

```
par(mar =  c(5,7,4,2),mgp = c(4,1,0), xpd =  NA)

# Create a plot. X axis is genome position & y axis will be the associa
ted species names
plot(
  musstackSNPs$Location,musstackSNPs$Name,
  main  =  "Your Title Here", #title of plot. This plot displays SNPs o
n a chromosome
  yaxt  =  'n', #Use this option to not display the y axis ticks and la
bels
  ylab  =  "Your species", # y axis label
  xlab  =  "Genome Position (bp)", #x axis label
  xlim  =  c(genomic_start_position, genomic_end_position), #sets range
 for x axis. Put base-pair value of genomic start and end position of c
hromosome for species of interest
  pch = 20, #sets the plot marker shape -- circle
  col = SNPstate # Colour the plot points by SNP state factor
  )

# Next line allows axis labels to be printed horizontally. value of 1
=  horizontal always.
par(las = 1)

# add y axis in. value of 2 represents y axis. use 'at' to add labels a
t a regular sequence from 1-8 becuase I have 8 mice samples. I added a
vector of the mouse species' names as the tick labels.
#I adjusted the axis font size to be smaller using cex.axis
axis(2, at = seq(1:18),
     labels  =  c("M. musculus", "M. m. castaneus 1", "M. m. castaneus
2", "M. dunni 1", "M. dunni 2", "M. famulus 1", "M. famulus 2", "M. fam
ulus 3", "M. fragilicauda 1", "M. fragilicauda 2", "M. fragilicauda 3",
 "M. fragilicauda 4", "M. caroli 1", "M. caroli 2", "M. caroli 3", "M.
cervicolor 1", "M. cervicolor 2", "M. cookii"),
     cex.axis = 0.5
     )

#Add a legend.
#legend is comprised of the four possible MDGA genotype results (-1, 0,
 1, 2)
legend(-2829834,20.94821,
       legend  =  c("No Call", "AA", "AB", "BB"),
       pch  =  20, #Set legend symbols
       ncol  =  2, # split genotype symbols and corresponding colours i
n two columns
       cex  =  0.75, # reduced size of legend
       col  =  c(1, 8, 5, 6) #added colours of genotype values
       )
```

Appendix D. Online *In silico* genome matches and Ensembl Gene ID matches

Please see Appendix D online for *in silico* MDGA probe matches obtained using E-MEM and associated Ensembl gene ID lists. (https://www.dropbox.com/sh/ma2gwckh9ik711h/AADcd0f8Kr9pCNUcaSYZaGnya?dl = 0)

Appendix E. Top DAVID functional associations

Please see online appendix E for full list of enriched KEGG pathways (p<0.001) from DAVID functional annotation tool results. (https://www.dropbox.com/sh/la2jzk26519ltu7/AAC4xUW3tZKFGABjd46zXu7Ua?dl = 0)

7 *Curriculum Vitae*

Rachel Kelly

Department of Biology
The University of Western Ontario, 1151 Richmond Street, London, ON

Education

MSc Biology 2017-2019
Cell and Molecular Biology, Western University, London, ON

BSc Honours Specialization: Genetics 2013-2017
Western University, London, ON

Published Work

Environmental Mutagenesis and Genomics Society 49th Annual Meeting (59) 13 2018

- Published abstract in Environmental and Molecular Mutagenesis (EMM) Journal for the international Environmental Mutagenesis and Genomics Society Annual Meeting

Presentations at Scientific Meetings

EMGS 49th Annual Meeting, San Antonio, USA 2018

- Presented a talk based on MSc thesis work titled “Adapting a Mouse Genotyping Technology to Survey Dynamics of Rodent Genomic Diversity”

3 Minute Thesis Western, University of Western Ontario, CA 2018

- Presented MSc thesis work at the 3 Minute Thesis Western 2018 competition “Into the Wild: Adapting genotyping technology to monitor the pulse of the planet”

Fallona Family Research Showcase, University of Western Ontario, CA 2018

- Presented MSc thesis work titled “From the lab to the field: Applying a lab mouse genetic technology to wild species” as an interdisciplinary poster presentation

Biology Graduate Research Forum Lightning Talk, University of Western Ontario, CA 2017

- Delivered a talk on my MSc thesis work titled “Pioneering use of mouse genotyping arrays as an instrument for comparative genomics” in the cell and molecular division of the BGRF lightning talk competition

Western Biology Day, University of Western Ontario, CA 2017

- Presented a research talk based on honours thesis research at Western Biology Day 2017 titled “Assessing Genetic Diversity Cross-Species Using the Mouse Diversity Genotyping Array: Can I Use Rhino DNA?”

Ontario Biology Day, Laurentian University, CA 2017

- Presented a research talk based on honours thesis research at Ontario Biology Day 2017 conference in Sudbury, Ontario

Awards and Honours

EMGS Travel Award ($750 USD) 2018

- Competitive research award based on abstract review granted by the Environmental Mutagenesis and Genomics Society for the 2018 EMGS 49th Annual Meeting in San Antonio, Texas

Departmental Travel Award ($150 CAD) 2018

- Competitive research award based on abstract review granted by the Department of Biology, Western University for the 2018 EMGS Annual Meeting

Best Lightning Talk ($75 CAD) - Biology Graduate Research Forum (BGRF) 2017

- Awarded top oral presentation at the 2017 Biology Graduate Research Forum at Western University

Western Scholarship of Excellence ($2,000 CAD) 2013

- Grade-based entrance scholarship awarded by Western University

Science Case Competition Certificate of Achievement 2013

- Certificate of achievement – placed in the top 25% of the Western Science Case Competition

Teaching Experience

Graduate Teaching Assistant 2017-2019

Western University, London, ON

- Third-year Human Genetics:
  - Nominated for a teaching assistant award
  - Designed lectures, led workshops, and taught core genetics concepts and techniques in tutorials to a class of 150 students; proctored and marked midterms and final exams
- Third-year Genetics: DNA Organization, Mutagenesis, and Repair
  - Nominated for a teaching assistant award
  - Designed lectures, led workshops, and taught genetic concepts and mutation detection techniques in tutorials to a class of 80 students; proctored and marked midterms and final exams
- First-year biology:
  - Taught core lab concepts, led pre-lab discussions, and marked assignments for groups of 20-30 students at a time
  - Proctored and marked midterms and final exams

Department of Biology Service

Chair and Judge of Ontario Biology Day Conference 2019

Western University, London, ON

- Acted as a chair and judge of undergraduate talks at the Ontario Biology Day conference for the categories of cell and molecular biology and genetics.

Western Synthetic Biology Symposium Conference Organization 2017, 2018

Western University, London, ON

- Organized and ran registration at the symposium, and acted as a liaison with representatives of industry for the 2017 Western Synthetic Biology Symposium
- Organized and coordinated judging of posters and talks, and acted as a liaison for representatives of industry at the 2018 Western Synthetic Biology Symposium

Volunteer Mentor of Western Synthetic Biology Case Competition 2018

Western University, London, ON

- Mentored a group of four undergraduate students on a weekly basis in creating a synthetic biology solution to solve a world problem

Volunteer/Guide at Science Fall Preview Day 2018

Western University London, ON

- Mentored incoming undergraduate students and aided in successful outreach services for the Department of Biology