7-23-2013 12:00 AM

Array-based genomic diversity measures portray Mus musculus phylogenetic and genealogical relationships, and detect genetic variation among C57Bl/6J mice and between tissues of the same mouse

Susan T. Eitutis, *The University of Western Ontario*

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

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

Array-based genomic diversity measures portray *Mus musculus* phylogenetic and genealogical relationships, and detect genetic variation among C57Bl/6J mice and between tissues of the same mouse

(Thesis format: Monograph)

by

Susan T Eitutis

Graduate Program in Biology

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

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

Abstract

Mouse models lack affordable genomic technologies slowing the identification of candidate variants contributing to complex phenotypes. The Mouse Diversity Genotyping Array (MDGA) is a low cost, high-resolution platform permitting genomic diversity assessment. Using a validated list of >500,000 single nucleotide polymorphisms (SNPs), we applied the first comprehensive analysis of SNP differences to detect genetic distance across 362 *Mus musculus* samples. Genetic distance measured between distantly and closely related mice correlates with known phylogeny and genealogy. Variation detected between C57BL/6J mice is consistent with previous reports of variants within this strain. Putative genetic variation detected between and within tissues indicates somatic mosaicism. Genotype differences detected within a mouse are a complex mixture of technical errors and biological differences. Detailed reconstruction experiments are therefore required to determine array sensitivity at detecting true biological variants. The MDGA shows promise for analyzing mutation accumulation with development, aging, environmental mutagenesis and diseases such as cancers.

Keywords

Single nucleotide polymorphism, Genotyping, Genomic, Mutation, Genetic Distance, Genetic Background, *Mus musculus*, Mouse Diversity Genotyping Array, genetic variation, somatic mosaicism.

Co-authorship statement

Susan Eitutis completed this work under the supervision and financial support from Dr. Kathleen Allen Hill. Susan completed the work presented in the thesis with assistance in the creation of phylogenetic trees and distance matrices. Marjorie E Osborne Locke, assisted with coding used to generate the distance matrices and phylogenetic trees. Mark Daley performed Mantel tests to determine if the genetic distance matrices were significantly different between genotyping lists and proposed the analysis of SNP list permutations. Andrea Wishart provided instruction in the use of Circos, a visualization program for the mouse genome.

Acknowledgements

The journey to completing my Masters thesis was not without its difficulties, and my success would not have been achieved without the help of many influential people along the way. I would like to thank first and foremost my supervisor Dr. Kathleen Hill for all the guidance and support she has given me over the last two years. Her patience and encouragement has enabled me to successfully complete a rather unique project, establish a foundation for bioinformatics research in her laboratory, and helped me effectively articulate my research to others in the scientific community.

I would like to thank my advisory committee Dr. Shiva Singh and Dr. Mark Daley, who always provided excellent advice when I encountered difficulties over the course of my thesis. I would like to thank David Carter for the microarray processing and DNA isolation that was key to the completion of my project.

I would like to thank all the past and present members of the Hill Lab for all the good times that we had together. To my friends and family who stuck by me through all the ups and downs of this project, I couldn’t have done it without you.

To my best lab buddy and project partner Andrea Wishart, I would not have been able to complete this project without you. You were the support that I needed when everything was falling apart, and the person who was there to celebrate with me when things actually worked. I cannot imagine having completed this with anyone else, thank you!

The thesis research was supported by funding to Dr. Kathleen Hill from the Natural Sciences and Engineering Research Council of Canada and the Canadian Institutes of Health Research.

Table of Contents

List of Figures

List of Tables

List of Appendices

List of Abbreviations

aCGH	array Comparative Genomic Hybridization
Aif	*Apoptosis-inducing factor*
B6	C5BL/6J
bp	base pairs
BWA	Burrows-Wheeler Alignment
CC	Collaborative Cross
CNV	Copy Number Variant
DO	Diversity Outbred
hq	*harlequin*
IGP	Invariant Genomic Probe
kb	kilobase
Mbs	megabases
MDGA	Mouse Diversity Genotyping Array
Mhc	*Major histocompatibility complex*
MPD	Mouse Phenome Database
MSM	*Mus musculus molossinus*
ROS	Reactive Oxygen Species
SEM	Standard error of the mean
SNP	Single Nucleotide Polymorphism
WT	Wild-Type

Chapter 1 : Introduction

1.1 General Introduction

The phenotypic variation observed between individuals results from the interaction between genetics and environment. Variation at the genetic level causes different phenotypic outcomes and is one of the primary causes for variation between individuals within a population. The sum total of the genetic variation between individuals arises from inherited and acquired mutations, and the study of the origins and mechanisms of this variation is relevant to understanding development and aging. We know that mutation accumulation begins early in development; however, the extent to which tissues and organisms vary in mutation accumulation remains unknown from a genomic perspective. By expanding mutation research to understand how and where mutations accumulate across the genome in a tissue-specific manner, researchers may be able to identify mutation signatures or patterns with specific phenotypic variations.

1.2 Genetic Variation

Genetic variation refers to differences in the DNA nucleotide sequence that occur between individuals of a population. Genetic variation encompasses a broad spectrum of genetic differences, ranging from modifications affecting single base pairs to those affecting megabases (Mbs – 1,000,000 bps) of DNA. The most common variants in the genome are single nucleotide polymorphisms (SNPs).[1] SNPs are single base pair substitutions that occur within one percent or more of the individuals within a population. However, the discovery of copy number variants (CNVs) has identified a new form of genetic variation that is prominent in both healthy and diseased phenotypes.[2,3] Copy number variants are large segmental duplications or deletions of genomic regions, typically 1 kb or greater, that alter the usual ploidy of the genome.

1.3 The origin of genetic variation: mutations and how they occur

Genetic variation results from mutations occurring during the lifespan of an individual. These mutations arise as a result of DNA damage that is not repaired upon replication.[1] DNA damage can occur from both endogenous and exogenous factors, and can affect both the germline and the soma. When DNA is exposed to mutagenic factors, the structure of the DNA is altered. During DNA replication, this altered structure can result in the changing of a single nucleotide pair, or facilitate mechanisms that induce large modifications. Such mechanisms known to facilitate genomic modifications include single strand breaks, double strand breaks, strand slippage, and fork stalling and template switching result in the addition or deletion of nucleotides from the DNA sequence.[4] The additions and deletions introduced by these mechanisms can modify single base pairs to many Mbs of DNA.

1.4 The nature of genetic variation

Genetic variation can contribute to the distinct differences observed between individuals and within populations. Humans from different ancestries often have distinct characteristics associated with distinct ancestries. These characteristics are the result of mutations becoming fixed within a population. For example, the redheaded phenotype prominent in those of Celtic ancestry resulted from a mutation in the *melanocortin 1 receptor* (*MCR1*) gene.[5] Such mutations are inherited and are often associated with other distinct characteristics of that population such as the association between red hair and fair skin. When specific allelic combinations associated with these characteristics are often inherited together, they are known as haplotypes.

However, differences contributing to genetic variation may result from the accumulation of mutations in somatic tissues.[6] Mutations occurring in somatic tissues are not inherited, but when frequent in the population they are referred to as recurrent mutations.[7,8] Recurrent mutations occur at mutational hotspots where the nucleotide sequence or DNA conformation is more susceptible to DNA damage in comparison to other locations in the genome. Variants resulting from recurrent mutations are often not associated with ancestral haplotypes. Recurrent mutations are often linked with specific phenotypes associated with health and disease, such as those mutations associated with cancers.

1.5 Inheritance of genetic variation

The inheritance of genetic variation was first described by Gregor Mendel in the mid 19th century. By selecting for different phenotypic traits, namely colour and shape of peas, Mendel identified that these traits, each of which were determined by a single gene, were passed between parent and offspring in a manner that could be predicted.[9] Mendel predicted for each of the selected traits, whether peas would be yellow or green in colour, and round or wrinkled in shape, based on the phenotypes of the parental plants. The predictivity of single gene inheritance of binary traits was later coined as Mendelian Genetics. However, not all traits follow the principles of Mendelian genetics.

1.6 Complex Phenotypes

Rather than a single gene directly impacting a phenotype, complex traits arise as a result of many genes contributing small effects, and the interactions between these genes and their environment.[1] Like Mendelian genetics, specific allelic combinations/interactions impact the phenotypic outcome of complex phenotypes. However, because there is no clear cause and effect relationship with complex phenotypes, it is difficult to make an association between a genetic variant and a specific phenotypic outcome. When you have multiple genes interacting to contribute to an overall phenotype, it becomes difficult to dissect out which genetic variants contribute to each specific characteristic of that phenotype. The addition of environmental variation adds an additional level of complexity to understanding complex phenotypes. The environment to which an individual is exposed can impact how substantial the effects of the genetic variants are on the observed phenotype. Therefore, when studying complex traits it is important to minimize the amount of genetic and environmental variation between individuals in order to understand how the interactions between genes and environment affect specific complex traits.

1.7 Genetic variation in human populations

1.7.1 The genetics of complex phenotypes

Studying complex phenotypes in humans has traditionally been accomplished by analyzing diseases. By looking at diseases in populations where many individuals are affected and disease inheritance can be traced, associations between genetic variants and the disease can be identified. Identification of genotype-phenotype associations is easiest in populations that contain minimal genetic and environmental variation. In humans, these populations are uncommon; however, popular case studies include Amish, Ashkenazi Jewish, and Acadian populations. Amish populations have a high prevalence of recessive disorders including those associated with dwarfism, anemia, and epilepsy.[10] Similarly, in Azhkenazis Jewish populations, mutations causing diseases such as Gaucher's, hemophilia, and Tay-Sachs are common.[11–13] Acadians also show prevalence for single gene disorders as a result of the founder effect within these populations.[14,15] The founder effect causes a decrease in genetic diversity ultimately increasing the prevalence of single gene disorders like Tay-Sachs. The limited genetic variation within these populations enables the analysis of haplotypes with minimal genetic variation, making them ideal models to study complex diseases. Therefore, inbred populations are becoming case studies for complex phenotypes because the limited genetic variation simplifies the identification of genotype-phenotype interactions.[12,16]

1.7.2 Single gene detection of genetic variation

The identification of genetic variation has traditionally focused on traits following the principles of Mendelian genetics. Genotype-phenotype associations were used to study medical phenomena by tracing phenotypes through family pedigrees.[17] Some of the earliest association studies completed in humans were conducted to understand the clotting of blood during blood transfusions in the early 1900's.[18] This led to the identification of four blood groups. These four blood groups were later classified as the ABO blood types, and could be used for paternity testing.[19] In 1940, Landsteiner and Wiener identified the genetic basis of the Rh groups which added a positive and negative factor to each of the four blood types, and explained the effects of mother-fetus incompatibility.[20]

1.7.3 Genomic detection of genetic variation

As research advanced, and studies progressed from single gene analysis towards technologies that can query across the genome, the ability to study complex traits became increasingly easier. The major improvement in studying complex traits came with the development of the Human Genome Project. The Human Genome Project, was a 13-year project focused on understanding the human genome and was completed in 2003 with the first full sequence read.[21] With a goal of identifying all genes as well as decoding the entire sequence of the human genome, the Human Genome Project became one of the biggest efforts to improve our understanding of human genetics. By 2002, the human genome project was nearing completion and the next step was to determine how genetic variation affected the genome. This led to the development of the HapMap project that focused on identifying variation and mapping haplotypes across human populations.[22]

Mapping genetic variation across the human population required the participation of individuals from a broad range of ancestries. To understand how genetic variation affects the human population, researchers focused on identifying haplotypes within each ancestral group. By identifying combinations of genes predominant to these populations, researchers could associate specific allelic combinations to specific phenotypes. Researchers could also speculate as to the accumulation of genetic variants within human populations dependent on geographic location and common ancestries. Therefore, when studying complex phenotypes, it is important to understand the ancestral origin of genetic variants and how the accumulation of these variants affects the genome over time.

1.7.4 Genomic technologies for studying complex traits

The rapid discovery of genetic variants from the human genome project and HapMap project led to major advances in genomic technologies, namely high-resolution microarrays and next-generation sequencing.[1,23–25] High-resolution microarray technologies from Affymetrix® (Affymetrix®, Santa Clara, CA) and Illumina® (Illumina®, San Diego, CA) became available during the late 1990's. These included a variety of SNP genotyping arrays, tiling arrays, and array comparative genomic hybridization (aCGH) technologies. SNP genotyping arrays allow for the analysis of allelic variation at hundreds of thousands of locations across the genome. Using hybridization technologies, SNP genotyping arrays are used to determine which allele an individual has at specific locations across the genome. Tiling arrays are used to intensively investigate whole segments of DNA rather than using SNPs that are distributed across the genome.[26] Tiling arrays contain probe sequences that overlap with each other or are spaced closely together so that an entire stretch of DNA can be queried by the probes. aCGH is used for the identification of CNVs between samples.[27,28] Comparative genomic hybridization relies on the use of a test sample and a reference sample labeled with two different dyes to identify variation in copy number. aCGH uses an array to determine the relative difference in fluorescence intensities between a reference and experimental sample. The differences in fluorescence intensities between the reference and sample are then used to determine overall copy number in comparison to the reference.

However, due to limitations in the amount of variation detectable with microarray technologies, researchers are turning to a more comprehensive analysis of the genome.[29] Next-generation sequencing technologies allow for the detailed analysis of the whole genome sequence. Next-generation sequencing provides the most complete picture of genetic variation between and within genomes, and is now the primary technology used in studies related to health and disease.[30–32] As whole genome sequencing becomes the new frontier for medical research, rapid advancements in next-generation sequencing technologies have made whole genome sequencing to a single base pair resolution financially feasible for most laboratories.[25,32] A variety of techniques have been developed, including synthesis by hybridization, nanopore sequencing, and sequencing by synthesis, that have allowed for rapid advancements in understanding the human genome.[33–35]

1.7.5 The Human 6.0 array

The most recent and comprehensive microarray technology that is currently available for humans is the Affymetrix® Genome-wide Human Array 6.0 (Human 6.0 array).[36] The

Human 6.0 array was released in 2008 and can detect variation at more than 906,000 SNP sites across the human genome. The Human 6.0 array also screens over 946,000 locations across the genome to assay copy number status. The Human 6.0 array was designed using only perfect match probes to maximize the number of variable regions that could be detected across the genome. SNPs selected, for inclusion on the Human 6.0 array, were based on the performance of SNPs used on the Human 5.0 array and Affymetrix® 500K mapping array. Additional SNPs identified from the HapMap project were used to supplement the SNPs selected from previous Affymetrix® microarrays in order to maximize the SNP coverage across the human genome. The Human 6.0 array is bi-allelic for each SNP, meaning that for every SNP location queried there is an A and a B allele represented by the SNP probe sequences. Copy number probes were also selected to obtain uniform coverage across the genome, which when combined with the SNP probes, are used to detect an increase or a decrease in copy number from the diploid state of the human genome. Copy number probes were not restricted by the location of SNPs across the genome. With the capability to both genotype a sample, as well as determine copy number, the Human 6.0 array has been implemented in a variety of different research studies focused on the identification of genetic variants within the human population.[37–39] In particular, the Human 6.0 array has been a popular choice for studying CNVs in relation to diseases.

1.7.6 Limitations of studying complex phenotypes in humans

Understanding complex traits and diseases has been made possible with the development of genomic technologies. Genome-wide association studies have made their mark on the scientific community, and have been responsible for the identification of many genotype-phenotype associations.[40] Unfortunately, short of identifying these associations, humans prove to be poor models to truly elucidate the mechanisms underlying the interactions between genes and their environment. To understand the basics for why specific genotypes result in certain characteristics, researchers must be able to limit the amount of variation between subjects. The inability to control a human's environment, therefore, makes it difficult to minimize the variation between samples. Similarly, traits affecting specific organs become difficult to study, as researchers are limited to non-invasive sampling of genetic material. Finally, understanding the inheritance and development of phenotypes, especially those associated with diseases, is challenging due to the long generation time of the human population.

An attempt to study complex phenotypes of individual tissues has been made in humans with the use of stem cell research.[41,42] Stem cell research allows for the study of individual organs in an *in vitro* environment. This allows for the analysis of genetic variation within a specific tissue. However, when studying complex phenotypes it is important to understand how these tissues interact with other organs in the body from a holistic view. Stem cell research does not recreate *in vivo* interactions that would occur within a whole organism.[43] Similarly, the ethical concerns regarding how these stem cells are obtained often limit the amount of research conducted using stem cells as a model. Therefore, researchers generally turn to model organisms to overcome the limitations of *in vivo* and *in vitro* experiments on human tissues.

1.8 Genetic variation in mouse populations

1.8.1 Mouse as a model for human complex phenotypes

Mice have been used as a model organism to study complex phenotypes for well over a century. The synteny of genome size and composition between mice and humans, make mice ideal for studying complex phenotypes and diseases that are observed in the human population. Additionally, mice naturally develop many of the same diseases as humans, including cancer, diabetes, glaucoma, addiction, and hearing loss because of their similarity in cellular physiology and development.[44] For diseases that do not naturally affect mice, mouse models have been created that mimic these diseases by manipulating the mouse genome. Transgenic technologies are used to insert, remove, or modify genes, to induce disease characteristics in mice to mimic specific diseases.[44] Coupled with the short generation time of a mouse, and the ability of researchers to control environmental conditions, mice prove to be a strong model for studying complex phenotypes. Therefore, many researchers examining complex traits choose to use mice as their model organism.

1.8.2 The origins of the laboratory mouse

Inbred mouse strains have been in development since the turn of the 20th century. With an interest of studying inheritance, the development of inbred mice was initially based on selecting for coat colour.[44] Mice were selected to be homozygous at loci for agouti (*a*), brown (*b*/*Tyrp1*) and dilute (*d/Myo5a*). To select for homozygosity at these loci, offspring were brother-sister mated for many generations, ultimately reducing the genetic variation. After nearly 20 generations of brother-sister mating, mice no longer contained variation between offspring, as all mice were homozygous at every location across the genome. This effectively creates mice that were genetically identical and contain the isogenic background that is characteristic of an inbred mouse strain.

By 1909, the first inbred mouse strain known as DBA was created by Dr. C.C. Little, the founder of The Jackson Laboratory.[44,45] The DBA inbred mouse was joined a short time later by several other inbred mouse strains, including C57, C3H, CBA, and A. Each of these strains contained their own unique genetic background with alleles primarily found in the *Mus musculus domesticus* subspecies. These mouse strains formed the basis from which many of the classical laboratory inbred mouse strains currently used in biological research were created.

The isogenic nature of inbred mouse strains makes them ideal for studies that require a large number of replicates that contain as little genetic variation as possible.[46,47] Because inbred mice contain no genetic variation between individuals within a strain, it is possible to repeat experiments while keeping the variation between samples at a minimum. By reducing the amount of variation between samples, conclusions may be made using a smaller sample set as compared to similar studies conducted between samples containing high genetic variation.[46] Inbred mouse strains exist for a variety of different phenotypic characteristics; this allows researchers to tailor their strain selection in order to answer specific biological questions.

Some of the first research conducted on complex phenotypes using these inbred mouse strains was aimed at understanding the basics of cancer research and immunology. One such study, conducted by Haldane in 1933, proposed the alloantigenic hypothesis of tumor rejection.[46] The alloantigenic hypothesis of tumor rejection functions on the premise that antigens are produced for an allele within a strain rather than as an immune response that is directly targeting the cells. Haldane showed that when tumors were transplanted within mice of the same strain, the tumor cells were not rejected. However, when transplanted to mice of a different strain, the tumor cells were rejected. The rejection of tumors between mouse strains indicates that there is a genetic basis to tumor rejection. The newfound understanding for the genetic basis of tumor rejection explained transplant rejection observed in human populations. This principle is still used today with respect to blood transfusions and organ donations, where blood and tissues are matched based on the patients genome.

Therefore, studies relating to complex phenotypes within laboratory mice have been applied to human populations.

1.8.3 The C57BL/6J mouse strain

One of the most commonly used laboratory strains is the C57BL/6J (B6) mouse. This strain was developed in 1921 and quickly became the strain of choice for many researchers. The B6 mouse strain was developed by Dr. C.C. Little using a parental line of C57 black mice, of which a defining characteristic is their black coat colour.[45] The B6 mouse strain, in particular, has been inbred for over 200 generations. The B6 mouse is also the mouse strain from which the mouse genome sequencing project was based.[48]

1.8.4 Congenic and consomic mice

Congenic and consomic mice were created to study the effects of single genes when different allelic variants were present on an otherwise isogenic background.[45] Congenic and consomic mice maintain the genetic background of their parental strains; however, they contain a single gene (congenic) or a single chromosome (consomic) from a mouse strain that is genetically distinct from their parents. The incorporation of a gene or chromosome from a different inbred mouse strain increases the genetic variation at a specific location in the mouse genome, while the remainder of the genome remains unchanged. Some of the most common congenic and consomic mice are on a B6 genetic background. However, these mice can be created for any combination of genetic backgrounds from the available mouse strains.

1.8.5 F1 mice

First filial (F1) mice are offspring from two different parental inbred mouse strains. These mice contain heterozygosity at all locations in the genome where the maternal and paternal inbred strains contain different alleles.[45] F1 mice display characteristics that result from a combination of two distinct genetic backgrounds. This increases the genetic variation within F1 mice in comparison to mice from an isogenic strain. The benefit of F1 mice is that studies can be reproduced. Therefore, by continually breeding two mice from two different isogenic backgrounds, all offspring contain the same combination of alleles. This allows for studies to be reproduced on a genetic background that is not isogenic; however, is reproducible.

1.8.6 Recombinant inbred strains

Recombinant inbred strains are derived from crossing two genetically distinct founder strains.[47] After breeding the F1 offspring from the original two founder populations, recombination occurs at meiosis as a result of homologous chromosomes crossing over. When loci are heterozygous, containing two different alleles, recombination can result in the creation of different haplotypes. The generation of new haplotypes in recombinant inbred strains results in new allelic combinations that are inherited together. After nearly 20 generations of brother-sister mating, mice are considered inbred and contain homozygosity at all locations in the mouse genome. The resulting recombinant inbred strain, contain regions of homozygosity derived from the maternal strain as well as regions of homozygosity derived from the paternal strain. By incorporating alleles from two different mouse strains into a single recombinant inbred mouse strain, the genetic diversity of recombinant inbred strains is higher than that of classical inbred strains. Therefore, recombinant inbred strains have all the benefits of a classical inbred mouse strain, with the added benefit of increased genetic diversity.

1.8.7 Wild-derived laboratory mice

Wild-derived laboratory mice, are inbred mouse strains that were derived from mice caught in the wild.[45] Wild-derived laboratory mouse strains are created by brother-sister mating the offspring from wild caught parents. Wild-derived laboratory mice, are isogenic mouse strains; however, the alleles within these strains are derived from different *Mus musculus* subspecies (*Mus musculus musculus, and Mus musculus castaneus*) in comparison to classical laboratory strains (*Mus musculus domesticus*).[49] The haplotypes associated with each of the founder strains reflect the genetic variation present in the wild caught parental mice. Alleles reflect those that are most common based on the subspecies and geographic locations of the wild caught parents.[45]

1.8.8 Mutant mice

Mutant mouse strains that display characteristics of diseased phenotypes are also available for many of the classical and wild-derived mouse strains. Mutant strains can arise from spontaneous mutations occurring during breeding, or from induced mutations using chemical or transgenic technologies.[45,50] Spontaneous mutations arising during breeding may result in phenotype changes. When this occurs, a new mouse strain is created when the mutation is non-lethal in a homozygous state. This creates a new mouse strain available for mutation analysis. Inducing mutations by modifying the genome with transgenic technologies or by chemically inducing mutations also creates mutant mouse strains. Mouse strains derived from induced mutations are also maintained at a homozygous state. Transgenic mutant mice can be created for any non-lethal combination of modifications; this allows for a broad spectrum of research on mutant phenotypes.

1.8.8.1 The *harlequin* (*hq*) mouse as a mutant mouse strain

The *harlequin* (*hq*) mouse is a mouse model for neurodegeneration.[51] The distinct phenotypic characteristics of the *hq* mouse include low body weight, a patchy coat, and severe ataxia. The *hq* mutation originally arose spontaneously on a CF1 outbred stock and was transferred to the current genetic background of B6CBACaA^{w-J}/A-Pdcd8^{Hq}/J[51]. The *hq* mutation results from a 1 kb proviral insertion into intron 1 of the *Apoptosis-inducing factor* (*Aif*) gene, resulting in an 80% downregulation of *Aif* expression.[51,52] AIF is involved in programmed cell death and it functions, in conjunction with complex I of the electron transport chain, as a NADH reductase.[53] The reduction of NADH results in an increase in the production of reactive oxygen species (ROS). However, recent studies have shown that *hq* mice do not show an increase in ROS.[54] AIF also functions as a mitochondrial hydrogen peroxide scavenger, which prevents damage caused by ROS.[53,55] The reduction of AIF in the *hq* mouse is associated with the degeneration of neurons in the cerebellum and retina in comparison to age-matched wild-type mice.[51,54,56] The *hq* mouse has also displayed a reduced level of AIF in the spleen which causes a reduction in neglect-induced death in T-cells.[57] Therefore, the *hq* phenotype not only affects neuronal tissues, but also displays a spleen-specific phenotype that has implications in immunity. The direct assessment of tissue-specific degradation in the *hq* mouse provides an excellent model for studying diseases of neurodegeneration affecting the human population. The implications of the cerebellar and retinal degeneration make the *hq* mouse an ideal model to study the effects of tissue-specific mutation accumulation in neurodegenerative diseases. By analyzing the accumulation of mutations in specific tissues, researchers can gain a better understanding of the mutational profiles including the frequency and distribution of mutations affecting specific tissues with neurodegenerative disorders.

1.8.10 Genetic variation within a mouse

Genetic variation not only exists between mouse strains, but also within mice as somatic mosaicism and germline mosaicism. Some of the most obvious forms of within mouse variation are associated with cancers. Cancers display mutator phenotypes that cause cells within a tumor to be genetically distinct from those outside of the tumor.[58] Although mutator phenotypes that are associated with cancer typically occur within a single tissue type, other research indicates that genetic variation exists between tissue types and within tissues as a result of point mutations.[30,59–61] There is also evidence that mutations accumulate in the germline of mice.[60,62,63]

1.8. 11 Traditional mouse models have minimal genetic variation

Currently, mouse models are used to study the mechanisms underlying a broad spectrum of complex phenotypes; including those associated with development and disease progression. Studies on inbred mouse strains have been used to identify the interactions between genes and how the interactions contribute to an observable phenotype. However, when genetic variation is incorporated into the genome, as seen within human populations, the interactions between genes are altered ultimately impacting the observable phenotype. [64] Gene-gene interactions are affected by genetic variation. Inbred mouse strains contain a minimal amount of genetic variation, as all mice within a strain are genetically identical. The human population is studied globally and contains high amounts of genetic variation because humans are typically not inbred. Increasing the variation within the genome, adds an additional level of information. Different combinations of genetic variants are now interacting to display a wide spectrum of phenotypes. Therefore, the murine model that is being used to study these interactions must be modified to mimic the genetic diversity of humans.

1.8.12 Next-generation recombinant inbred mice

The first step in developing a mouse model that mimics the genetic diversity of the human population was initiated with the development of the Collaborative Cross (CC) mouse, a “next-generation recombinant inbred strain”.[65] The CC mouse contains the genetic diversity of eight founder populations rather than the traditional two used in recombinant inbred strains. The eight founder populations include five classical laboratory strains (A/J, C57BL/6J, 129/SvImJ, NOD/LtJ, and NZO/H1LtJ) as well as three wild-derived laboratory mouse strains (CAST/EiJ, PWK/PhJ, and WSB/EiJ).[66] By breeding these eight founder strains together, mice were generated that contain alleles from each of the eight genetic backgrounds.[67] Currently, there are 160 CC mouse lines available, all of which contain different combinations of haplotypes derived from alleles from each of the founder strains. However, being an inbred mouse strain, each of these individual lines is homozygous at every locus in the genome. The CC mouse contains all the benefits of an inbred mouse strain with the addition of increased genetic variation. Although increasing the genetic variation across the genome is a step in the right direction, the next-generation recombinant inbred mouse still does not capture the heterozygosity observed within human populations.

1.8.13 The Diversity Outbred mouse

Developing a mouse model that mimics the human population requires the mouse to be heterozygous at many locations across the genome. The Diversity Outbred (DO) mouse was produced using a novel outbreeding strategy in order to maximize the allelic variation within a mouse strain.[68] This breeding strategy was derived from the idea of heterogeneous stock populations.[69,70] The unique feature of these mice, in comparison to classical inbred, recombinant, and next-generation recombinant inbred strains, is that they maintain a level of heterozygosity across the genome that replicates the levels of heterozygosity found within the human population. The 160 CC mouse lines were used to generate the DO mouse. Each line contains the genetic diversity from eight founder populations, therefore in the resulting DO mouse, there are various allelic representations from these eight founder strains. Because these mice need to maintain a high level of heterozygosity, the standard brother-sister mating technique is not applied. Rather, mice are bred using a random mating strategy, to generate mice that are genetically unique.[71] Therefore, each DO mouse that is bought and bred will have its own combination of founder alleles. Because each mouse is unique, the heterozygosity and recombination makes tracking genetic background using breeding records challenging, studies conducted with DO mice require genotyping at all loci across the genome to identify allelic origin.

1.8.14 The origins of somatic mosaicism with development

Mammalian development begins with a single cell, which during development replicates to form each of the tissue types within an organism. As this cell divides and organogenesis begins, each tissue will follow a unique developmental history with respect to proliferation, differentiation, cell type, apoptosis, and mutagen exposure. As an embryo develops it undergoes gastrulation, a process that divides the embryo into the three germ layers; the ectoderm, the mesoderm, and the endoderm. After gastrulation, organogenesis initiates and the primordial cells for each tissue begin to differentiate. Each of these primordial tissues begins differentiation at a distinct embryonic day at which point the primordial tissues are subjected to replication rates, apoptosis, and mutagen exposure specific to their tissue type. Therefore as tissues differentiate, they begin to accumulate genetic differences that reflect their developmental histories. The accumulation of these genetic differences results in somatic mosaicism, as tissues within the mouse are no longer genetically identical. Three candidate tissues for the study of somatic mosaicism, that represent each of the three germ layers, are the spleen, the cerebellum, and the liver. Each of these tissues has shown different rates of accumulation of spontaneous mutations over the lifespan of a mouse that are unique to the developmental and functional histories of each tissue.[59–61,72–78]

The spleen is a mesodermal tissue that begins development at gestational day 12.5.[79] It is divided into two distinct compartments, the red pulp and the white pulp. The cell types found within the spleen include lymphocytes, hematopoietic cells, granulocytes, and circular mononuclear cells in the red pulp, and lymphocytes, macrophages, dendritic cells, and plasma cells in the white pulp. In the mouse, the spleen accounts for about 0.2% of the weight of the mouse and acts as a blood filter. The spleen also provides hematopoietic functions beginning at gestational day 17, as it is involved in the production of blood cells. Therefore, the spleen plays an intricate role in the immune system of the mouse. The replication rate of cells within the adult spleen ranges between 1 and 21 days depending on cell type.[80] Within the spleen there is evidence of somatic mosaicism associated with immunology and the hematopoietic function of the tissue.[81,82]

The cerebellum is an ectodermal tissue that develops at embryonic day 10.5. The cerebellum is 90% neurons, including granular neurons, Purkinjie cells, and oligodendrytes.[83] Other non-neuronal cell types within the cerebellum include parvalbumin-positive fast-spiking basket cells, somatostatin-positive regular-spiking bipolar and multipolar cells, and cholecystokinin-positive irregular-spiking bipolar and multipolar cells.[84] The cerebellum is a post-mitotic tissue meaning that cells are no longer dividing in adult mice, resulting in little cell turnover in this tissue. It contributes to roughly 2% of the body weight of a mouse and the primary function is the regulation of motor control and balance. Previous research in humans has identified chromosomal aneuploidy in the developing fetal brains.[85] In mice, mobile elements in the brain have been identified during development and in granular neurons of adult brain tissues that contribute to somatic mosaicism within the cerebellum.[86,87] The cerebellum is also prone to genomic modifications as a result of circular DNA occurring early in development around embryonic day 16-17.[88] The presence of circular DNA molecules facilitates somatic recombination within cells of the developing embryonic brain leading to genomic alterations.

The liver develops from endodermal tissues at embryonic day 9.5 and contributes to about 7% of the overall body weight. It is composed of nearly 70% hepatocytes (parenchymal cells), which replicate every 480-620 days.[80,89] Hepatocytes are involved in the synthesis and storage of proteins. The liver also contains non-parenchymal cells including Stellate and Kupffer cells.[90] The liver is also responsible for the removal of cellular waste. Previous reports of somatic mosaicism in the liver has been associated with highly unstable minisatellites early in development.[91] There is also evidence for somatic mutations occurring during the processing of metabolites,[92] and recent literature has identified CNVs associated with metabolic stress.[93,94]

1.8.15 Tissue-specific mutation frequency

Studies of tissue-specific mutation frequency using transgenic mice have reported the different rates at which tissues accumulate mutations throughout development.[60,61,73,78,95] Beginning as a zygote containing no mutations, mutations have been reported to accumulate as early as the fetal stage.[60,61,73] By post-natal day 10, mutation frequency has been shown to be tissue-specific.[60] As mice mature into late adulthood, tissue-specific mutation frequencies become more distinct.[60,73,78] Neuronal and germ tissues maintain a constant mutation frequency after 3 months of age, whereas mutation frequency in the liver, spleen, and adipose tissues increases as the mouse ages.[60] The extent to which spontaneous mutations accumulate in a tissue depends upon the development of the tissue and how they are affected by aging.

1.8.16 Genome-wide mutation detection

Genome-wide mutation detection is the gold standard for mutation studies. In humans, high-density microarray technologies such as the human 6.0 array (Affymetrix®), array Comparative Genomic Hybridization (aCGH), and the cytoscan HD array (Affymetrix®) incorporate probes that represent regions across the genome to detect both small and large genomic modifications.[28,96] Next-generation sequencing technologies also provide affordable genome-wide screening of the entire genome to identify genomic modifications.[32] Unfortunately in mice, the cost of sequencing the genome is far from affordable for most laboratories, transgenic mice screen limited genomic areas, and microarray technologies lack genome-wide coverage compared to human microarray standards.[97] Therefore, a high-resolution, genome-wide mutation detection technology is in high demand for the mouse.

1.8.17 Analysis of complex traits by single gene mutation detection

Traditionally, complex traits in mice are studied using mutation detection approaches, as many of these traits are related to aging and disease progression. Since mutations are known to cause aging and disease progression, research on mouse models focused on the identification and characterization of mutations.[98–100] Conventional mutation detection systems in the mouse rely on the use of single gene analyses to study the frequency and nature of mutation accumulation. Single gene mutation studies allow for *in vitro* examination of mutations and can be conducted using either endogenous genes, like *HPRT*, or transgenes such as *cII*, *lacI*, and *lacZ.*[101] Mutations are identified by packaging the isolated genes into a vector, infecting these vectors into *Escherichia coli* and screening for mutant colonies or plaques.

Transgenic mice, namely Muta^{TM} Mouse and the BigBlue$^{®}$ mouse have been developed to study mutations using transgenes. These mice contain copies of the 48 kb lambda genome which contains the *cII*, *lacI*, and *lacZ* genes.[102,103] Muta^{TM} Mouse was developed on a BALB/C X DBA/2J genetic background and contains 40 copies of the lambda phage genome on chromosome 3. Muta^{TM} Mouse was traditionally used to study the *lacZ* transgene; however more recently has been used for the *cII* assay as well.[102,104,105] The BigBlue$^{®}$ mouse is a transgenic mouse model that contains 40 copies of the lambda genome on chromosome 4. This mouse was developed on a B6 genetic background and is traditionally used to study the *lacI* and *cII* transgenes.[103,105] Although these transgenic mouse models prove to be effective at identifying the frequency and type of mutation accumulation throughout development, these mutation assays lack the ability to detect mutation accumulation outside the region of these exogenous genes.

1.8.18 Genomic technologies for the mouse

Unfortunately, genomic technologies for the mouse are limited in comparison to humans, in their ability to detect genome-wide allelic differences. Genomic technologies for the mouse include microarrays and next-generation sequencing. However, the advancement in genomic technologies specific to the mouse lags behind those that are available for humans. The lack of demand for mouse specific genomic technologies resulted in poor technological advancement and a much higher price tag for genome-wide studies in comparison to humans. As a result, microarrays, the predominant genomic technology for the mouse, were limited in the number of locations that could be studied across the genome. Until as recently as 2006, microarray technologies could detect only 14,000 SNPs across the mouse genome.[106] Next-generation sequencing technologies, although available for the mouse, are cost prohibitive for most laboratories costing approximately $35,000 to sequence a single mouse genome. Traditional methods used to identify allelic variation, including polymerize chain reaction (PCR), although effective, are inefficient as a genome-wide strategy. Currently, mouse models for studying complex diseases, namely the DO mouse, would require PCR amplification for hundreds of thousands of locations across the genome in order to determine the genetic background of each allele. Therefore, a high-resolution technology for the mouse that is efficient and cost effective was developed.

1.8.19 The Mouse Diversity Genotyping Array

In 2009, the highest resolution, genome-wide microarray was developed by The Jackson Laboratory in collaboration with Affymetrix®.[107] Modeled after the Human 6.0 array, the Mouse Diversity Genotyping Array (MDGA) was designed to capture as much genetic variation as possible within the mouse population. The MDGA incorporates probes for 623,124 single nucleotide polymorphisms (SNPs) identified from inbred and wild-derived mouse strains (Table 1-1). This is also the first mouse array capable of detecting large structural variants with an additional 916,296 invariant genomic probes (IGPs), conserved across mouse subspecies. All probes are distributed across each chromosome of the mouse genome to give the highest-resolution microarray technology for the mouse to date.

Probes for SNPs and IGPs have two distinct designs. SNP probes were designed in sets of eight (probe set), where all eight of these probes are required to detect the genotype at a single SNP location and must meet the strict criteria outlined in Table 1-2. These eight probe sequences were designed to detect two possible alleles on the array, four detecting Allele A and four detecting Allele B (Figure 1-1). Two probes for each allele detect a sense (forward/+) target sequence and two probes for each allele detect an antisense (reverse/-)

target sequence. The four sequences on the forward strand only differ in the allele that is detected, meaning that the two sequences detecting Allele A are identical and only differ from the two sequences detecting Allele B by a single nucleotide. Likewise, the four sequences representing the reverse strand are identical with the exception of the SNP detected. Although probes for the forward and reverse strand of DNA detect the same SNP, they may be offset up to 10 bps. Probes representing Allele A and Allele B of the forward strand are printed on the array in pairs at two separate locations on the array to account for the duplication of the probe sequences. Probes for the reverse strand are printed in the same manner.

Using strict criteria for probe design is essential for accurate genotyping. A combination of fluorescence intensities from eight probe sequences is used to determine a genotype call.[108,109] By incorporating two identical probe sequences for each strand and for each allele on the microarray, redundancy is created and SNP genotype calling becomes more accurate. However, when inconsistencies occur in the design of these probe sequences, they manifest as technical errors in genotyping calls. Because microarrays are designed to detect genotype calls over hundreds of thousands of locations across the genome, probes detecting these locations are designed using a strict set of criteria. Probes are designed so that each probe sequence on the microarray works at an optimal hybridization temperature, ideally for allele specificity. Therefore, for probes that do not meet basic design criteria, such as probe length and sequence context, the genotyping calls determined using these probe sets are unreliable.

Table 1-1: The total number of SNPs in the original SNP list annotation by strategy of SNP identification

Mouse Chromosome	Probe identification strategies[a]							
	Classical[b]	NIEHS[c]	B6[d]	Wild[e]	MSM[f]	Mt[g]	Y[h]	Total
1	12246	35799	278	46	2882	0	0	51251
2	11313	27968	112	30	2825	0	0	42248
3	9845	26298	203	22	2498	0	0	38866
4	9369	24908	192	17	2262	0	0	36748
5	9039	25896	169	14	2393	0	0	37511
6	9097	25707	156	19	2158	0	0	37137
7	8249	24573	199	70	1891	0	0	34982
8	7982	22226	152	21	1726	0	0	32107
9	7755	21202	164	87	1948	0	0	31156
10	7970	19084	144	46	2112	0	0	29356
11	7660	18709	144	41	1966	0	0	28520
12	7049	21616	155	48	1639	0	0	30507
13	7125	20597	97	36	1797	0	0	29652
14	7288	17456	117	56	1525	0	0	26442
15	6448	18119	102	55	1555	0	0	26279
16	6093	15335	121	28	1593	0	0	23170
17	5710	18224	122	50	1368	0	0	25474
18	5544	16437	55	13	1550	0	0	23599
19	3633	11513	86	40	907	0	0	16179
X	8238	12565	5	60	1002	0	0	21870
Y	0	0	0	0	0	0	51	51
Mitochondrial	0	0	0	0	0	19	0	19
Total	157653	424232	2773	799	37597	19	51	623124

[a] how the SNPs used on the MDGA were identified. Grouped by the type of laboratory strain, chromosome, or project that identified the SNPs.

[b] SNPs identified from classical laboratory strains

[c] SNPs identified from studies conducted by the National Institute of Environmental Health Sciences (NIEHS)

[d] SNPs identified in the B6 mouse that are absent from the NIEHS

[e] SNPs that are not shared with A/J, DBA/2J, 129S1, and MSM/Ms strains.

[f] SNPs identified in *Mus macedonicus, Mus spretus, Mus spicilegus,* and *Mus musculus*

[g] SNPs identified that are on the mitochondrial chromosome

[h] SNPs identified that are on the Y chromosome

Table 1-2: Criteria outlined by Yang *et al*, 2009 to define a "chippable probe" on the Mouse Diversity Genotyping Array

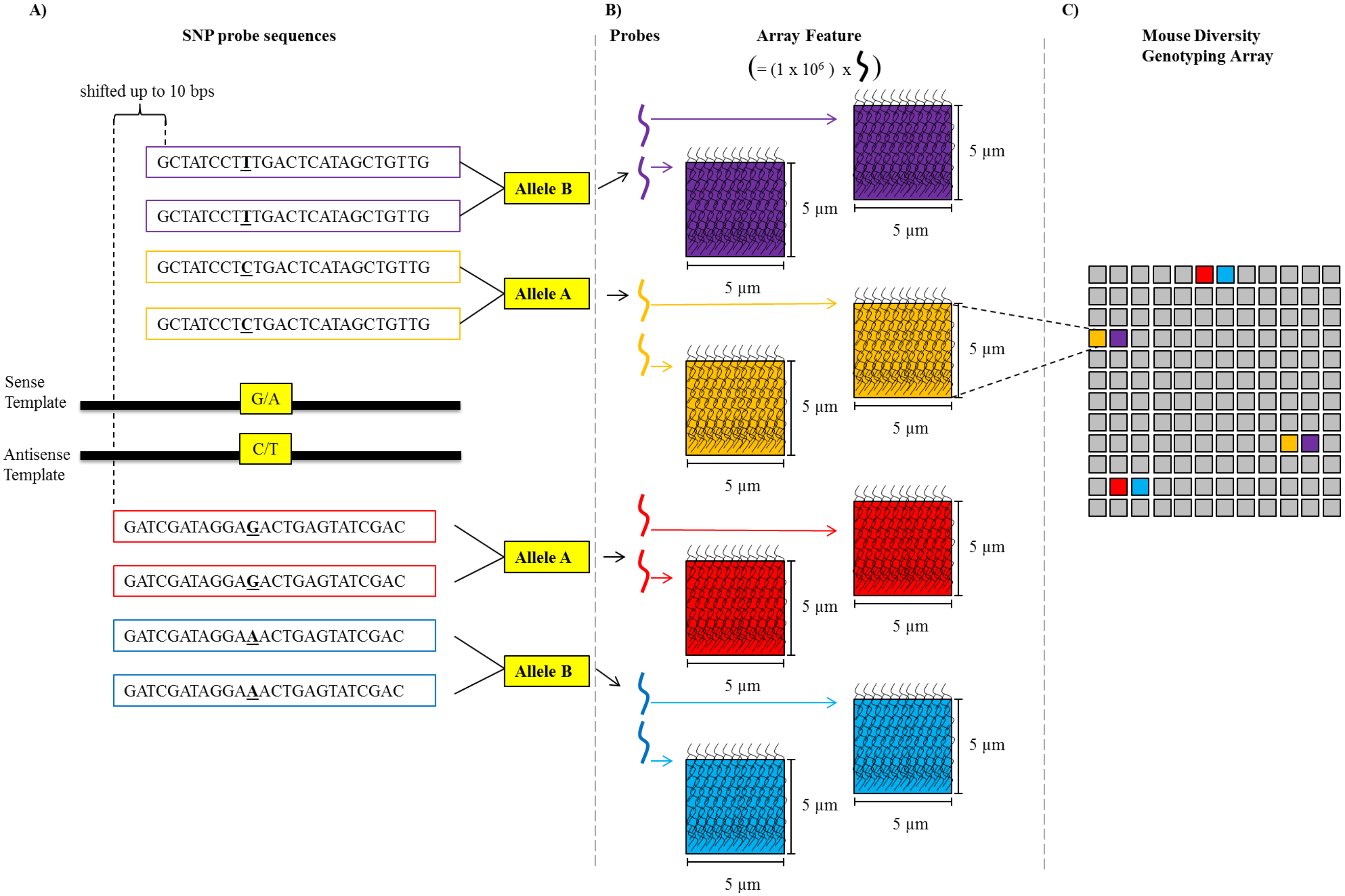
Criteria[a] probes must meet to be chippable
1. SNP must be located on a *Nsp*I or *Sty*I fragment[b] between 50 bps and 1 kb
2. SNPs should be at least 10 bps away from a *Nsp*I and *Sty*I cut site
3. No other SNP ± 12 bps of the target SNP
4. 33 mer centered on the SNP must BLAT[c] as unique

[a] All probe selection criteria are based on the C57BL/6J mouse reference genome sequence build 36 (NCBI, mm8:Ensembl) and build 37 (NCBI, mm9:Ensembl)

[b] *Nsp*I and *Sty*I are the restriction enzymes used to digest the genomic DNA

[c] sequence alignment tool in Ensembl

Figure 1-1: Design of Single Nucleotide Polymorphism (SNP) probes on the Mouse Diversity Genotyping Array (MDGA). A) Each SNP selected for the array is detected using a combination of 8 probes that are represented on the array. The 8 probes consist of a combination of 4 unique probe sequences (outlined in purple, yellow, red, and blue), each represented twice on the array. Two sequences (purple and yellow) detect sense DNA and two sequences (red and blue) detect antisense DNA. The two sequences on the sense strand differ by a single base pair (SNP in bold), for Allele A (yellow) or Allele B (purple). Two sequences on the antisense strand differ by a single base pair (SNP in bold), for Allele A (red) or Allele B (blue). Probes detecting sense and antisense sequences can be shifted up to 10 bps from each other. **B)** Each probe sequence (purple, yellow, red, and blue) is represented on two separate features of the array that are printed in pairs. **C)** Probes detecting sense alleles are paired together and probes detecting antisense alleles are paired together. With each separate sequence represented twice on the array, each pairing occurs at two separate locations on the MDGA.

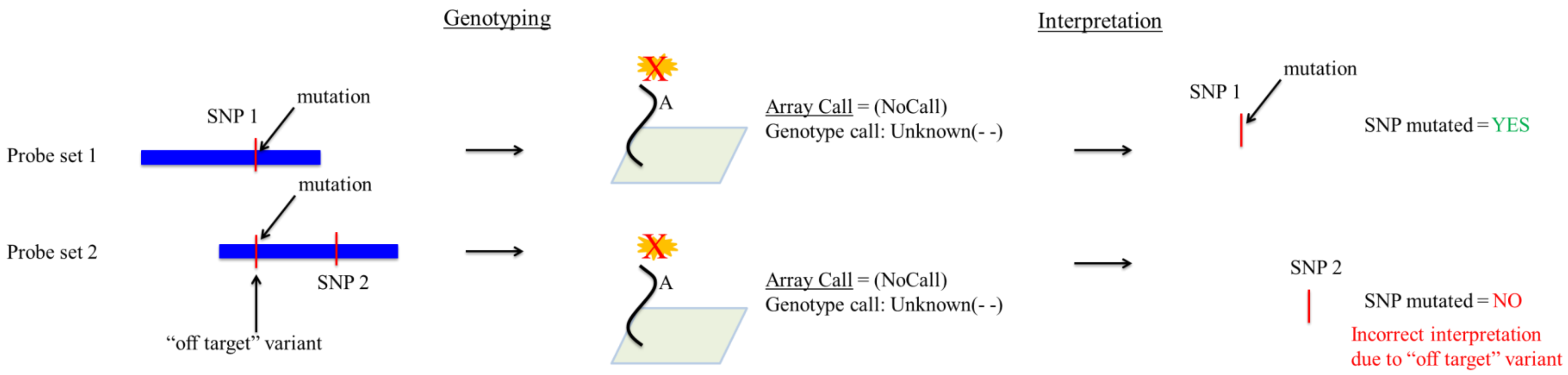

A)
SNP probe sequences
shifted up to 10 bps
GCTATCCTTTGACTCATAGCTGTTG
GCTATCCTTTGACTCATAGCTGTTG
Allele B
GCTATCCTCTGACTCATAGCTGTTG
GCTATCCTCTGACTCATAGCTGTTG
Allele A
Sense Template
G/A
Antisense Template
C/T
GATCGATAGGAGACTGAGTATCGAC
GATCGATAGGAGACTGAGTATCGAC
Allele A
GATCGATAGGAAACTGAGTATCGAC
GATCGATAGGAAACTGAGTATCGAC
Allele B
B)
Probes
Array Feature
(= (1 x 10⁶) x)
5 µm
C)
Mouse Diversity Genotyping Array

Technical error resulting from flaws in probe design can affect genotyping accuracy because the probes no longer perform under the “optimal conditions” to which they were designed. Probes that are not 25 nucleotides long do not match the optimal hybridization temperature calculated for the microarray.[109] Shortened probe lengths are also more susceptible to mismatch alignment causing a reduction in the genotyping accuracy of the microarray. Probes that align to more than one location in the mouse genome will determine the genotype of a SNP based on hybridization to multiple locations across the genome. Therefore, genotype calls, are no longer indicative of the SNP for which the probe set was designed to detect. Probes containing recognition sequences for the digestion enzymes used to prepare the samples for hybridization cannot detect genotypes because their target DNA will have been digested. SNPs detected by multiple probe sets on the array potentially double the amount of genetic variation detected within a sample. This is because when a SNP is detected by its assigned probe set, as well another probe set that is associated with a different SNP, it causes an “off target” variant in this second probe set (Figure 1-2). Because the SNP is detected twice on the array any differences detected will be double counted. This results in an overestimate of the number of NoCalls (no genotype call determined) or the genetic distance values, when the SNP contains an allele that contributes to a genetic difference between samples.

Design of the IGPs is slightly simpler than the SNPs, with only two probes querying each invariant genomic site represented on the array. Invariant genomic probes were designed to be within the exons of all genes in the mouse genome. Probe sequences selected in each of the exons for representation on the array, had to be invariant within the mouse genome (containing no SNPs) and occur only once in order to be used for copy number detection. Each exon that was selected contains a total of 6 probes, two detecting a proximal site within the exon, two detecting a medial site within the exon, and two detecting a distal site within the exon. Ideally the proximal, medial, and distal probe sequences do not overlap, however, constraints based on sequence context of the mouse genome may result in overlapping probe sequences.

The MDGA was designed with the intent of detecting genetic variation between mice of different genetic backgrounds. To date, the MDGA has been used primarily by The Jackson Laboratory to detect variation and map quantitative traits within outbred mouse strains, the DO mice, and the CC lines.[67,68,110,111] Researchers have also been using the MDGA to detect CNVs across the mouse genome.[112,113] The MDGA has been used to study quantitative traits and determine genetic variation across the genome between different mouse strains. However, the MDGA has the potential to be the first high-resolution technology for studying genetic variants underlying complex phenotypes and the potential to be the first high-resolution mutation detection assay for the mouse.

Figure 1-2: Overestimate of the number of genotype differences when SNPs are detected by multiple probe sets. The genotype call for any given probe set will be affected by “off target” variants, variants that occur in the probe sequence other than the SNP for which that probe was designed. SNPs, such as the mutated SNP in probe set 1, will affect the genotyping calls of both probe set 1 and probe set 2. When interpreting genotype calls, the first probe set was correctly identified as a variant, as the SNP that the probe set was designed contains a mutation. However, probe set 2 was designed for a SNP that did not contain a mutation and should therefore have had a genotype different from a NoCall. The off target variant in the second probe set, however, caused a mutation in the target sequence for the probe which caused the incorrect genotyping of the second SNP.


Genotyping
Interpretation
Probe set 1
SNP 1
mutation
A
Array Call = (NoCall)
Genotype call: Unknown(- -)
SNP 1
mutation
SNP mutated = YES
Probe set 2
mutation
SNP 2
"off target" variant
A
Array Call = (NoCall)
Genotype call: Unknown(- -)
SNP 2
SNP mutated = NO
Incorrect interpretation
due to "off target" variant

1.8.20 Genetic distance as a measure of genetic diversity detected using genotyping microarray technologies

The most advanced microarray technologies have been designed to capture genetic variation at a high-resolution across the genome. The ability to capture genetic variation at the genomic level has enabled researchers to analyze phylogenetic and genealogical relationships within a variety of species. To date, genotyping microarrays have been used to recreate representations of phylogenies for many species including humans, horses, dogs, mice, plants, and viruses.[112,114–119] Phylogenetic representations of the relationships are created using neighbour-joining algorithms with data taken from either similarity matrices (total number of genotype calls that are the same between two samples divided by the total number of genotypes compared with the differences between the two samples subtracted) or distance matrices (the total number of different genotype calls between two samples divided by the total number of genotypes compared between the samples). These matrices are then used to create phylogenetic trees describing how distant or closely related samples are within a species or subspecies.

1.8.21 Mutation detection with the Mouse Diversity Genotyping Array

The MDGA was designed as the mouse equivalent of the Human 6.0 array. It is an affordable high-resolution technology for the mouse that bridges the gap in genomic technologies between mice and humans. Because the design on the MDGA was based off of the Human 6.0 array, it should have the capability to be used as a mutation detection system. With the ability to screen more than 1.5 million locations across the mouse genome, the MDGA has the potential to be the first high-resolution mutation detection assay for the mouse. Probes associated with each SNP can detect both “on target” and “off target” mutations as either differences at the SNP location itself or differences within nucleotides surrounding the SNP. It also provides the first genotyping microarray for the mouse that combines SNP probes and copy number probes for the identification of CNVs. Ultimately, the development of the MDGA fills the gap in technological advances between humans and mouse. This allows for mouse models to be used to study complex phenotypes such as aging and disease progression with a comparable technology to that available for humans.

To determine if the MDGA can be used as a next-generation mutation detection system, we must first determine the ability of the microarray to detect genetic variation between and within laboratory mice. By determining how effective the microarray is at detecting genetic diversity, we can conclude whether or not the microarray is sensitive and specific enough to be used for detecting somatic mutations. Comparisons of mutation frequency and distribution of mutations between tissues of the same mouse will allow me to compare genome-wide mutation accumulation in tissues that have different developmental origins and timelines. Analysis of mutation accumulation in tissues with such different developmental histories, will allow me to establish a genome-wide mutation signature for each tissue during development and aging.

1.9 Central Hypothesis and Specific Aims

Given that SNP-based genotyping microarrays have high-resolution for assessing genetic diversity, the Mouse Diversity Genotyping Array can be used to investigate genetic distance not only between different mice but also between tissues of the same mouse and within a single tissue.

Specific Aim 1: To assess if all probes on the MDGA have an optimal design for hybridization and genotyping.

Specific Aim 2: To measure genetic distance using the MDGA to distinguish between distantly and closely related mice.

Specific Aim 3: To measure genetic distance using the MDGA to distinguish between mice ranging in genetic background from 100% B6 to 100% CBA/CaJ genetic backgrounds.

Specific Aim 4: To assay tissue-specific genotypes to distinguish between the spleen, cerebellum, and liver within a mouse.

Specific Aim 5: To assay intra-tissue genotype variation using replicates for the spleen and cerebellum of a B6 mouse.

Chapter 2 : Materials and Methods

2.1 Generating a SNP probe list

2.1.1 Understanding probe design

A SNP is a site within the genome that contains variation at a single nucleotide within at least one percent of the population. Variants at these single nucleotide positions are known as alleles. On the MDGA, each SNP is detected by a set of eight probe sequences, also known as a probe set. These probe sets contain probe sequences that can detect the forward strand of DNA and probe sequences that can detect the reverse strand of DNA. These sequences also have the capability to bind to DNA containing one of two alleles, an A allele or a B allele. To create a list of probe sets that are most effective at genotyping, each probe within a probe set was filtered using a strict set of design criteria. The identification of probes that fail to meet these design criteria resulted in the removal of the entire probe set from analysis. Removal of a SNP, results in the removal of the associated probe set from genotyping analysis.

2.1.2 Making a consensus list

Two separate genotyping lists were released with “poorly preforming SNPs” removed from the original 623,124 SNPs that are printed on the array. The first list released by The Jackson Laboratory contained 549,683 SNPs after removal of all “poorly performing SNPs” and the second list released by Genotyping Console (Affymetrix®, Santa Clara, CA) contained 584,726 SNPs after removal of all “poorly performing SNPs”. Both of the updated lists were compared to determine all SNPs present in both The Jackson Laboratory list and the Genotyping Console list. The overlap between the two lists was used to generate a consensus list of SNPs. Probe filtering was performed on this consensus list, where SNPs were removed based on their probe design.

2.1.3 SNP filtering by probe design

SNPs in the consensus list were subjected to strict probe design analysis. In addition to the original guidelines for creating "chippable probes" outlined by Yang *et al*, 2009, each SNP probe was tested against seven parameters (Table 2-1). SNPs with one or more of their 8 corresponding probes failing to meet these criteria were removed from the final SNP list used for genotyping (Figure 2-1).

Probe sequences obtained from the Didion *et al*, 2012 annotation file were checked against the CDF library file (Affymetrix®), which provides the coordinates for each of the probes (623,124 SNPs * 8 probe sequences each) on the MDGA. Annotations from the Didion *et al*, 2012 file were determined to be correct by comparison with locations checked by Fadista *et al*, 2012 and were used as the final annotations for the chromosome and chromosome position in the remainder of the analysis.[117,120]

1) Probe lengths were checked for both the forward and reverse probe sequences for the A and the B alleles. Sequences were taken from the CDF file available from The Jackson Laboratory. Probe lengths were calculated in Microsoft Excel™ (Microsoft, Redmond, Washington) by determining the number of characters in the string that denoted the probe sequence. For inclusion in the genotyping list, all eight probes representing a single SNP were 25 nucleotides long. If one or more of the eight probe sequences did not meet this criterion, the SNP was removed from genotyping analysis.

2) The alignment scores for forward and reverse probes were checked for each SNP. Alignment scores were taken from the most recent publication of MDGA

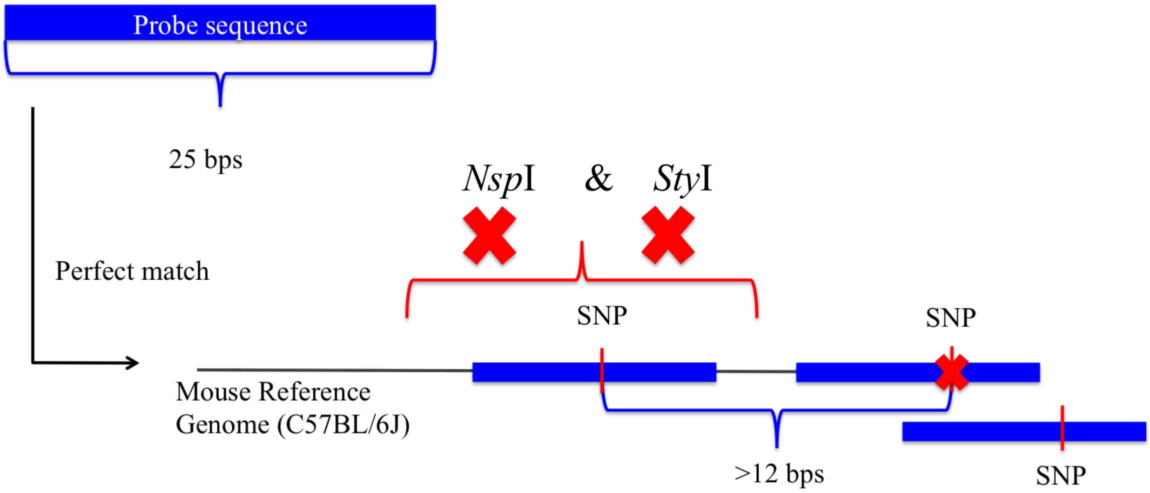
Table 2-1: Criteria SNP probes must meet for inclusion in the final genotyping list

Inclusion criteria for SNP probes in genotyping list[a]
1. Probe length must be 25 nucleotides
2. Alignment score must be 2, “Perfect match” [b]
3. SNPs must be present in the mouse SNP database (Mouse Phenome Database)
4. Probes must not be cut with *Nsp*I and *Sty*I
5. 10 bps upstream and downstream of the probe must not be cut with *Nsp*I and *Sty*I
6. SNPs must be at least 12 bps apart
7. SNPs must not be detected with more than one probe set

[a] All probe selection criteria are based on the sequence of the C57BL/6J mouse reference genome build 37 (NCBI, mm9:Ensembl)

[b] sequences that align to a single location in the mouse reference genome build 37

Figure 2-1: **Probe removal criteria when filtering the original SNP list.** All probe sequences on the array must be 25 bps in length, and align as a perfect match to the mouse reference genome. They cannot be cut with both *Nsp*I and *Sty*I restriction enzymes within the sequence of the probe or the 10 bps flanking the probe. Consecutive SNPs must be at least 12 bps away from each other and any SNP that was detected by multiple probe sets on the MDGA was removed.


Probe sequence
25 bps
Perfect match
NspI & StyI
SNP
SNP
Mouse Reference
Genome (C57BL/6J)
>12 bps
SNP

Annotations.[117] Probes were given a score of -3, -2, -1, 1, or 2 depending on their alignment to GRCm37 (build 37) of the mouse reference genome.[117] These alignment scores were determined using Burrows-Wheeler Aligner (BWA).[121] A score of -3 indicates that the probe set (all probes pertaining to a particular SNP) was not uniquely aligned to the reference genome. A score of -2 indicates that the probe sequence did not align to the reference genome. A score of -1 indicates that the probe sequence was not unique to the reference genome. A score of 1 indicates that the probe aligned to another location in the genome with a single mismatch. Finally, a score of 2 indicates that the probe sequence aligned to a single location in the genome with no off target alignments. All probes for the forward and reverse orientation of a SNP must be "2" for inclusion in the final genotyping list.

3) Genotyping output using the MOUSEDIVm520650 library file (Affymetrix®) identified SNP probes that did not have a chromosome location associated with them. SNP probes were then checked for inclusion in the Mouse Phenome Database (MPD) using the chromosome and chromosome base pair location. If the location in the genome was not available, the rsNumber was used to search for the SNP. SNPs that were not included in the MPD were removed from analysis.

4) Forward and reverse sequences for each SNP were checked for the full *Nsp*I and *Sty*I restriction enzyme recognition sites. *Nsp*I cuts at 5' RCATGY 3' (5' ACATGT 3' and 5' GCATGC 3') and *Sty*I cuts at 5' CCWWGG 3' (5' CCATGG 3' and 5' CCTAGG 3'). Probes were searched for all possible cut sites of *Nsp*I and *Sty*I in the CDF file and SNPs were removed if any of the eight sequences were cut by both restriction enzymes.

5) Sequences 10 bps upstream and downstream of the forward and reverse probe sequences for each SNP were checked for the full *Nsp*I and *Sty*I restriction enzyme recognition sites. Additional sequences were obtained using an in-house shell script which

extracted 50 bps before and after the SNP using its chromosome and base pair location from build 37 of the mouse reference genome (NCBI). Probe start and end positions taken from Didion *et al*, 2012 were used to determine the 10 bps before the start of the probe and the 10 bps after the end of the probe. SNPs where any probe contained a cut site in these additional 20 bps of sequence for both *Nsp*I and *Sty*I were removed from genotyping.

6) The nucleotide spacing between consecutive SNPs, or inter-SNP distance, represented on the array was determined using the chromosome positions provided in the reannotation by Didion *et al*, 2012. Probes that had an inter-SNP distance of less than 12 bps were removed from the SNP list for genotyping.

7) Forward and reverse probe sequences were checked to determine if there was overlap with more than one SNP represented on the array. SNPs were ordered by chromosome and chromosome position (bps). Consecutive SNP probes were checked to determine if the SNP directly before or directly after could be detected within the range of the probe for the SNP of interest. If the SNP fell between the probe start and end positions of the previous or subsequent probe sequences, the SNP detected by multiple probe sets was removed.

2.2 Annotating the filtered SNP list

Annotations from the MOUSEDIVm520650 library file were cross-referenced with the most recent annotation of SNP probes for the MDGA.[117] Chromosome location and base pair position were compared between these two files to identify any discrepancies. Differences to chromosome or chromosome position (bps) had the reference sequence ID numbers searched in dbSNP (NCBI) against build 37 of the mouse reference genome. Chromosome and chromosome position (bps) were corrected according to the NCBI database, build 37 of the mouse reference genome. Genomic locations were checked against the Fadista *et al*, 2012 SNP locations, where SNPs were identified to contain incorrect annotations in the original annotation file.[120] All SNPs were then plotted on a karyotype of the mouse genome to visualize the distribution of SNPs across the genome using Circos (Canada's Michael Smith Genomic Sciences Centre, Vancouver, BC, Canada).[122] Chromosome and chromosome position (bps) were used to compare the distribution of SNPs across the genome and the interSNP distances between the original SNP list and the filtered SNP list.

Chromosome location was also checked for each SNP using the Mouse Phenome Database (MPD), a publically available database maintained by The Jackson Laboratory®.[123] A SNP/variation query was performed for each SNP within the CGD-MGA1 dataset (MDGA dataset) which includes over 546,000 SNPs for 151 mouse strains.[124] At each SNP, the alleles for two commonly used laboratory strains, the C57BL/6J (B6) mouse strain and the CBA/CaJ mouse strain were recorded. The SNP for the *Mus musculus molossinus* (MSM) strain from Japan was also determined to be used as a possible control. MSM probes can be used as controls to determine the accuracy of genotyping because mice containing a B6 and a CBA/CaJ genetic background should not genotype with the MSM allele. The A allele and the B allele for each SNP on the array was identified as either a B6 allele, a CBA/CaJ allele, or a MSM allele using MPD. If the allele was not B6, CBA/CaJ, or MSM, the annotation was left blank. Each SNP was then assigned a "B6:CBA/CaJ annotation" based on their capability to detect the genetic backgrounds of the B6 and CBA/CaJ mouse strains. SNPs where only the B6 allele was detected were annotated as "B6", where only the CBA/CaJ allele was detected as "CBA", where both the B6 and CBA/CaJ alleles were detected as "Heterozygous", and where neither the B6 nor CBA/CaJ alleles were detected as "negative". SNPs where neither the B6 nor CBA/CaJ alleles were detected were annotated as negative because mice used in the somatic mosaicism study only contained the genetic backgrounds for a B6 and CBA mouse and would therefore be expected to have NoCalls at these negative control probes.

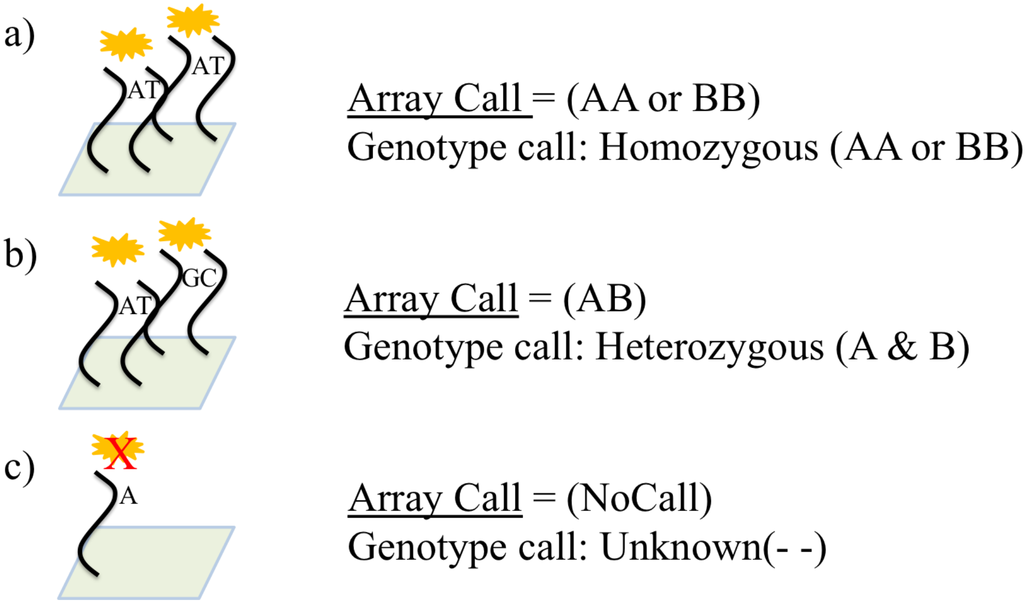
Chromosome number and chromosome position were used to identify whether a SNP was located within the region of a gene. Two databases were downloaded from UCSC, the first containing all known genes (reference genes) and the second containing putative genes (known genes) in build mm9 (Ensembl) of the mouse reference genome. All genes, known and putative, that span the genomic location of the SNP were recorded.

Probe sequences for the forward and reverse probes were taken from the CDF file and included in the final annotation, along with the chromosome, chromosome position (bps), SNP, alternate SNP, start position of the forward probe sequence, end position of the forward probe sequence, start position of the reverse probe sequence, end position of the reverse probe sequence, and the target sequences for the forward and reverse probes taken from the Didion *et al*, 2012 annotation. All additional information from the Didion *et al*, 2012 annotation file was included in the corrected annotations. This additional information was not checked given that the focus of SNP list filtering and reannotation was based on probe design.

2.3 Genotyping of the Mouse Diversity Genotyping Array CEL files

Genotyping was performed using Affymetrix® Genotyping Console (4.1.2) (Affymetrix®, Santa Clara, CA), a publically available program which can be downloaded from Affymetrix® (http://www.affymetrix.com/estore/browse/level_seven_software_products_only.jsp?productId=131535#1_1). Genotyping Console runs using the Mouse Diversity Array, AGCC library files (CD_MOUSEDIVm520650_rev2) which can also be downloaded from the Affymetrix® website or directly through the Genotyping Console program. A project file was created defining the location of the library and the type of array that is being used, MOUSEDIVm520650 (MDGA). Genotyping results were generated for a total of 384 samples (351 samples downloaded from Centre for Genome Dynamics and 33 samples from in-house experiments). Each genotyping run was completed by loading all CEL files into Genotyping Console; along with a list denoting the gender for each sample and a specific SNP list to be used for genotyping. The algorithm used to genotype the MDGA is the BRLMM-P algorithm, which is specific to perfect match probes, with default settings and a score threshold of 0.1.[108] Using this algorithm, one of four genotypes is assigned to each SNP in the genotyping list for every sample (Figure 2-2). The first two genotype calls are homozygous genotypes, one homozygous for the A allele denoted in the genotyping results as "AA", and the other homozygous for the B allele denoted as "BB" in the genotyping results. The third possible genotype call is the presence of both the A and the B alleles and is annotated as "AB". Finally, the fourth possible genotype call is when no genotype can be determined and is shown as a "NoCall" in the genotyping results. After one round of genotyping, all samples that failed to pass a minimum threshold of 97% overall call rate were considered to have failed genotyping. Overall, genotype call rate is the total number of SNPs with an identified genotype; the total number of AA, AB, and BB calls. All failed samples were identified and removed from analysis. A second round of genotyping was performed on all samples that passed this 97% cutoff, a threshold recommended in the Genotyping Console users guide. Results from the second round of genotyping again yielded an AA, AB, BB, or NoCall for each SNP in the genotyping list. Genotypes for each sample were exported into Microsoft Office Excel® 2007 (Microsoft, Redmond, Washington) and cross-referenced with the annotated SNP list of filtered SNP probes to determine the corresponding chromosome and locus. These genotyping results were used for analysis of genetic distance, genetic background, and somatic mosaicism.

Figure 2-2: **Possible genotype calls from the Mouse Diversity Genotyping Array.** The four possible genotype calls are listed. a) The first two possibilities are homozygous genotypes which include either homozygous A (AA) or homozygous B (BB) genotypes. b) the third possible genotype is heterozygous (AB). c) the fourth possibility is the failure of a genotype to be determined, or unknown genotype (NoCall).



<u>Array Call</u> = (AA or BB)
Genotype call: Homozygous (AA or BB)

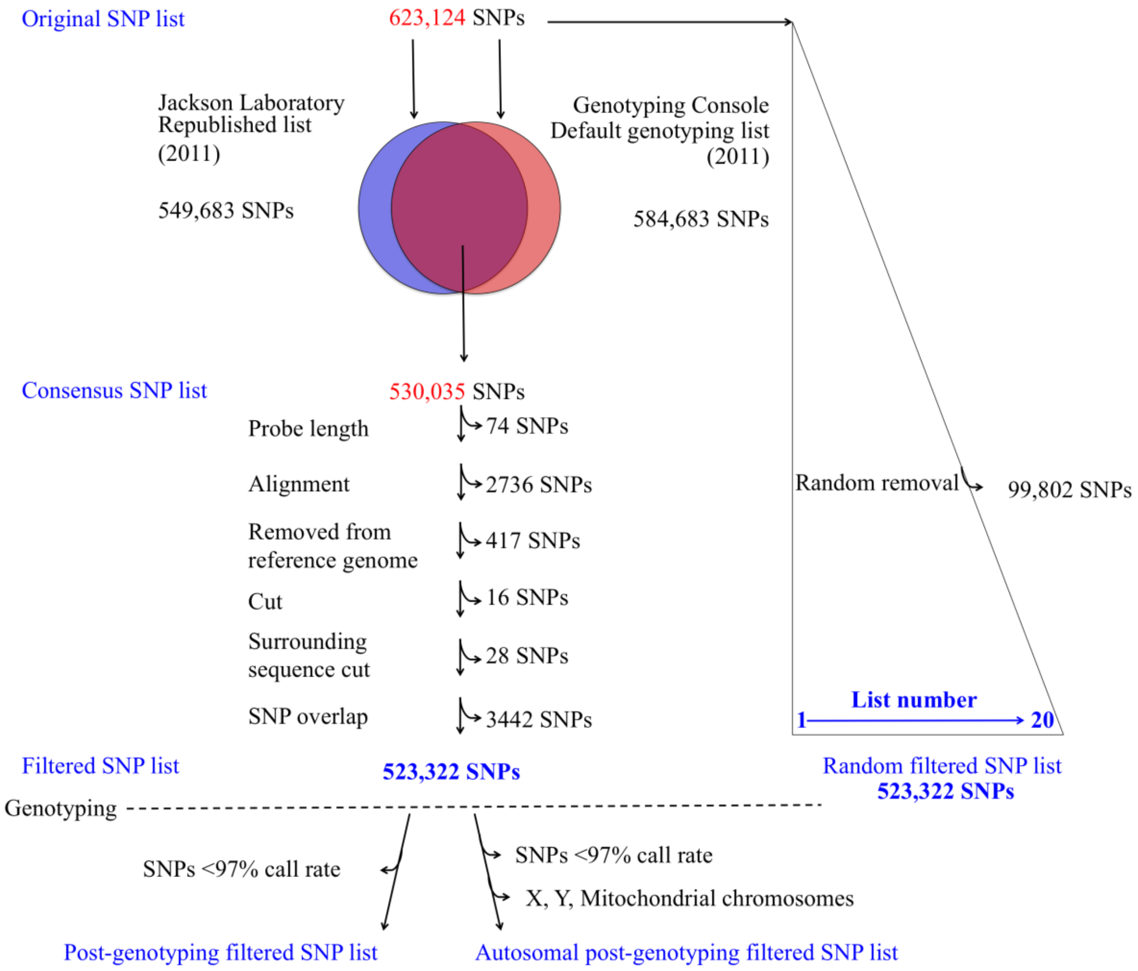
<u>Array Call</u> = (AB)
Genotype call: Heterozygous (A & B)

<u>Array Call</u> = (NoCall)
Genotype call: Unknown(- -)

2.4 Post-genotype filtering

Each SNP in the genotyping list was given an overall genotyping call rate. Overall genotype call rate is defined as the total number of AA, AB, and BB genotypes associated with a SNP across all samples that were genotyped. SNPs that failed to meet an overall SNP call rate of 97% were removed from the analysis after the second round of genotyping as recommended by the Genotype Console analysis guide. Samples were not regenotyped after removing SNPs that did not meet this overall call rate. All passing SNPs were carried through for further analyses of genetic distance, genetic background, and somatic mosaicism. For comparisons made using samples from The Jackson Laboratory dataset, only autosomal SNPs were used. This is because the dataset contains both male and female samples and the inclusion of SNPs from the X and Y chromosomes will make samples appear more different based on gender rather than genetic background. For genetic background comparisons of the Hill laboratory samples, the unknown origin of the X chromosome from the maternal founder mouse resulted in the removal of the X chromosome from these analyses. Comparisons for somatic mosaicism detection included the X chromosome in analysis because comparisons were made only between male mice (Figure 2-3).

Figure 2-3: **Lists of SNP probes generated to be used before and after genotyping.** The original SNP list contains all SNPs that were printed on the array (Yang *et al,* 2009). This original SNP list was filtered by The Jackson Laboratory and Genotyping Console to generate two separate lists that had poorly performing probes removed. From The Jackson Laboratory 2011 list and the Genotyping Console 2011 list a consensus list was generated taking all probes that overlapped between the two lists. Probes were removed based on specific probe design criteria to generate a filtered SNP list. Twenty randomly filtered SNP lists were also created, each containing the same number of SNPs as the filtered SNP list that had been removed from the original SNP list. After genotyping, two lists were created, one removing all SNPs that contained a genotyping call rate of less than 97% "post-genotyping filtered SNP list", and a second removing all SNPs that contained a genotyping call rate of less than 97% as well as the sex and mitochondrial chromosomes "autosomal post-genotyping filtered SNP list".


Original SNP list
623,124 SNPs
Jackson Laboratory
Republished list
(2011)
549,683 SNPs
Genotyping Console
Default genotyping list
(2011)
584,683 SNPs
Consensus SNP list
530,035 SNPs
Probe length
74 SNPs
Alignment
2736 SNPs
Removed from
reference genome
417 SNPs
Cut
16 SNPs
Surrounding
sequence cut
28 SNPs
SNP overlap
3442 SNPs
Random removal
99,802 SNPs
List number
1
20
Filtered SNP list
523,322 SNPs
Random filtered SNP list
523,322 SNPs
Genotyping
SNPs <97% call rate
SNPs <97% call rate
X, Y, Mitochondrial chromosomes
Post-genotyping filtered SNP list
Autosomal post-genotyping filtered SNP list

2.5 Comparing genotyping accuracy

Genotyping accuracy, defined as the total number of CEL files that passed the genotyping threshold of 97% after one round of genotyping, was compared between the original SNP list, the Genotyping Console SNP list, and the filtered SNP list. All 351 CEL files, each representing a different mouse, from The Jackson Laboratory dataset of CEL files from the MDGA were used to test the genotyping accuracy of the three SNP lists. Genotype analysis was completed three separate times, one time for each of the SNP lists. After genotyping, the total number of samples that failed to meet the overall call rate of 97% was compared between the original SNP list, the Genotyping Console SNP list, and the filtered SNP list.

2.6 Testing genotyping accuracy

The genotyping accuracy of the filtered SNP list was tested against 20 lists where probes were removed at random. These 20 lists all contained the same number of SNPs as the filtered list; however, removal of SNP probes was completely random. These 20 randomly generated SNP lists had probes removed at random using a random number assignment in Excel. Numbers were randomly assigned to each SNP between a range of 1 and 623,124, they were then sorted based on the random number assignments, and the first 523,322 SNPs were included in the random filtered SNP list. Each random filtered SNP list was then imported into Genotyping Console and 20 rounds of genotyping were preformed, one round with each list. Each list was used to genotype 351 samples and had a gender file denoting the gender for each sample. The total number of samples that failed to pass the 97% genotyping threshold was counted for each of the randomly generated lists. This number was compared to the total number of samples that failed to pass the 97% genotyping threshold with the filtered SNP list after a single round of genotyping.

2.7 Animal Care

All protocols were approved by The Canadian Council on Animal Care 2009-033 (Appendix A). Mice selected for the study were 8-month and 15-month-old males hemizygous for the *harlequin* (*hq*) mutation ($X^{hq}Y$), 11-month-old female carriers hemizygous for the *hq* mutation ($X^{hq}X$), and 8-month-old male mice that carry the wild-type allele for the *Apoptosis-inducing factor* (*Aif)* genotype (XY) (mixed genetic background: B6CBACaA^{w-J}/A-Pdcd8^{Hq}/J) (Jackson Laboratories, Bar Harbour, ME). Mice on a pure C57BL/6J (B6) (Taconic Farms, Germantown, NY) genetic background were also selected at 10 months-of-age and carried the wild-type allele for the *Aif* gene (XY). Mice were housed at 21±1°C on a 14/10 hour light/dark cycle. Diets provided *ad libitum* for all mice were standard (PMI Foods, St. Louis, MO) with water provided *ad libitum*.

2.8 Genotyping for the *Apoptosis-inducing factor* gene

Mice were ear notched at 10-12 days of age for identification purposes. Tail clippings or ear pieces were used to determine the genotype at the *Aif* locus in all samples to determine the presence of the mutated allele of samples as *Aif*-proficient WT mice or *Aif*-deficient *hq* mice. Genotyping was completed using the Terra-PCR kit (Qiagen, Valencia, CA) with crude tissue extracts and a tri-primer mixture to amplify a portion of both the WT *Aif* sequence and the proviral insertion responsible for the *hq* phenotype.[51] Primers included an exonic forward primer, *Aif* 3468 5’ AGT GTC CAG TCA AAG TAC CGG G 3’and reverse *Aif* 4000 5’ CTA TGC CCT TCT CCA TGT AGT T 3’ primer. A third primer *hq*LTR2 5’ GAA CAA GGA AGT ACA GAG AGG C 3’was used to amplify the long terminal repeat of the murine C-type ecotropic virus inserted into *Aif* intron one of mice hemizygous for the *Aif* insertion. Cycling conditions included an initial denaturing step at 98°C for 2 minutes, followed by 38 cycles of 98°C for 10 seconds, 60°C for 15 seconds, and 68°C for 30 seconds (GeneAmp

PCR System 9600, Applied Biosystems, Foster City, CA). Amplicons were electrophoresed through a 1.5% agarose gel stained with SYBR® Safe (Invitrogen, Life Technologies, Burlington, ON, Canada) for visualization of amplicons indicative of *Aif* genotypes.

2.9 Tissue harvesting

Mice were euthanized at 8, 10, or 15 months-of-age with carbon dioxide (CO_2) inhalation according to Animal Care and Veterinary Services (ACVS standard operating procedure (SOP)) 320-02 in an enclosed chamber prior to harvesting tissues. After euthanization, the spleen, cerebellum, and liver were harvested and snap-frozen in liquid nitrogen to prevent the degradation of DNA. Tissues were later transferred to a -80°C freezer for long-term storage.

2.10 DNA extractions

Harvested tissues were subjected to one of two high molecular weight DNA extraction procedures at The Jackson Laboratory (Bar Harbour, MA), London Regional Genomics Center (LRGC) (Robarts Research Institute, London, ON, Canada), or in-house (Table 2-2). Samples that were extracted at The Jackson Laboratory used the Wizard® Genomic DNA Purification Kit (Promega, Madison, WI) for 40-70 mg of animal tissues. London Regional Genomics Center isolated tissues with both the Gentra® Puregene® Kit (Qiagen, Mississauga, ON) and the Wizard® Genomic DNA Purification Kit, both with 7-11 mg of tissue. In-house DNA extractions were completed using the Wizard® Genomic DNA Purification Kit by homogenizing 6-10 mg of thawed tissues. In-house modifications made to the Wizard® Genomic DNA Purification Kit protocol include two additional centrifugations

Table 2-2: Summary table of the 27 samples from the Hill laboratory

Mouse ID[a]	Tissue Type[b]	Age (months)[c]	Sex[d]	*Aif* genotype[e]	DNA Extraction Protocol[f]	Amount of Tissue (mg)[g]	Processing Location[h]
911.143	Cl	15.2	M	$X^{hq}Y$	Wizard	40.0	JAX
911.148	Cl	15.2	M	$X^{hq}Y$	Wizard	40.0	JAX
911.50	Sp	8.7	M	XY	Wizard	30.0	JAX
911.50	Cl	8.7	M	XY	Wizard	70.0	JAX
911.49	Sp	8.7	M	XY	Wizard	30.0	JAX
911.49	Cl	8.7	M	XY	Wizard	60.0	JAX
904.9	Sp	7.7	M	$X^{hq}Y$	Wizard	50.0	JAX
904.9	Cl	7.7	M	$X^{hq}Y$	Wizard	60.0	JAX
904.11	Sp	7.7	M	$X^{hq}Y$	Wizard	30.0	JAX
904.11	Cl	7.7	M	$X^{hq}Y$	Wizard	40.0	JAX
300.6	Cl	10.4	M	XY	Gentra	42.0	LRGC
900.3	Sp	11.3	F	$X^{hq}X$[i]	Gentra	10.0	LRGC
900.3	Li	11.3	F	$X^{hq}X$[i]	Gentra	9.4	LRGC
911.50	Sp-2	8.7	M	XY	Gentra	8.6	LRGC
911.50	Cl-2	8.7	M	XY	Gentra	11.0	LRGC
911.50	Li	8.7	M	XY	Wizard	10.0	LRGC
300.7	Cl-1	10.4	M	XY	Wizard[j]	7.0	LRGC
300.7	Cl-2	10.4	M	XY	Wizard[j]	9.0	LRGC
300.7	Cl-3	10.4	M	XY	Wizard[j]	8.5	LRGC
300.7	Sp-1	10.4	M	XY	Wizard[j]	6.7	LRGC
300.7	Sp-2	10.4	M	XY	Wizard[j]	8.8	LRGC
300.7	Sp-3	10.4	M	XY	Wizard[j]	7.0	LRGC
911.17	Sp	8.9	M	XY	Wizard[j]	7.5	LRGC
911.17	Li	8.9	M	XY	Wizard[j]	8.8	LRGC
911.17	Cl	8.9	M	XY	Wizard[j]	6.6	LRGC
911.49	Li	8.7	M	XY	Wizard[j]	8.8	LRGC
911.50	Li-2	8.7	M	XY	Wizard[j]	7.0	LRGC

[a] the specific identifier that was assigned to each of the mice

[b] the tissue samples analyzed using the MDGA (Sp = Spleen, Cl = Cerebellum, Li = Liver), replicates are indicated in by “- #”

[c] the age of the mouse at euthanization and tissue harvest

[d] M=male, F= female

[e] refers to the presence or absence of the *harlequin* mutation

[f] the method of DNA extraction. Wizard refers to Wizard® Genomic DNA Purification Kit (Promega, Madison, WI) and Gentra refers to the Gentra® Puregene® Kit (Qiagen, Mississauga, ON).

[g] mass of tissue (mg) that was used for the high molecular weight DNA extraction

[h] where the genomic DNA was processed, hybridized to the array, and scanned. The Jackson Laboratory (JAX) and London Regional Genomics Centre (LRGC) .

[i] indicates a carrier for the *harlequin* mutation

[j] refers to extractions performed in house

of DNA samples during the protein precipitation step, one at 16,000xg for 4 minutes and one at 16,000xg for 8 minutes. Pellets were dried overnight and resuspended at room temperature for 8 hours before refrigerating (4^oC) until use. All samples were electrophoresed through a 1.5% agarose gel with a 1 kb ladder (Invitrogen, Life Technologies, Burlington, ON, Canada) to verify high-molecular weight DNA extraction. DNA quality ratios, 260/280 and 260/230, and quantity were also tested using a Nanodrop ND-1000 spectrophotometer (Thermo Scientific, Ottawa, ON, Canada). CEL files obtained from The Jackson Laboratory dataset were tail DNA samples isolated at The Jackson Laboratory using the Wizard® Genome DNA purification Kit for animal tissues.

2.11 Preparation of extracted DNA

DNA extractions completed at The Jackson Laboratory were prepared for microarray hybridization and scanned at the same location. DNA extractions completed at the LRGC or in-house were prepared for microarray hybridization and microarrays were scanned at LRGC. Scanning of the microarrays using the GeneChip® Scanner 3000 7G (Affymetrix®, Santa Clara, CA) generated a CEL file containing the fluorescence intensities for each of the features on the array.

All samples were prepared for microarray hybridization following the protocol outlined for the Affymetrix® Genome-Wide Human SNP *Nsp/Sty* 6.0 Assay Kit 5.0/6.0 (Figure 2-4).[125] High-molecular weight DNA was digested in two separate restriction enzyme digests, one with *Nsp*I and the second with *Sty*I. Adaptors recognizing the overhangs produced by the restriction enzymes were then ligated to the ends of the digested DNA fragments. Primers that recognize the adaptor sequences were used to PCR amplify fragments ranging between 200-1100 bps. Amplicons from each of the digests were then pooled together and fragmented using a fragmentation buffer and reagent (DNase 1). This generated an average fragment size smaller than 180 bps. These fragments were end labeled using biotin at the 3’ end of the DNA fragment and denatured at 95 ^{o}C for 10 minutes. After denaturation the DNA was hybridized to the MDGA at 50 ^{o}C for 16 to 18 hours before scanning.

Microarrays were they stained so that a fluorescence signal was generated for each of the SNPs to be used for genotyping (Figure 2-5). Staining of the MDGA involves a three step process. The first staining is performed using streptavidin-phycoerythrin (SAPE). Phycoerythin is a chromophor which emits light at 578 nm. The emission of light by phycoerythin is used to produce the CEL image of fluorescence intensities. The second step in staining uses an anti-streptavidin antibody that is bound to a biotin molecule. This anti-streptavidin antibody recognizes the streptavidin in SAPE and provides additional binding sites for a second round of staining with SAPE. The third step in the staining process is the second round of staining with SAPE. The addition of SAPE a second time amplifies the initial signal that was created with the binding of SAPE directly to the DNA molecule. Microarrays were then scanned with the GeneChip® Scanner 3000 7G to generate a CEL file.

Figure 2-4: Overview of genomic DNA preparation for the array. DNA was digested in two separate restriction enzyme digests (*Nsp*I and *Sty*I) to produce fragments for PCR amplification. Adaptors were ligated to the ends of the digested DNA fragments that were recognized by PCR primers. PCR amplicons from each of the digests were then pooled together, fragmented with DNase 1, and 3' end labeled with biotin. Labeled fragments were then denatured and hybridized to the MDGA. After hybridization, microarrays were scanned and fluorescence intensities were interpreted into genotype calls.

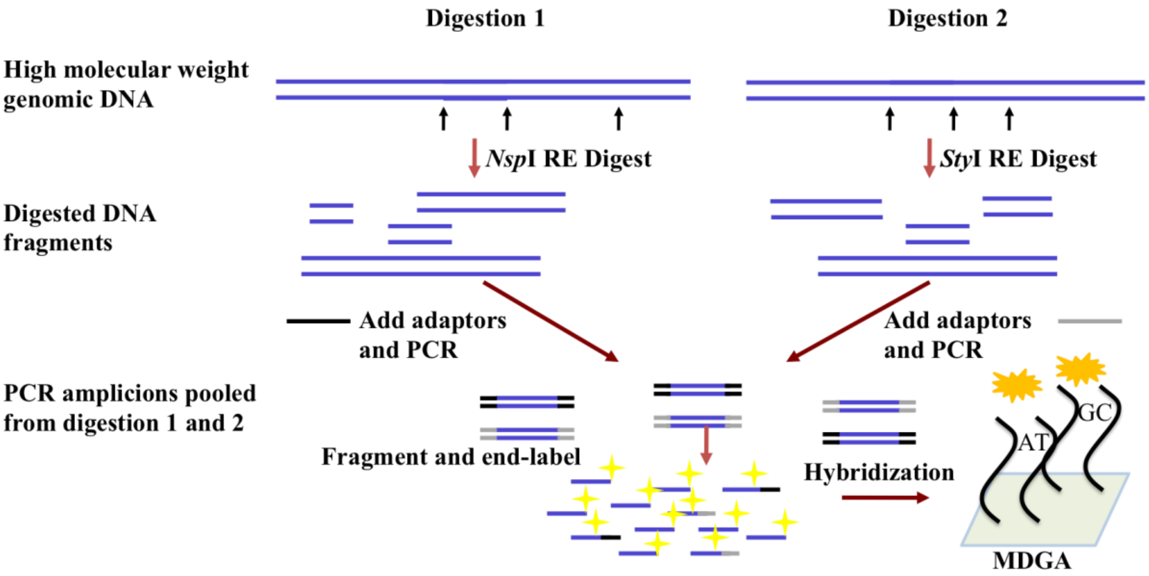
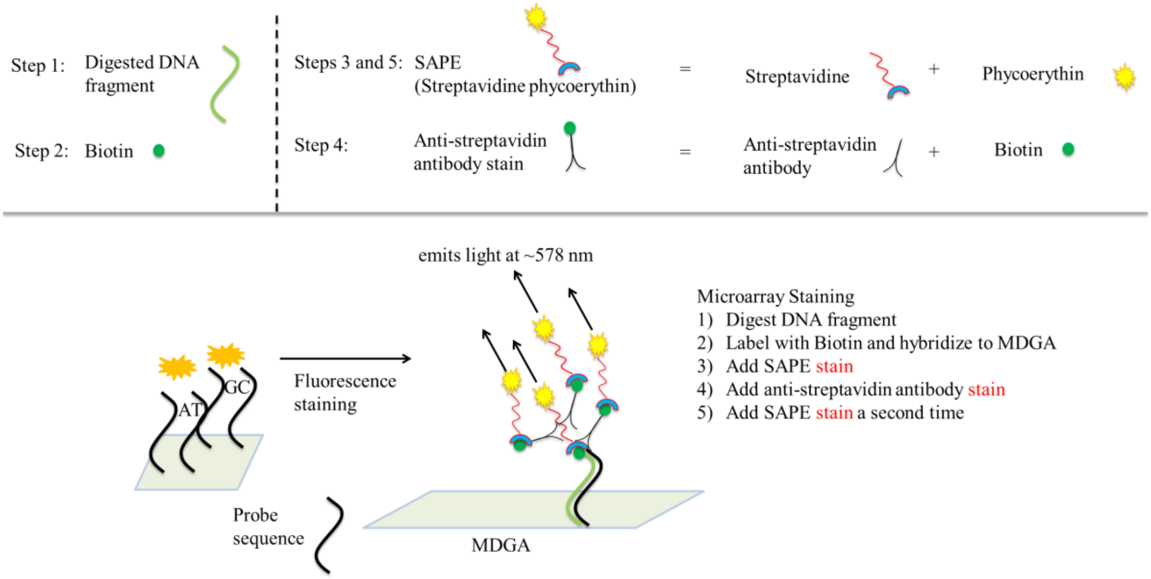

Digestion 1
Digestion 2
High molecular weight genomic DNA
*Nsp*I RE Digest
*Sty*I RE Digest
Digested DNA fragments
Add adaptors and PCR
Add adaptors and PCR
PCR amplicions pooled from digestion 1 and 2
Fragment and end-label
Hybridization
AT
GC
MDGA

Figure 2-5: The staining process for the Mouse Diversity Genotyping Array. Digested DNA fragments are biotin labeled and hybridized to the MDGA. This is followed by three rounds of staining. Staining steps are highlighted in red. The first stain is with Streptavidine phycoerythin (SAPE). SAPE contains the phycoerythin chromophore. An anti-streptavidin antibody stain is then applied which recognized the streptavidin component of SAPE. The antibody stain contains another biotin molecule which is recognized a second time by SAPE in the third staining step.


Step 1: Digested DNA fragment
Step 2: Biotin
Steps 3 and 5: SAPE (Streptavidine phycoerythin)
=
Streptavidine
+
Phycoerythin
Step 4:
Anti-streptavidin antibody stain
=
Anti-streptavidin antibody
+
Biotin
emits light at ~578 nm
GC
AT
Fluorescence staining
Probe sequence
MDGA
Microarray Staining
1) Digest DNA fragment
2) Label with Biotin and hybridize to MDGA
3) Add SAPE stain
4) Add anti-streptavidin antibody stain
5) Add SAPE stain a second time

2.12 Experimental Design

2.12.1 Genetic Distance

The genotyping capability of the MDGA was tested using a set of 351 publically available CEL files. These 351 CEL files represent samples genotyped using the MDGA by The Jackson Laboratory and are available from the Centre for Genome Dynamics (http://cgd.jax.org/datasets/diversityarray/CELfiles.shtml). The samples used in this dataset represent mice that are similar to classical laboratory mouse strains, primarily derived from the *Mus musculus domesticus* subspecies to more distantly related subspecies of mice, including *Mus musculus musculus*, *Mus musculus molossinus*, and *Mus musculus castaneus*. Each sample was grouped into one of eight categories depending on genetic background (Table 2-3). These samples were used to determine if the array was capable of differentiating between mouse strains that are genetically similar and genetically distant by comparison of genetic distance measures. Genetic distance measures were calculated in two ways, first as a comparison to a reference genotype and second through pairwise comparisons between each sample (Figure 2-6).

Table 2-3: Summary of sample types for the 351 CEL files available from The Jackson Laboratory

Type[a]	Number of CEL files
Classical Laboratory Strain[b]	120
Congenic[c]	1
Consomic[d]	10
BXD[e]	44
Wild-Derived Laboratory Strains[f]	58
F1 Hybrid[g]	55
CC-UNC G2:F1[h]	40
Wild Caught[i]	23

[a] Samples are ordered based on their potential to be divergent from the reference genome

[b] Mouse strains traditionally used in the laboratory such as the B6 mouse

[c] inbred strains that contain a single gene with a genetic background different from the rest of the genome

[d] inbred strains that contain a single chromosome with a genetic background different from the rest of the genome

[e] BXD mice inbred mouse strains that are derived from a B6 mouse crossed with a DBA mouse

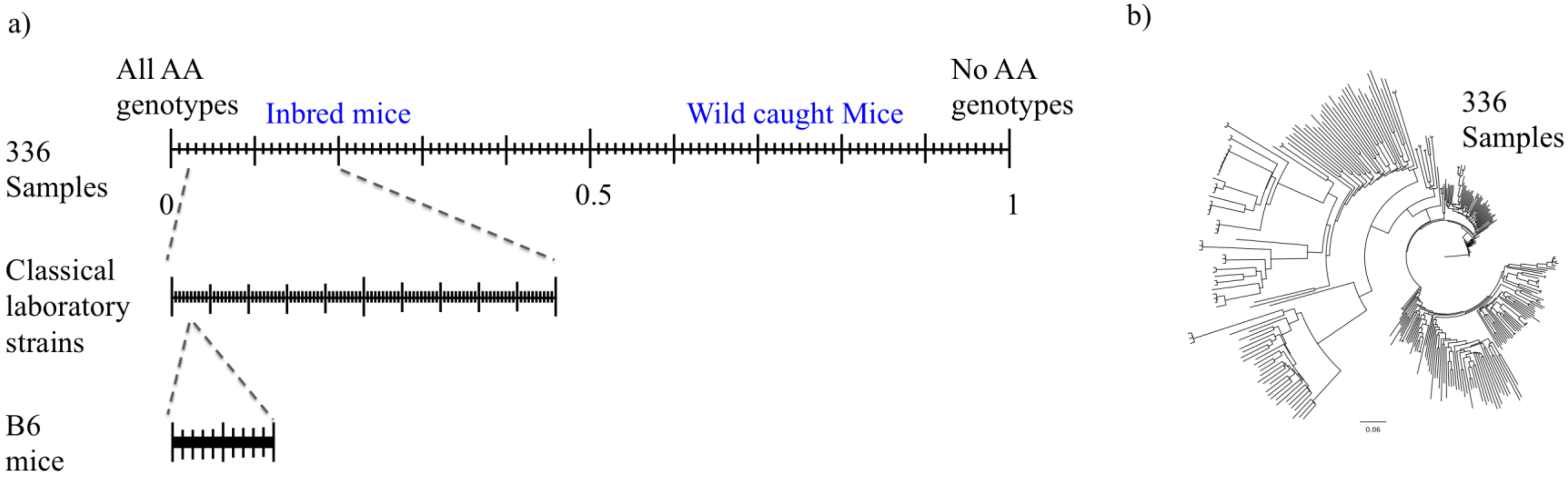
[f] inbred mouse strains that were derived from wild caught mice

[g] the first generation of offspring from parents that are from two different mouse strains

[h] collaborative cross mouse strain developed to include the genetic diversity of 8 inbred mouse strains

[i] mice that were captured from fields across Europe and Asia

Figure 2-6: Predicted relative genetic distance between the mouse strains from The Jackson Laboratory dataset. Predicted relative genetic distance between the 351 samples available from The Jackson Laboratory. a) shows the genetic distance as compared to a reference of all homozygous (AA) genotypes. b) shows the genetic distance as a phylogenetic tree resulting from pairwise comparisons between each sample.


a)
All AA genotypes
Inbred mice
Wild caught Mice
No AA genotypes
336 Samples
0
0.5
1
Classical laboratory strains
B6 mice
b)
336 Samples
0.06

2.12.2 Genetic Background

Twenty-seven additional samples ranging in genetic background from pure B6 to varying degrees of B6 and CBA/CaJ genetic background ratios (B6:CBA), along with eight B6 mice and one CBA/CaJ mouse from The Jackson Laboratory dataset were analyzed to determine the sensitivity of the MDGA in distinguishing a spectrum of genotypes (Figure 2-7). Genetic diversity was measured as the genetic distance between the mouse samples ranging across a spectrum of genetic backgrounds from a pure B6 genetic background to pure CBA/CaJ genetic background. Genetic distance measures were determined as a comparison to a reference genotype of all homozygous A calls, as well as with pairwise comparisons between each of the samples. The additional 27 samples (Table C-1), include a total of 10 mice, with tissues taken from the spleen (Sp) (n=10), liver (Li) (n=5), and cerebellum (Cl) (n=12). Mice with two or more different tissue types were used to determine if genetic differences can be detected between tissues of a single mouse. Replicates of the same tissue sample from the same mouse were used to determine the variation in genotype within a single tissue for samples 300.7 Sp, 300.7 Cl, 911.50 Sp, and 911.50 Cl.

2.12.3 Somatic Mosaicism

Somatic mosaicism was measured for the 27 additional samples by analyzing the genetic differences detected between tissues of the same mouse (Figure 2-8). The distribution of differences for each tissue was compared across chromosomes to determine if there was a tissue-specific profile to the accumulation of differences. Tissue-specific profile refers to the overall number of differences as well as the distribution of differences across the mouse genome in reference to specific tissues.

Figure 2-7: Experimental design for detecting the genetic distance between samples ranging from 100% C57Bl/6J mouse strain to 100 % CBA/CaJ mouse strain. The experimental design used to determine the range of genetic backgrounds from samples prepared in-house. a) Comparison of genotypes to a reference of all homozygous A (AA) genotypes to determine genetic distance measures. Genetic distance measures were used to determine the distance for mixed genetic backgrounds that ranged from 100% C57Bl/6J (B6) and 100% CBA/CaJ (CBA) mouse strains. Samples highlighted in blue indicate they are from the database of 351 Jackson Laboratory samples. Pure B6 genotypes were used to determine the sensitivity to detect genetic differences between tissues of the same mouse, and three replicates of the spleen (Sp) and three replicates of the cerebellum (Cl) were used to determine intra-tissue variation. b) Pairwise comparisons of genotype were used to generate a phylogenetic tree for the samples ranging in genetic backgrounds from pure B6 to pure CBA/CaJ.

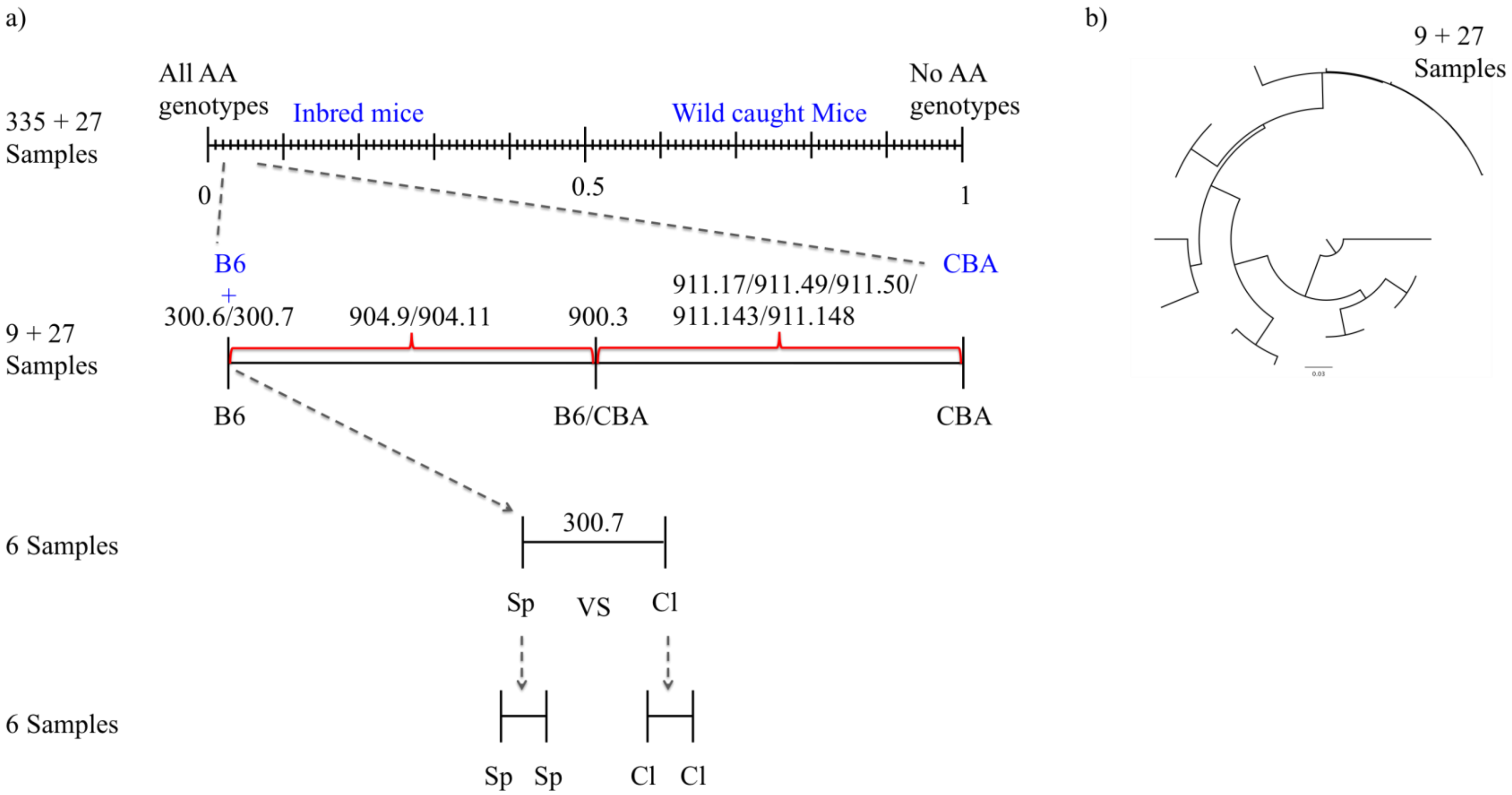
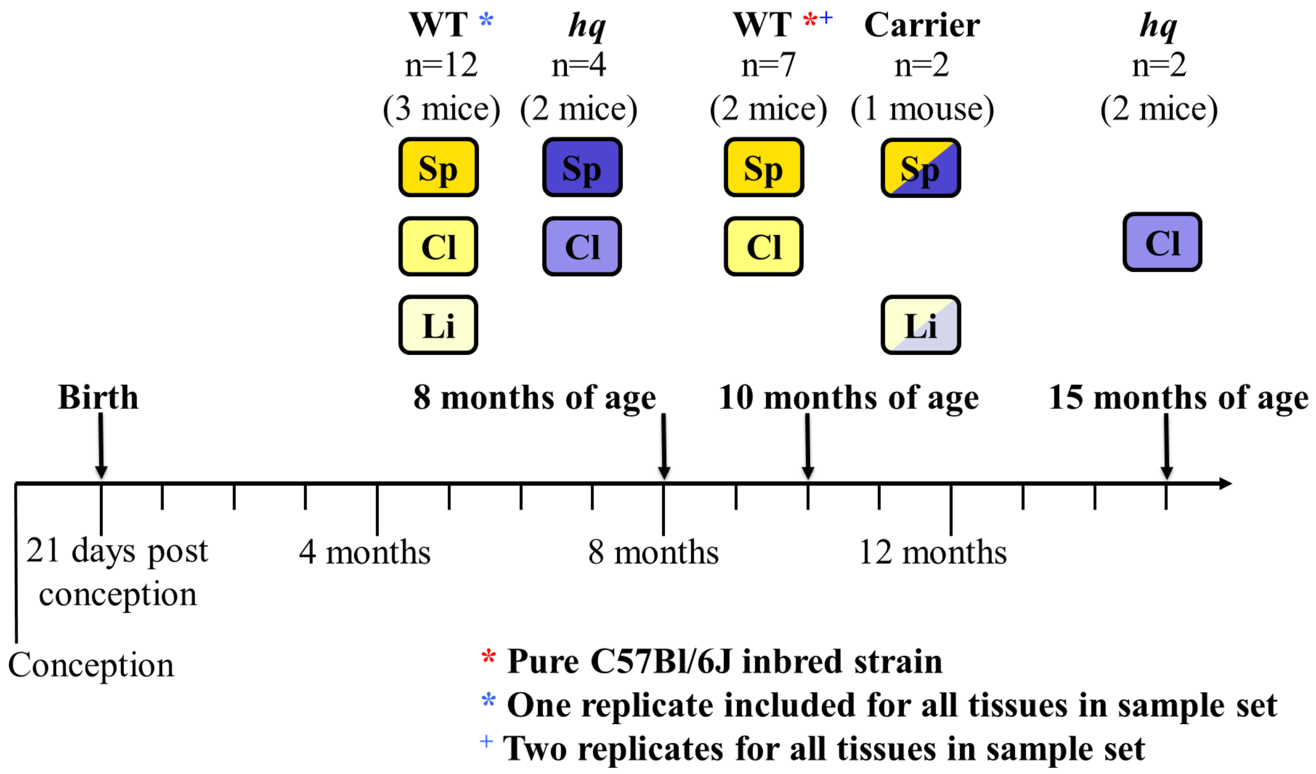

a)
335 + 27
Samples
All AA
genotypes
Inbred mice
Wild caught Mice
No AA
genotypes
0
0.5
1
B6
CBA
+
300.6/300.7
904.9/904.11
900.3
911.17/911.49/911.50/
911.143/911.148
9 + 27
Samples
B6
B6/CBA
CBA
6 Samples
300.7
Sp
VS
Cl
6 Samples
Sp
Sp
Cl
Cl
b)
9 + 27
Samples
0.03

Figure 2-8: Experimental design for the comparison of somatic mosaicism between the spleen, liver, and cerebellum of wild-type and *harlequin* mice. The experimental design used to infer estimates of somatic mosaicism between the spleen (dark shade), cerebellum (medium shade), and liver (light shade). Mice used in this study were wild-type (WT) (yellow) or *harlequin* (*hq*) (purple). Carrier mice are denoted as half purple and half yellow. Tissues were harvested from mice at 8 months-of-age, 10 months-of-age, and 15 months-of-age. Two replicates (totaling three samples per tissue) were included for one WT B6 mouse 10 months-of-age as well as a single replicate (totaling two samples per tissue) for one WT mouse 8 months-of-age. The second WT mouse at 10 months-of-age has only one tissue sample for the spleen and no sample for the cerebellum.


WT *
n=12
(3 mice)
Sp
Cl
Li
hq
n=4
(2 mice)
Sp
Cl
WT *+
n=7
(2 mice)
Sp
Cl
Carrier
n=2
(1 mouse)
Sp
Li
hq
n=2
(2 mice)
Cl
Birth
8 months of age
10 months of age
15 months of age
21 days post conception
4 months
8 months
12 months
Conception
* Pure C57Bl/6J inbred strain
* One replicate included for all tissues in sample set
+ Two replicates for all tissues in sample set

2.13 Determining genetic distance

2.13.1 Genetic distance measures from a reference

Genetic distance was calculated for each sample using genotypes determined after the second round of genotyping for the autosomal post-genotyping original SNP list and the autosomal post-genotyping filtered SNP list. Genetic distance was calculated as the total number of genotype differences from a reference containing homozygous A (AA) genotypes for all SNPs. Genotype calls contributing to a homozygous A genotype were based off the B6 reference mouse strain. Measures were taken for all autosomal SNPs that passed the 97% post-genotype filtering threshold. Genotype differences were identified as anything deviating from the homozygous A reference genotype (Homozygous B (BB), Heterozygous (AB), and NoCalls). The total number of differences from an AA reference was counted for each sample, and divided by the total number of autosomal SNPs that were compared to the reference. This generated a genetic distance value between 0 and 1 for each sample. For specific aims 2 and 3, genetic distance values were calculated using the autosomal post-genotyping filtered SNP list.

2.13.2 Pairwise genetic distance measures

Genetic distance measures were then calculated by pairwise comparisons for all samples using an iterative process in which each sample was set as a reference and compared to all remaining samples in the dataset. Again, genetic distance measures were calculated by determining the number of differences from the reference sample for all autosomal SNPs passing the 97% post-genotype filtering threshold. Any SNPs containing different genotype calls (AA, AB, BB, or NoCall) between two samples for all pairwise sample comparisons were identified as differences. In addition, SNPs where a “NoCall” was shared between the reference and the sample were also counted as differences in genotype. NoCalls are not informative as to why no genotype call was determined; therefore, an overestimate of genotype differences was determined by including shared NoCalls as genotype differences. The total number of differences, including shared NoCalls, was then divided by the total number of autosomal SNPs that were compared. This generated a genetic distance value for each sample between 0 and 1. This was repeated until each sample served as a reference. Genetic distances were compared between all mouse strains to create a distance matrix. This distance matrix was used to create a phylogenetic tree using the modified neighbour-joining algorithm, bionj function from the analysis of phylogenetics and evolution (APE) package in R.[126,127] Phylogenetic trees were saved in the Newick format and visualized using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).[128] A comparison between the matrices was made using a Mantel Test to determine if the underlying matrices used to build the trees were globally different between the autosomal post-genotyping original SNP list and the autosomal post-genotyping filtered SNP list.

A second set of genetic distance measures was calculated using pairwise comparisons of genetic differences between samples; however, this time excluding shared NoCalls. Excluding shared NoCalls from the differences gave a conservative estimate for genetic distance between the samples. Therefore, genetic distance was calculated as all differences between two samples divided by the total number of SNPs that were compared. Again, a Mantel Test was used to determine if the structure of the trees between the autosomal post-genotyping original and filtered lists was statistically different. This second type of pairwise comparison is what was used for the remainder of the determinations of genetic distance in specific aims 2 and 3.

2.14 Comparing genotyping accuracy using genetic distance measures from a reference

Genetic distance measures compared to a reference of all homozygous A genotypes were calculated using the autosomal post-genotyping original SNP list as well as the autosomal post-genotyping filtered SNP list. Genetic distance values were separated by strain type and compared to determine if there was a significant difference between values calculated with each of the SNP lists. Genetic distance values for each strain type were tested using the Shapiro-Wilk test (Cytel®, Cambridge, MA) to determine if genetic distance values for each of the genotyping lists was normally distributed. Strain types were then compared between the original and filtered SNP lists to determine if genetic distance values shifted significantly using a Mann-Whitney U test because samples were not normally distributed (Wilcoxon-Mann-Whitney test) (R, Vienna, Austria).

2.15 Percentage of genetic variation detectable

The percentage of genetic variation detectable using the list of filtered SNPs was determined for all samples passing the first round of genotyping. The percentage of genetic variation detectable was calculated using the autosomal post-genotyping filtered SNP list. Percent genetic variation was determined as the total number of SNPs from this list that could detect a difference in genotype in at least one of the samples divided by the total number of SNPs queried. The percentage of genetic variation was then compared for each of the sample types to determine the amount of genetic variation captured within each group. For each sample type, all SNPs with one or more samples varying in genotype call were added. These values were then divided by the total number of SNPs capable of detecting variation within all of the samples regardless of sample type.

2.16 Analysis of genetic variation for C57BL/6J mice

2.16.1 Percentage C57Bl/6J genetic background

In the available set of 351 CEL files, there were tail samples from eight B6 mice. Genetic differences between these eight mice were measured using the autosomal post-genotyping filtered SNP list after two rounds of genotyping (Table 2-4). The total percentage of genotypes that are B6, as annotated in the SNP list, was calculated for each mouse. Only SNPs that were capable of detecting a B6 genotype as determined in the reannotations were included in the analysis.

2.16.2 Genetic variation within C57Bl/6J mice

Genotype calls were compared between the eight B6 samples for each SNP to determine the total number of genotype differences between the mice. Differences in genotype calls were then attributed to each sample by determining which mouse contained the "different" genotype call. The total number of differences observed between each mouse was compared as a ratio of the total number of genotype differences to the total number of SNPs queried using and compared as an average ± SEM. Each genotype difference is associated with a SNP, and therefore a chromosome and chromosome position. The total number of genotype differences for each of the samples was divided among the chromosomes based on the known chromosome location within the mouse reference genome. The observed number of genotype differences on each of the chromosomes was then compared between samples with a Fisher's Exact Test with Monte Carle Simulation (MCS) (Cytel®, Cambridge, MA). The distribution of differences across each of the autosomes was recorded and graphed using Circos for each of the eight samples to determine if the differences were localized to specific regions across each of the chromosomes.

Table 2-4: B6 samples from the 351 publically available CEL files from the Center for Genome Dynamics

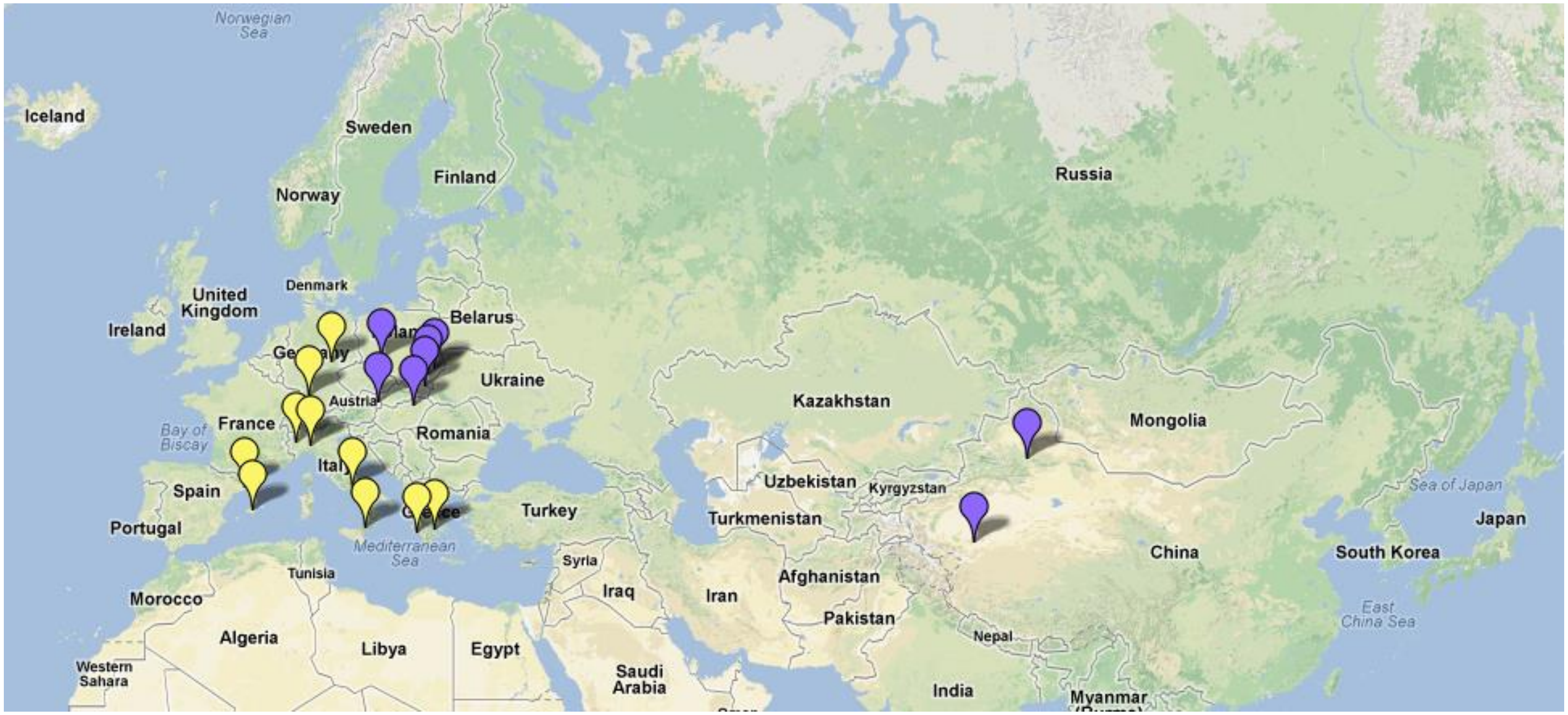
Cel file identifier	Sample Identifier
SNP_mDIV_A7-7_081308	B6 (1)
SNP_mDIV_A1-SNP08_001_103008	B6 (2)
SNP_mDIV_A2-SNP08_001_103008	B6 (3)
SNP_mDIV_A3-SNP08_001_103008	B6 (4)
SNP_mDIV_A4-SNP08_002_103008	B6 (5)
SNP_mDIV_A6-SNP08_002_103008	B6 (6)
SNP_mDIV_A5-SNP08_002_103008	B6 (7)
SNP_mDIV_A5-378_121608	B6 (8)

Samples containing a genotype call that was shared between less than four samples were considered to have a “genotype difference”. In cases where there was an even split containing four B6 mice with one genotype and four B6 mice of another genotype, the four samples containing NoCalls were determined to contain the genotype difference. All differences detected between the eight B6 samples were analyzed to determine SNPs where mice were divided between two genotypes in ratios of 4:4 (4 mice with one genotype : 4 mice with a second), 4:3 (4 mice with one genotype : 3 mice with a second), or 5:3 (5 mice with one genotype : 3 mice with a second). The distribution of these SNPs across the autosomes was visualized with Circos to determine if they were localized to specific regions across the genome.

2.17 Analysis of genetic variation for wild caught mice

Nineteen wild caught mice from The Jackson Laboratory dataset that passed the first round of genotyping were compared to identify all SNPs with genotype differences. SNPs that contained differences in genotype calls between any of the wild caught samples were identified as containing genotype differences. The total number of SNPs on each chromosome that detected a genotype difference was totaled. These differences were not attributed to any particular samples because all of the wild caught mice were from different locations across Europe and Asia (Figure 2-9).

Figure 2-9: The geographic location where wild caught mice included in The Jackson Laboratory database were caught. Samples in yellow indicate *Mus musculus domesticus* and samples in purple indicate *Mus musculus musculus* (Center for Genome Dynamics; Google Maps).


Norwegian Sea
Iceland
Sweden
Finland
Norway
Russia
Denmark
United Kingdom
Ireland
Belarus
Germany
Poland
Ukraine
Austria
Bay of Biscay
France
Romania
Kazakhstan
Mongolia
Italy
Spain
Uzbekistan
Kyrgyzstan
Sea of Japan
Greece
Turkey
Portugal
Turkmenistan
Japan
Mediterranean Sea
China
South Korea
Tunisia
Syria
Iraq
Iran
Afghanistan
Morocco
Pakistan
East China Sea
Algeria
Libya
Egypt
Nepal
Western Sahara
Saudi Arabia
India
Myanmar (Burma)

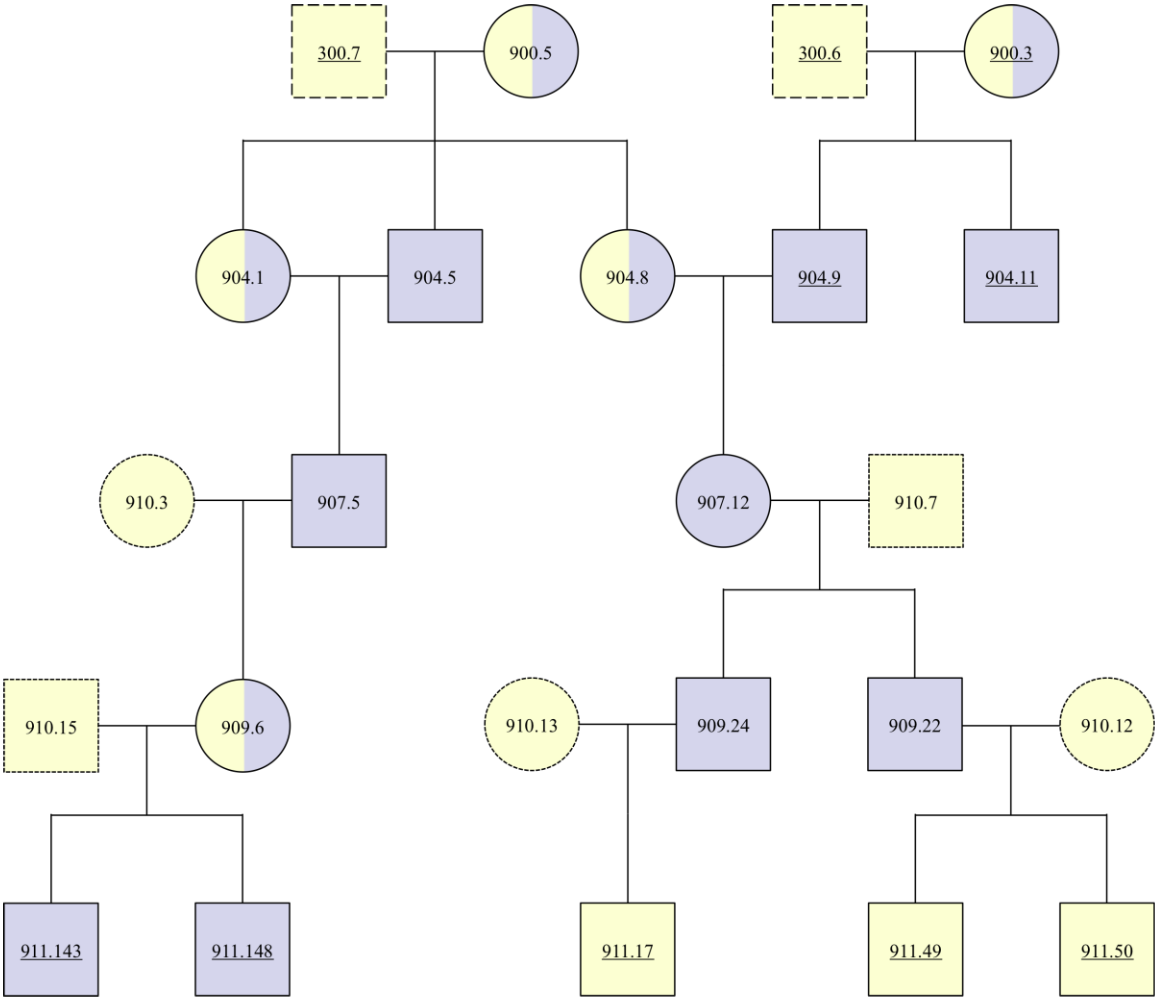
2.18 Comparison of C57BL/6J genotyping differences to wild caught genotyping differences

Differences identified between the eight B6 mice as well as the wild caught mice were compared to determine SNPs that were different between isogenic mice and wild caught mice. The list of genetic differences identified between B6 mice at ratios of 4:4, 4:3, and 5:3 was also compared against the list of differences identified between wild caught mice. All SNPs in common between the two lists were identified and analyzed for genomic position and genic regions.

2.19 Distinguishing mixed C57Bl/6J and CBA/CaJ genetic backgrounds

The genotype calls for each SNP were used to identify whether an allele was derived from a B6 or a CBA/CaJ mouse strain for samples ranging from pure B6 to pure CBA/CaJ genetic backgrounds. Samples ranging in genetic background between these two strains include the 27 samples from the Hill laboratory, the eight B6 samples from The Jackson Laboratory dataset, and the one CBA/CaJ sample also from The Jackson Laboratory dataset. Annotations from the list of filtered SNPs were used to identify the percentage of B6 and CBA/CaJ genotypes for each mouse for the post-genotyping filtered SNP list. This percentage was then compared to the expected genetic background identified from breeding records (Figure 2-10). Mice were ordered by percentage B6 to percentage CBA/CaJ to determine the range of genetic diversity between the two mice strains. Pairwise comparisons were also made to generate a phylogenetic tree of the 27 samples from the Hill laboratory, the eight B6 samples from The Jackson Laboratory dataset, and the one CBA/CaJ sample from The Jackson Laboratory dataset to determine if genotype could be used to accurately discriminate between a spectrum of mixed genetic backgrounds.

Figure 2-10: Mouse pedigree of samples from the Hill laboratory. Samples are labeled with their specific identifier number. Samples are coloured based on *Aif* phenotype, with yellow representing WT and purple representing *hq*. Samples that are carriers for the *hq* mutation ($X^{hq}X$) are half yellow, half purple. Samples that are underlined have been selected for analysis in this study. Samples that are outlined with dashes indicate isogenic backgrounds, large dashes are pure C57BL/6J (B6) and small dashes are pure CBA/CaJ (CBA). Remaining samples have a range in their mixture of B6:CBA genetic background.


300.7
900.5
300.6
900.3
904.1
904.5
904.8
904.9
904.11
910.3
907.5
907.12
910.7
910.15
909.6
910.13
909.24
909.22
910.12
911.143
911.148
911.17
911.49
911.50

2.20 Genetic diversity and tissue type

Samples from the Hill laboratory include multiple tissues from the same mouse, which were compared to determine if the array was sensitive enough to detect genetic diversity between tissues. The sensitivity to detect differences between tissues of the same mouse was analyzed using the genetic distance measures and pairwise comparisons described previously by comparing the ranking of samples based on percentage B6 genetic background.

2.21 Genetic variation within tissues

Three replicates from the spleen and three replicates from the cerebellum of a pure B6 mouse were compared to determine the variation within each tissue. Differences were assigned to a single replicate by determining which of the three samples contained a genotype different from the other replicates. Replicates were compared using Fisher's Exact Test with Monte Carlo Simulation (MCS; 10,000 simulations), to determine if the total number of differences attributed to each sample (average ± SEM) and distribution of differences across the autosomes and X chromosome were significantly different. The distribution of differences for each replicate was then compared to a random distribution, in which differences were distributed at random across the genome, proportional to the length of the chromosome. Significance from a random distribution was determined using a Fisher's Exact Test with MCS.

2.22 Genetic variation between tissues

Comparisons were made between all combinations of spleen and cerebellum replicates, resulting in a total of nine pairwise comparisons (Sp-1 vs Cl-1, Sp-2 vs Cl-1, Sp-3 vs Cl-1, Sp-2 vs Cl-1, Sp-2 vs Cl-2, Sp-2 vs Cl-3, Sp-3 vs Cl-1, Sp-3 vs Cl-2, and Sp-3 vs Cl-3). For each of the nine pairwise comparisons, differences were then divided by tissue type, by determining which of the tissues contained a different genotype call (Table 2-5). When an AA, AB, or BB genotype of one tissue is compared to a NoCall in a second tissue, the tissue containing the NoCall contains the difference. When a homozygous genotype is compared to a heterozygous genotype, the tissue containing the heterozygous genotype contains the difference. Finally, when a homozygous A was compared to a homozygous B, the tissue containing the homozygous B genotype contained the difference. For each comparison, the total number of differences between the spleen and the cerebellum was compared to determine if there was a significant difference between the tissues (average ± SEM). The distribution of differences was also compared across all autosomes and the X chromosome, to determine if there was a significant difference between the tissues (Fisher's Exact Test with MCS).

2.23 Somatic Mosaicism

Genotype comparisons were made between the spleen, cerebellum, and/or liver tissue of each mouse. Differences in genotype between the tissues were identified for all autosomes and the X chromosome using the post-genotyping filtered SNP list, after the second round of genotyping. Genotypes were analyzed to identify locations where tissues did not share the same genotype calls. Differences in genotype calling were scored, and the tissue containing the unexpected genotype call was determined. The total number of differences that were attributed to each tissue was then compared to determine differences in frequency (average ± SEM) and distribution across the mouse genome (Fisher's Exact Test with MCS).

Table 2-5: Comparison of genotypes used to determine somatic mosaicism between pairs of tissues[a]

Tissue 1 genotype	Tissue 2 genotype	Genotype Difference
AA	AA	No difference
AA	AB	AB
AA	NoCall	NoCall
NoCall	AB	NoCall

[a] Two tissues of the same mouse should contain the same genotype calls. The tissue containing the difference contains the genotype that is not expected. Samples used for somatic mosaicism studies are on a B6 or mixed B6/CBA/CaJ genetic background and should therefore contain AA genotypes. When a NoCall occurs in only one tissue it is likely due to the failure of hybridization on the array due to a genetic variant.

Two-tissue and three-tissue comparisons were made depending on the mouse. All *hq* mice 8 months of age had the spleen and cerebellum tissues compared and all WT mice 8 months of age had the spleen, cerebellum, and liver compared. In two tissue comparisons, the genotypes of each tissue were compared to determine which tissue contained the different genotype call. The tissue containing the difference was identified as the tissue that contained the NoCall or non-homozygous AA genotype. For tissue triads of spleen, cerebellum, and liver genotype calls were compared to identify any differences between the three tissues (Table 2-6). Differences in tissue triad comparisons were assigned to the tissue containing the "one off" genotype. Where two tissues shared the same genotype call, and the third tissue differed from the other two. Any cases where all three tissues contained a different genotype call, a difference was counted for each of the three tissues. Total differences and the distribution of differences across the autosomes and the X chromosome were compared between tissues.

2.23.1 Comparisons between WT mice

The total number of differences within a tissue was compared between the spleen samples, between the cerebellum samples, and between the liver samples for the three WT mice. The distribution of differences across the autosomes and the X chromosome was compared between samples as well as in comparison to a random (or expected) distribution. A random distribution of differences was calculated for each separate sample, as the total number of observed differences divided by the total number of SNPs compared and multiplied by the length of each chromosome. This gives the expected number of SNP differences for each chromosome proportional to the number of SNPs on that chromosome.

Table 2-6: Comparison of genotypes used to determine somatic mosaicism studies between tissue triads[a]

Tissue 1 genotype	Tissue 2 genotype	Tissue 3 genotype	Genotype Difference
AA	AA	AA	No difference
AA	AB	AA	AB
AA	NoCall	NoCall	AA
NoCall	AB	AB	NoCall

[a] Three tissues of the same mouse are expected to have the same genotype calls. When these genotype calls are different the genotype call that is in the majority of tissues is the expected genotype. Therefore, the tissue containing the genotype not shared between the other tissues contains a difference.

2.23.2 Comparison of WT replicates

The total number of genotype differences between the spleen, cerebellum, and liver was compared for a replicate tissue sample of one WT mouse. The total number of differences between these three tissues (average ± SEM), as well as their distribution was compared using a Fisher's Exact Test with MCS. The distribution of the observed differences for each tissue was also compared to a random distribution (Fisher's Exact Test with MCS). Values for a random distribution in each of the three tissues were calculated as described above.

Tissue replicates were also compared to determine the variation within each of the tissues of a single mouse. Replicates for the spleen, cerebellum, and liver were compared to determine the number of differences between two samples of the same tissue. The distribution of these differences was also compared across the autosomes and the X chromosome to identify any significant differences in distribution.

2.23.3 Comparisons of *hq* mice

The total number of differences observed in the spleen of the two *hq* mice and between the cerebellum of the two *hq* mice was compared. The distribution of these differences across the autosomes and X chromosome was compared. The distribution of the observed differences was also compared for each sample to a random distribution as described previously.

Chapter 3 : Results

3.1 The filtered SNP list contains 523,322 SNPs

A consensus list of 530,035 SNPs was generated by taking the overlap between the list of 549,683 SNPs provided by The Jackson Laboratory and the list of 584,729 SNPs provided by Genotyping Console (Affymetrix$^{®}$). This consensus list was used to filter SNPs based on the seven probe criteria used for filtering probe sets.

A total of 74 SNPs were removed in the final filtered SNP list that had a probe length other than 25 bps. Another 2,736 SNPs had to be removed based on alignment and 417 SNPs were identified and removed as they were no longer present in dbSNP. Additionally, 16 SNPs were removed because their probe sequences contained a recognition sequence for both *Nsp*I and *Sty*I restriction enzymes. Another 28 SNPs were removed that contained recognition sequences for both enzymes within 10 bps surrounding the probe sequence. A total of 3,442 SNPs were removed because probe sequences detected more than one SNP represented on the array. A final SNP list containing 523,322 SNPs was generated which represents regions distributed across each of the chromosomes in the mouse genome (Figure 3-1; Appendix B). The 523,322 SNP list is the final list that should be used for genotyping.

3.2 Nearly all SNPs in the reannotated filtered SNP list can detect a B6 genotype

Overall, removal of SNPs was evenly distributed across the genome. No one chromosome showed a disproportionately high number of SNP removed (Figure 3-2). Each chromosome is represented proportionately on the array. The number of SNPs for each chromosome is proportional to chromosome length (Table 3-1). The average interSNP distance for the filtered SNP list increases from 4 kb to 4.9 kb from the original SNP list; however, the overall range remains similar (Figure 3-3).

Figure 3-1: The distribution of SNPs across the mouse genome for the filtered SNP list. SNP frequency is graphed alongside each of the chromosomes in blue. Regions where there are gaps in SNP coverage can be seen as areas with no colouring in the blue inner circle.

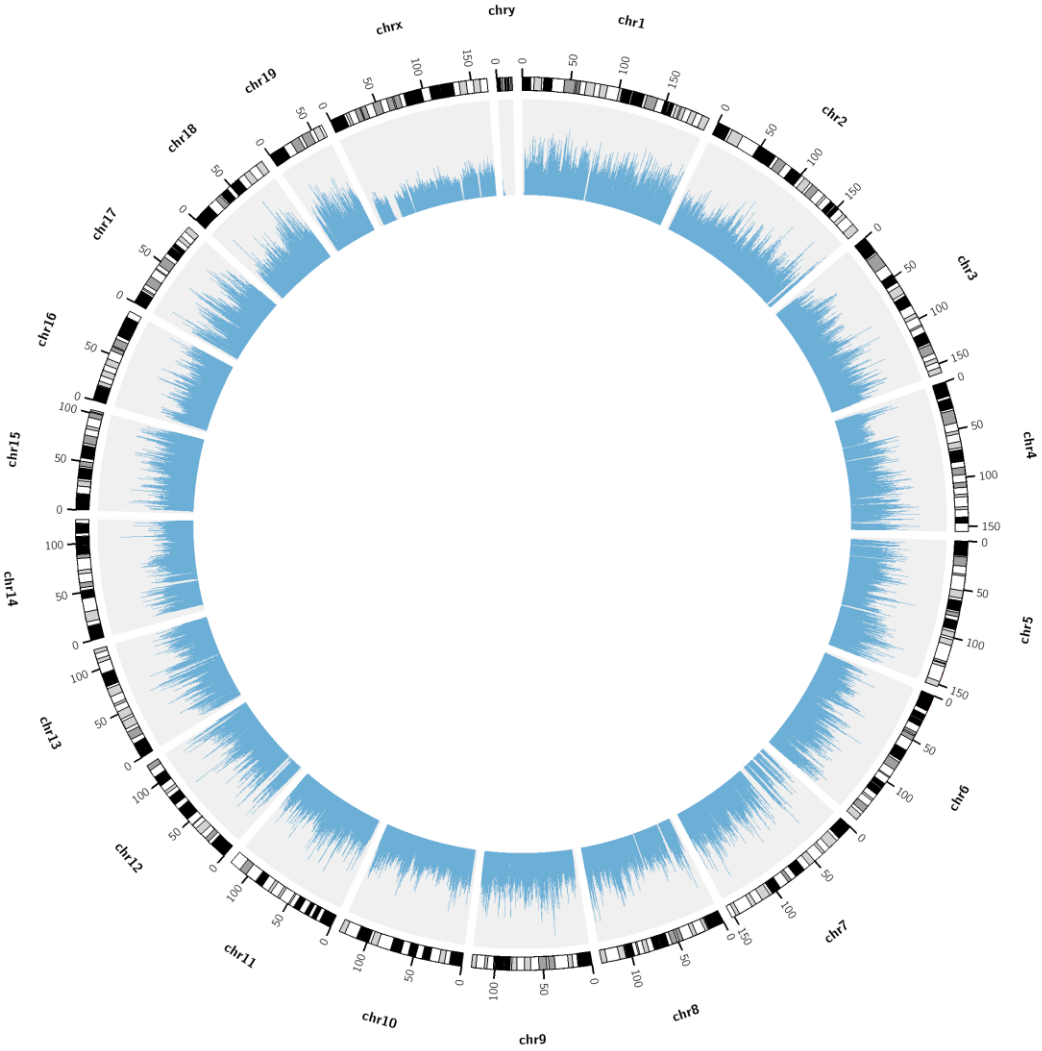

chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chrx
chry

Figure 3-2: **Total number of SNPs representing each chromosome in the mouse genome.** Grey plus black bars indicate the number of additional SNPs that are in the original SNP list. The filtered SNP list is represented by grey bars with a blue outline. The black bars represent the number of SNPs that have been removed during filtering.

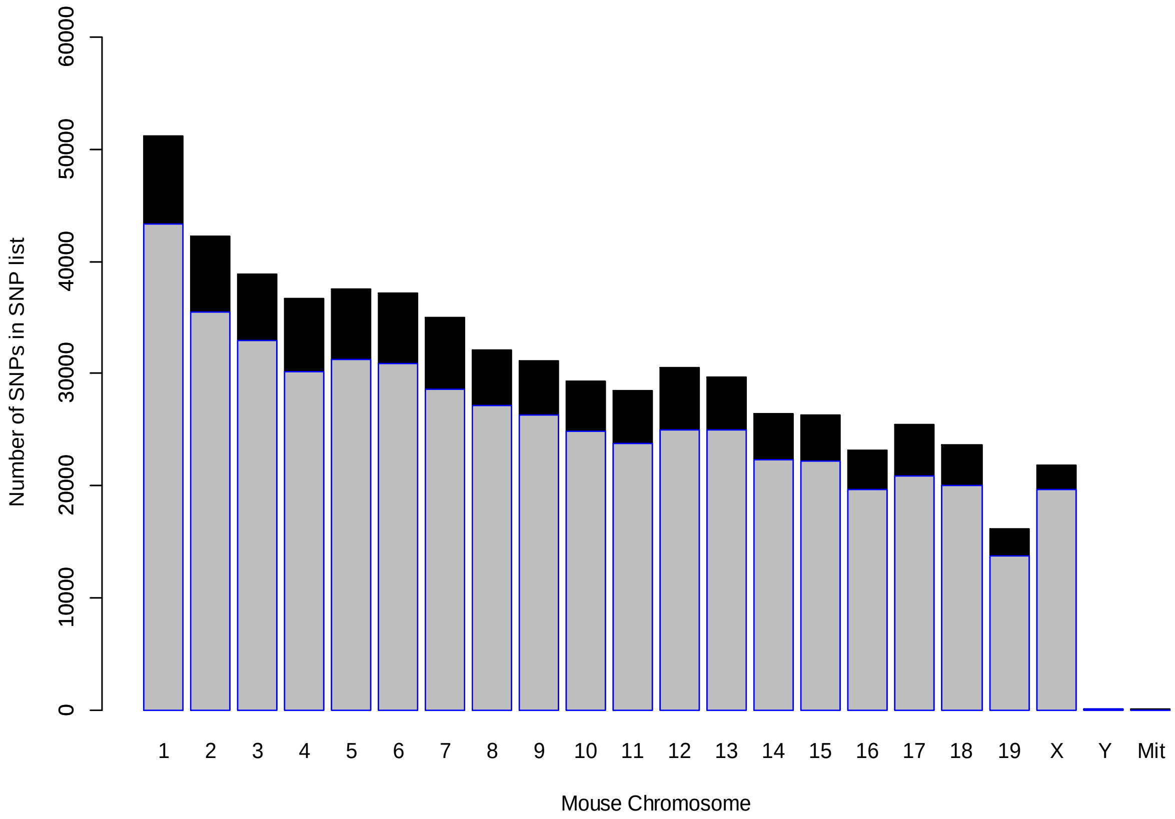

Number of SNPs in SNP list
0
10000
20000
30000
40000
50000
60000
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Y
Mit
Mouse Chromosome

Table 3-1: Overall representation of SNPs from each chromosome for the filtered SNP list

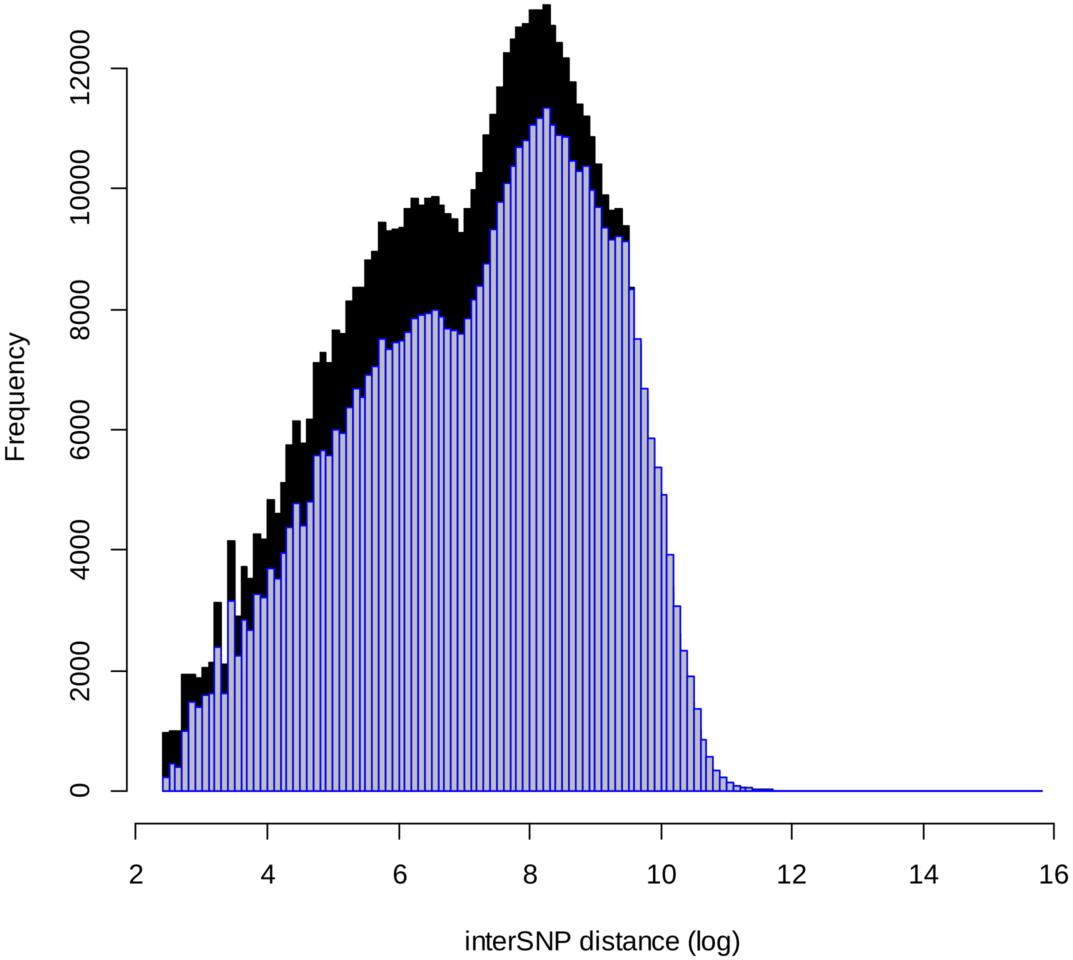
Mouse Chromosome	Chromosome Length (bp)	Number of SNPs	% coverage[a]	First SNP position (bp)[b]	Last SNP position (bp)[c]
1	197,195,432	43381	0.0220	3013441	197191885
2	181,748,087	35467	0.0195	3010110	181731538
3	159,599,783	32975	0.0207	3032787	159598996
4	155,630,120	30222	0.0194	3013031	155559551
5	152,537,259	31258	0.0205	3060235	152527863
6	149,517,037	30950	0.0207	3001551	149505351
7	152,524,553	28592	0.0187	3099583	152510110
8	131,738,871	27194	0.0206	3083611	131706890
9	124,076,172	26328	0.0212	3084105	124029861
10	129,993,255	24802	0.0191	3009197	129982856
11	121,843,856	23745	0.0195	3005934	121829163
12	121,257,530	24938	0.0206	3109153	121256667
13	120,284,312	24979	0.0208	3006383	120259651
14	125,194,864	22259	0.0178	6074266	125150241
15	103,494,974	22224	0.0215	3090615	103458457
16	98,319,150	19669	0.0200	3282623	98290818
17	95,272,651	20903	0.0219	3059769	95264893
18	90,772,031	20058	0.0221	3012495	90766409
19	61,342,430	13685	0.0223	3125547	61337203
X	166,650,296	19652	0.0118	4990476	166416379
Y	15,902,555	34	0.0002	37027	2634903
Mitochondrial	16,299	7	0.0429	2525	3443

[a] the total percentage of the chromosome that is represented by SNPs on the MDGA

[b] the chromosomal position of the first SNP on the array for each chromosome

[c] the chromosomal position of the last SNP on the array for each chromosome

Figure 3-3: InterSNP distance distribution for SNPs on the MDGA on a log scale. The black represents interSNP distances (bps) determined using the original SNP list. The grey with blue borders represents interSNP distances in the filtered SNP list. Each bar indicates an increment of 150 bps.


Frequency
0
2000
4000
6000
8000
10000
12000
2
4
6
8
10
12
14
16
interSNP distance (log)

The average interSNP distance for each chromosome varies depending on the chromosome (Table 3-2). The minimum average interSNP distance is 153 bps for the mitochondrial genome and maximum average interSNP distance is 787 kb for the Y chromosome. For autosomes, the minimum average interSNP distance is 4.2 kb and the maximum average genetic distance is 5.3 kb. The smallest interSNP distance is 12 bps and the largest interSNP distance is 7.3 Mb. The total number of SNPs representing each of the strategies used to identify the SNP has been broken down by chromosome for the filtered SNP list (Table 3-3). The number of SNPs that can detect a B6 allele, a CBA/CaJ allele, both a B6 and a CBA/CaJ allele, or neither a B6 or CBA/CaJ allele according to the B6:CBA/CaJ Annotations for each chromosome is listed in Table 3-4. The number of SNPs located in genic regions on each of the chromosomes ranges from 32% to 100% of SNPs detectable on the chromosome (Table 3-5). The proportion of SNPs capable of detecting a B6 allele differs with each chromosome and ranges from 99.10% to100% (Table 3-6).

3.3 Fewer samples fail when genotyping with the filtered SNP list

The total number of samples failing after one round of genotyping for the original SNP list, Genotyping Console SNP list, and filtered SNP list was 50, 29, and 15 respectively. The 20 randomly filtered SNP lists resulted in between 53 to 56 samples failing after one round of genotyping (Figure 3-4). The samples that failed using the filtered SNP list include classical laboratory mice, wild-derived laboratory mice, and wild caught mice (Table 3-7).

Table 3-2: InterSNP distance for each chromosome in the mouse genome for the filtered SNP list.

Mouse Chromosome	Average inter-SNP distance (bp) ± SEM	Smallest inter-SNP distance (bp)	Largest inter-SNP distance (bp)	Number of SNPs above average	% of SNPs above average + SEM
1	4476 ± 69	12	2610728	12573	28.98
2	5039 ± 83	12	2484127	10206	28.78
3	4748 ± 44	12	365224	9660	29.29
4	5048 ± 78	12	1170016	8693	28.76
5	4782 ± 87	12	2101102	9131	29.21
6	4734 ± 45	12	249605	9031	29.18
7	5226 ± 289	12	7268520	7218	25.24
8	4730 ± 109	12	2050016	7845	28.85
9	4594 ± 45	12	405837	7845	29.80
10	5120 ± 51	12	385400	7602	30.65
11	5004 ± 50	12	155954	6929	29.18
12	4738 ± 115	12	1751470	6854	27.48
13	4694 ± 73	12	905392	7176	28.73
14	5350 ± 130	12	1687052	6447	28.96
15	4516 ± 45	12	81398	6683	30.07
16	4831 ± 54	12	247483	5966	30.33
17	4411 ± 67	12	709221	5778	27.64
18	4375 ± 47	12	142451	5952	29.67
19	4254 ± 63	12	279160	3937	28.77
X	8215 ± 402	12	7033330	5632	28.66
Y	78724 ± 25825	47	606537	7	20.59
Mitochondrial	153 ± 58	12	435	1	14.29

Table 3-3: Summary of SNPs in the filtered SNP list by strategy used to identify each SNP

Mouse Chromosome	**Probe identification strategies**							
	Classical[a]	**NIEHS**[b]	**B6**[c]	**Wild**[d]	**MSM**[e]	**Y**[f]	**Mt**[g]	**Total**
1	10483	30598	188	13	2099	0	0	43381
2	9768	23521	91	9	2078	0	0	35467
3	8356	22655	116	7	1841	0	0	32975
4	7937	20530	121	9	1625	0	0	30222
5	7545	21929	83	8	1693	0	0	31258
6	7631	21711	88	4	1516	0	0	30950
7	6844	20323	139	31	1255	0	0	28592
8	6824	19075	87	8	1200	0	0	27194
9	6667	18085	107	25	1444	0	0	26328
10	6730	16433	73	19	1547	0	0	24802
11	6593	15672	92	6	1382	0	0	23745
12	5859	17785	82	16	1196	0	0	24938
13	6100	17512	77	10	1280	0	0	24979
14	6323	14792	72	20	1052	0	0	22259
15	5488	15500	64	20	1152	0	0	22224
16	5112	13266	59	6	1226	0	0	19669
17	4793	15035	82	15	978	0	0	20903
18	4696	14178	39	6	1139	0	0	20058
19	3120	9826	55	11	673	0	0	13685
X	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	7	0	7
Mitochondrial	7510	11294	3	25	820	0	34	19686
Total	134379	359720	1718	268	27196	7	34	523322

[a] SNPs identified from classical laboratory strains

[b] SNPs identified from studies conducted by the National Institute of Environmental Health Sciences (NIEHS)

[c] SNPs identified in the B6 mouse that are absent from the NIEHS SNPs that were not shared with A/J, DBA/2J, 129S1, and MSM/Ms strains.

[d] SNPs identified in *Mus macedonicus, Mus spretus, Mus spicilegus,* and *Mus musculus*

[e] SNPs identified in *Mus musculus molossinus*

[f] SNPs identified that are on the Y chromosome

[g] SNPs identified that are on the mitochondrial chromosome

Table 3-4: Summary of B6:CBA/CaJ annotations for the filtered SNP list

Chromosome	B6[a]	CBA[b]	Heterozygous[c]	Negative[d]	Total
1	33320	315	9738	8	43381
2	28962	250	6254	1	35467
3	25293	262	7417	3	32975
4	24445	211	5561	5	30222
5	24845	213	6198	2	31258
6	24694	239	6016	1	30950
7	22479	196	5913	4	28592
8	21226	224	5740	4	27194
9	20911	189	5226	2	26328
10	21504	217	3076	5	24802
11	18459	159	5124	3	23745
12	19121	170	5644	3	24938
13	19740	210	5024	5	24979
14	15853	188	6216	2	22259
15	17932	171	4119	2	22224
16	15383	140	4144	2	19669
17	15766	157	4978	2	20903
18	16661	155	3241	1	20058
19	10323	117	3245	0	13685
X	31	0	3	0	34
Y	7	0	0	0	7
Mitochondrial	17330	46	2273	3	19652
Total	414285	3829	105150	58	523322

[a] either the A allele or the B allele is present in the C57BL/6J (B6) mouse strain

[b] either the A allele or B allele is present in the CBA/CaJ mouse strain

[c] the A allele or B allele is present in the C57BL/6J (B6) mouse strain and the non-B6 allele is present in the CBA/CaJ mouse strain

[d] cannot detect the C57BL/6J allele or CBA/CaJ allele

Table 3-5: Summary of SNPs in genic regions in the filtered SNP list

Mouse Chromosome	Number of SNPs in Genes	% of SNPs in genes
1	18996	43.79
2	17637	49.73
3	12972	39.34
4	14120	46.72
5	15716	50.28
6	15526	50.16
7	14117	49.37
8	11851	43.58
9	13068	49.64
10	11326	45.67
11	12080	50.87
12	10799	43.30
13	10596	42.42
14	10074	45.26
15	9908	44.58
16	8815	44.82
17	10540	50.42
18	9079	45.26
19	7434	54.32
X	7020	35.72
Y	11	32.35
Mitochondrial	7	100.00

Table 3-6: The proportion of SNPs on each chromosome capable of detecting an allele that is present in the B6 genetic background.

Chromosome	**Number of SNPs that genotype a B6 mouse**	**% of SNPs that genotype a B6 mouse**
1	43058	99.26
2	35216	99.29
3	32710	99.20
4	30006	99.29
5	31043	99.31
6	30710	99.22
7	28392	99.30
8	26966	99.16
9	26137	99.27
10	24580	99.10
11	23583	99.32
12	24765	99.31
13	24764	99.14
14	22069	99.15
15	22051	99.22
16	19527	99.28
17	20744	99.24
18	19902	99.22
19	13568	99.15
X	19603	99.75
Y	34	100.00
Mitochondrial	7	100.00
Total	519435	99.26

Figure 3-4: Number of samples that failed after one round of genotyping for randomly generated SNP lists of 523,322 SNP probes. A total of 351 CEL files from the publically available Jackson Laboratory dataset were used to determine the genotyping accuracy of the filtered SNP list. The total number of samples that failed after one round of genotyping for the filtered SNP list is shown in red. The total number of samples that failed after one round of genotyping for the 20 randomly generated SNP lists is shown in black.

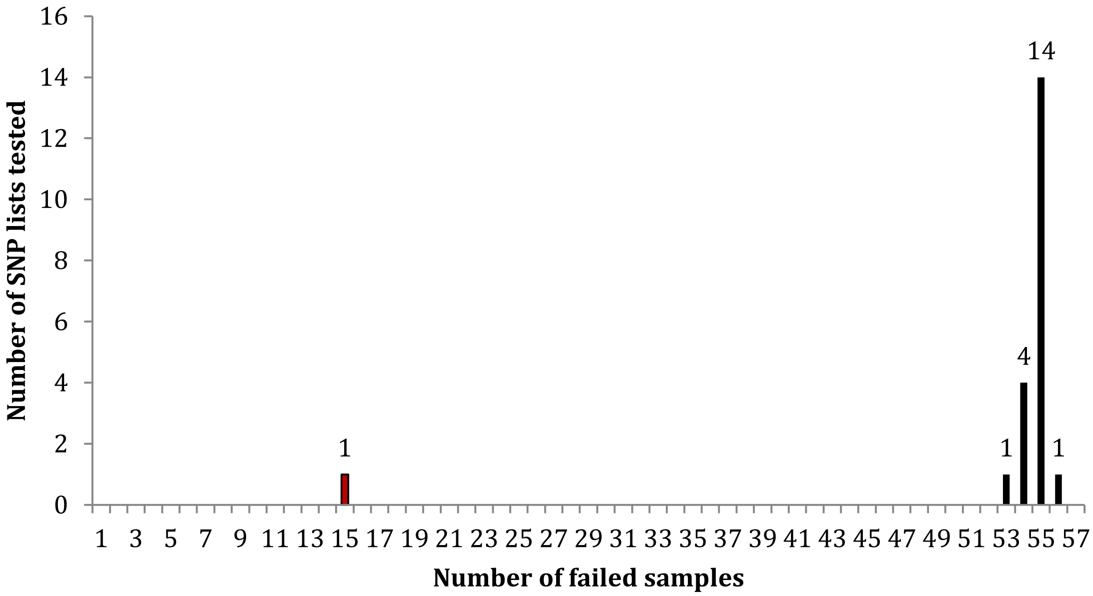

16
14
12
10
8
6
4
2
0
Number of SNP lists tested
1
1
4
14
1
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57
Number of failed samples

Table 3-7: Samples that failed round one of genotyping from the 351 publically available CEL files from The Jackson Laboratory dataset using the filtered SNP list

MDGA CEL file identifier	Mouse Gender	Overall call rate[a]	Sample Type	Mouse Strain
SNP_mDIV_A9-9_081308	Male	92.51	Wild-derived laboratory strain	CAST/EiJ
SNP_mDIV_D4-470_012209	Male	95.17	Wild-derived laboratory strain	MDGI
SNP_mDIV_B6-390_012709	Male	95.49	Classical laboratory strain	A/HeJ
SNP_mDIV_C10-121_090908	Male	95.98	Classical laboratory strain	ATEB/LeJ
SNP_mDIV_C11-35_081308	Male	96.29	Wild-derived laboratory strain	PWK/PhJ
SNP_mDIV_D6-133_090908	Male	96.31	Classical laboratory strain	MA/MyJ
SNP_mDIV_A6-54_082108	Male	96.32	Wild-derived laboratory strain	MSM/Ms
SNP_mDIV_B6-450_012209	Male	96.38	Wild-derived laboratory strain	PWK hybrid
SNP_mDIV_A4-157_091708	Female	96.52	Wild caught	RDS10105
SNP_mDIV_A7-55_082108	Male	96.60	Wild-derived laboratory strain	SKIVE/EiJ
SNP_mDIV_B4-191_082108	Female	96.69	Wild caught	BAG94
SNP_mDIV_B10-491_022709	Female	96.73	Wild caught	Yu2120f
SNP_mDIV_C10-405_012709	Male	96.87	Classical laboratory strain	NONcNZO10/LtJ
SNP_mDIV_B6-188_103008_3	Female	96.94	Wild caught	BAG102
SNP_mDIV_B2-446_012209	Male	96.96	Wild-derived laboratory strain	MPB

[a] overall call rate refers to the total number of AA, AB, and BB genotypes determined for each sample using all SNP probes included in the 523,322 SNP list used for genotyping.

3.4 There is little variation between genetic distance measures calculated for the original and filtered SNP lists

Genetic distance measures for the autosomal post-genotyping filtered SNP list are similar to those calculated for the autosomal post-genotyping original SNP list (Table 3-8). However, using the filtered SNP list, genetic distance values increase or decrease depending on their sample type. For wild caught mice the minimum genetic distance measure remained the same; however the maximum and average genetic distance measures increased. The minimum, maximum, and average genetic distance measures for collaborative cross mice and F1 hybrid mice increased. The genetic distance range for wild-derived laboratory mice broadens, with the minimum value decreasing and the maximum value increasing. The average genetic distance value for wild-derived laboratory strains also increased. The minimum and maximum genetic distance measures, as well as the average genetic distance value for both consomic and BXD mice decreased. The genetic distance of the congenic mouse sample also increased. The range of genetic distance values for classical laboratory strains decreased for the minimum genetic distance value. The maximum genetic distance value remained the same for classical laboratory strains; however, the average genetic distance value decreased. Genetic distances for both the collaborative cross mice and F1 hybrid mice were significantly different between the original and filtered SNP lists ($p<0.01$, $p<0.01$ respectively; Mann-Whitney U Test). The remaining sample types showed no significant difference between genetic distance values calculated with the original SNP list and those calculated with the filtered SNP list.

Table 3-8: Genetic distance measures determined for samples compared to a homozygous A reference using autosomal post-genotyping lists

Sample Type[a]	**Original SNP list**[b]				**Filtered SNP list**[cd]			
	Min[e]	**Max**[f]	**Mean**[g]	**SEM**[h]	**Min**[e]	**Max**[f]	**Mean**[g]	**SEM**[h]
Classical laboratory strains (114, 116)	0.010	0.226	0.157	0.007	0.000	0.226	0.153	0.007
congenic (1, 1)			0.237				0.241	
consomic (10, 10)	0.011	0.237	0.174	0.018	0.001	0.209	0.036	0.019
BXD (44, 44)	0.089	0.168	0.110	0.002	0.083	0.167	0.106	0.002
Wild-derived laboratory strains (42, 51)	0.136	0.521	0.339	0.017	0.133	0.529	0.354	0.017
F1 Hybrid (37, 55)[i]	0.190	0.587	0.462	0.022	0.191	0.646	0.495	0.019
CC-UNC G2:F1 (37, 40)[i]	0.363	0.456	0.411	0.003	0.374	0.473	0.428	0.003
Wild Caught (12, 19)	0.271	0.511	0.361	0.026	0.271	0.519	0.409	0.029

[a] Samples are ordered based on their potential to be divergent from the reference genome

[b] 521,841 SNPs used in genetic distance measure determinations

[c] 450,927 SNPs used in genetic distance measure determinations

[d] red values indicate a decrease in genetic distance towards 0 and blue values indicate an increase genetic distance towards1

[e] the minimum genetic distance values for each of the sample types

[f] the maximum genetic distance values for each of the sample types

[g] the average genetic distance values for each of the sample types

[h] standard error of the mean for each of the sample types

[i] significantly different distribution of genetic distance values between the original and filtered SNP lists (Mann-Whitney U test).

Genetic distance measures were then calculated for all samples that passed the first round of genotyping after all samples were regenotyped for both the original SNP list and the filtered SNP list to determine if there was a significant difference in phylogenetic trees (Figure 3-5; Figure 3-6). Only samples that appeared in both lists were compared and no significant difference was seen between the ordering of the nodes when genetic distance was calculated when NoCalls shared between samples were included as differences (Mantel score=0.9995; p<0.001; Mantel Test). When NoCalls shared between samples were excluded as genetic differences, the genetic distance values were also the same between the two genotyping lists (Mantel score=0.9995; p<0.001; Mantel Test) (Figure 3-7; Figure 3-8).

Figure 3-5: Phylogenetic tree for 301 samples created using genetic distance values determined with the autosomal post-genotyping original SNP list. The colours for each of the mouse samples indicate which sample type they are according to their mouse strain. NoCalls shared between samples were counted as differences when calculating the genetic distances used to create this tree.

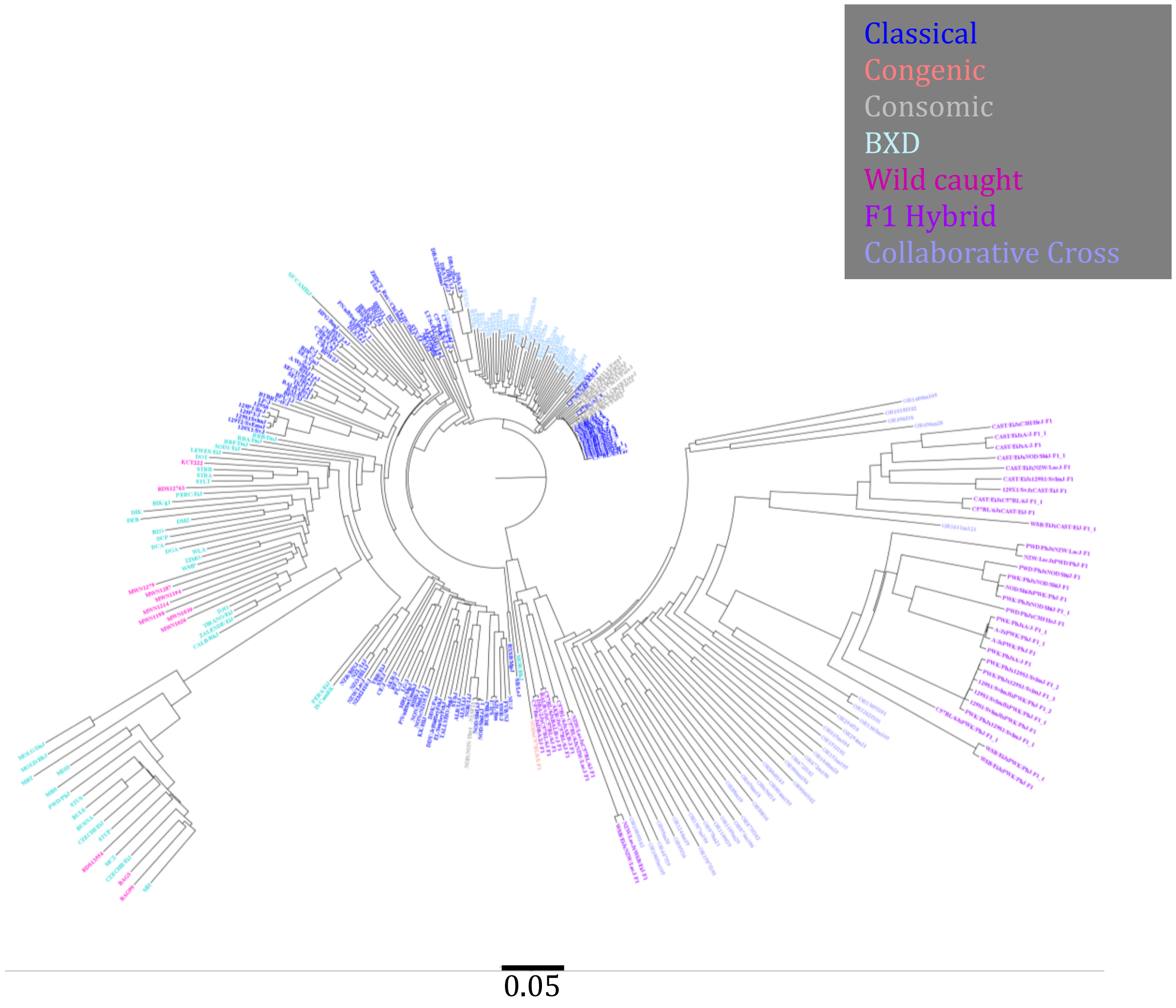

Classical
Congenic
Consomic
BXD
Wild caught
F1 Hybrid
Collaborative Cross
0.05

Figure 3-6: Phylogenetic tree for 336 samples created using genetic distance values determined with the autosomal post-genotyping filtered SNP list. The colours for each of the mouse samples indicate which sample type they are according to their mouse strain. NoCalls shared between samples were counted as differences when calculating the genetic distances used to create this tree.

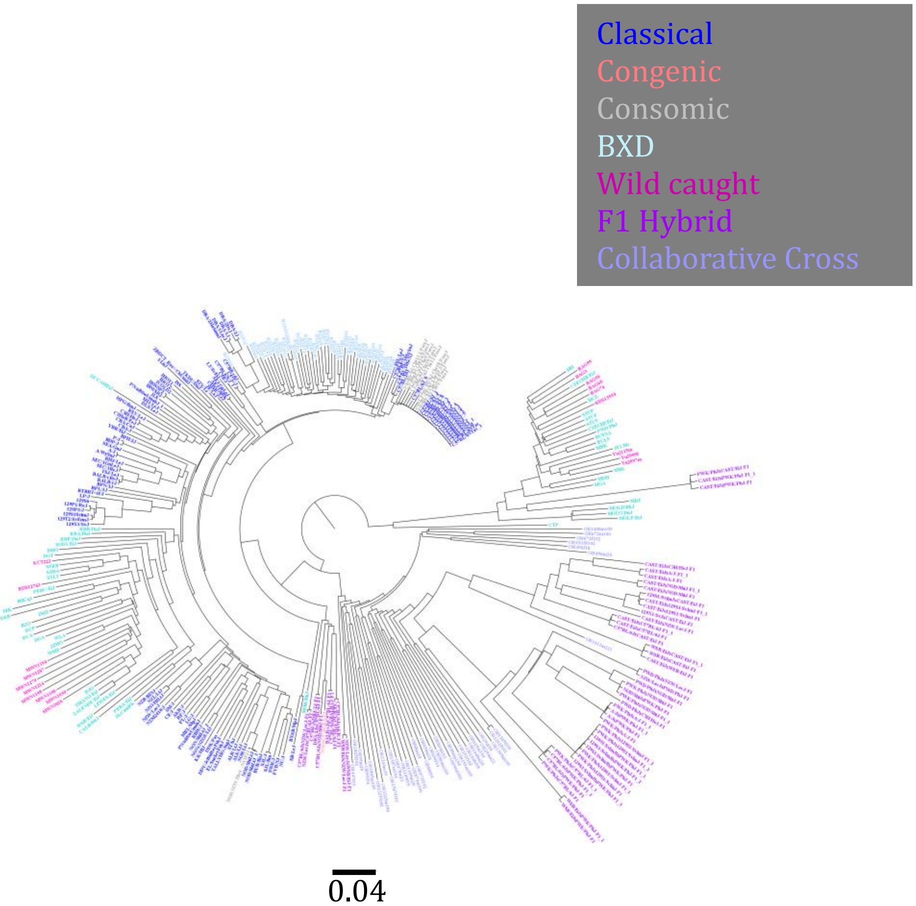

Classical
Congenic
Consomic
BXD
Wild caught
F1 Hybrid
Collaborative Cross
0.04

Figure 3-7: Phylogenetic tree for 301 samples created using genetic distance values determined with the autosomal post-genotyping original SNP list. The colours for each of the mouse samples indicate which sample type they are according to their mouse strain. Shared NoCalls between samples were not counted as differences in the genetic distance measures used to create this tree.

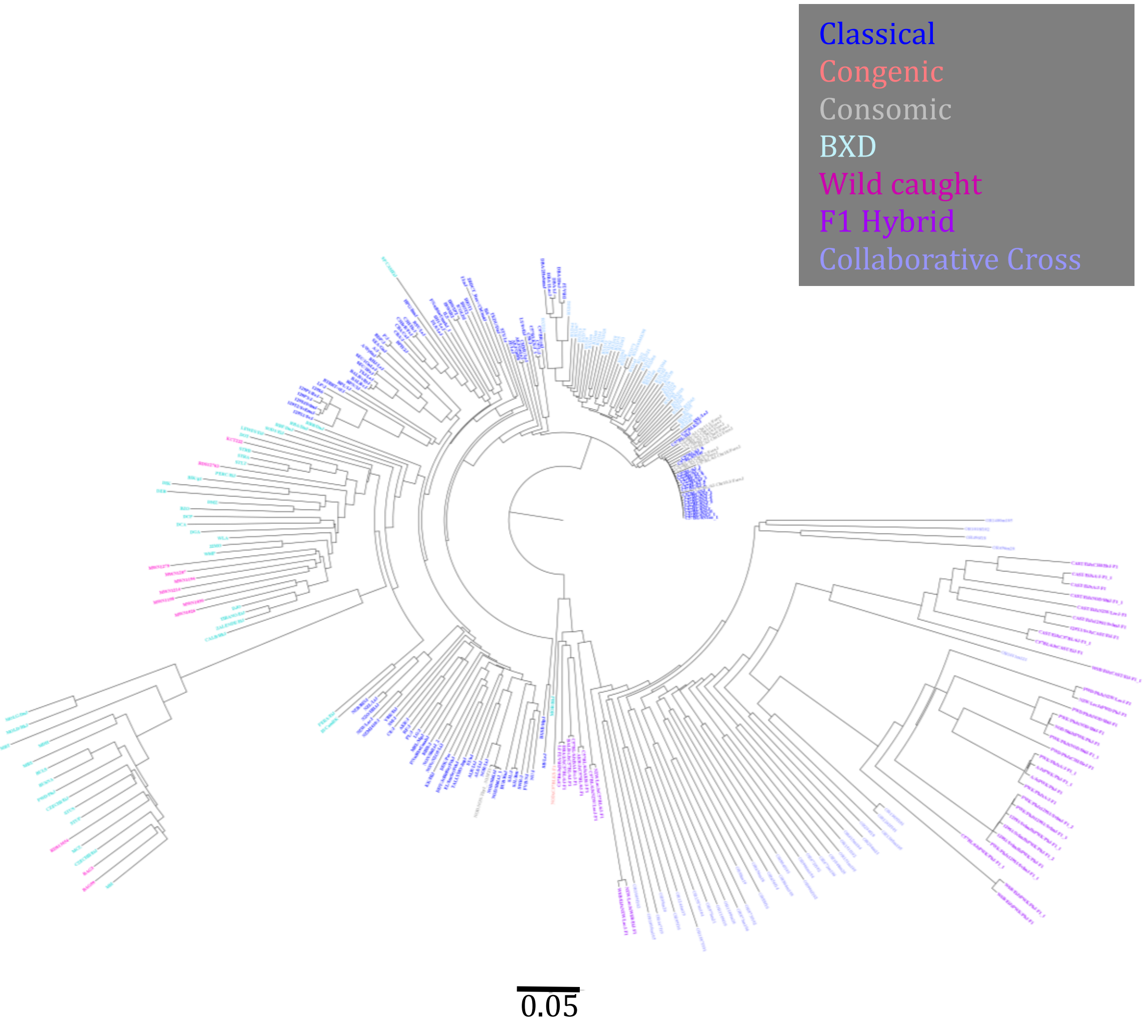
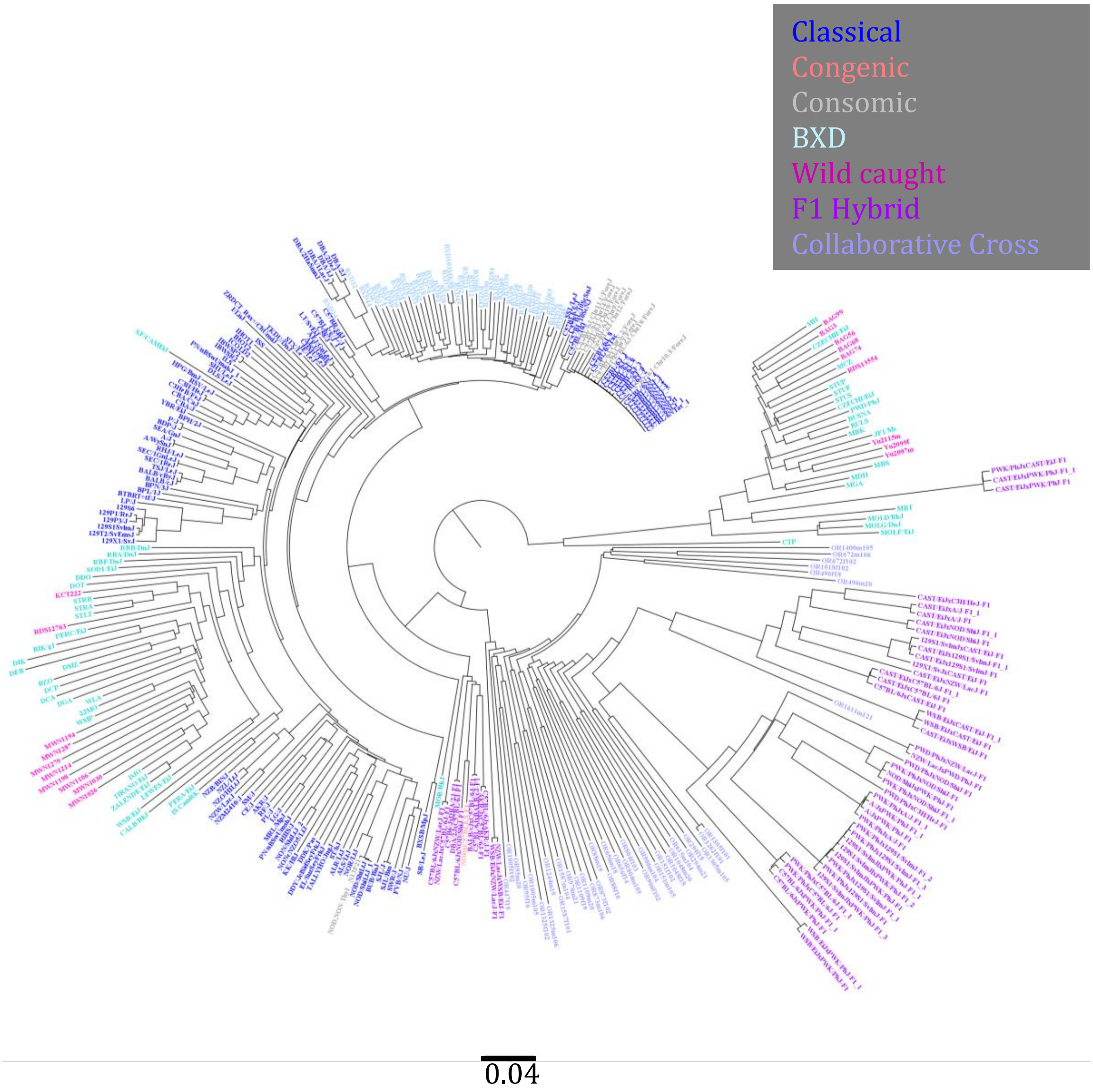

Classical
Congenic
Consomic
BXD
Wild caught
F1 Hybrid
Collaborative Cross
0.05

Figure 3-8: Phylogenetic tree for 336 samples created using genetic distance values determined with the autosomal post-genotyping filtered SNP list. The colours for each of the mouse samples indicate which sample type they are according to their mouse strain. Shared NoCalls between samples were not counted as differences in the genetic distance values used to create this tree.

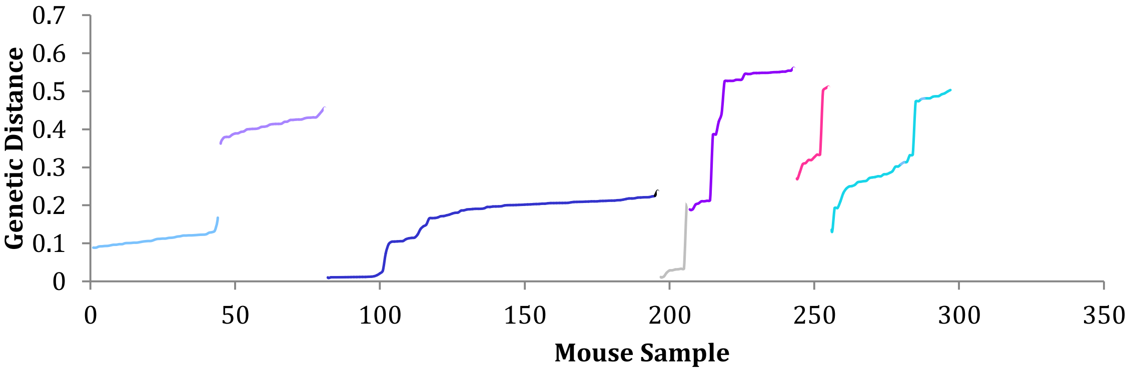

Classical
Congenic
Consomic
BXD
Wild caught
F1 Hybrid
Collaborative Cross
0.04

3.5 Genetic Distance

3.5.1 Genetic distance measures can distinguish between samples as compared to a reference genome homozygous for the A allele at all loci

After one round of genotyping, 15 samples were identified as having failed to pass the genotyping threshold of 97% (Table 3-7). The remaining 336 samples were subjected to a second round of genotyping with the filtered SNP list and the genotyping call rate for each sample is listed (Table C-2). Genetic distance was calculated for the 336 samples using the autosomal post-genotyping filtered SNP list (450,927 SNPs) to obtain a genetic distance measure for each of the samples compared to a reference genome homozygous for the A allele at all loci (Figure 3-9). The mouse strains in the 336 samples have been divided by sample type; wild caught, collaborative cross, F1 hybrid, wild-derived laboratory, BXD, consomic, congenic, and classical laboratory. Mice that are more genetically distant from the homozygous A reference have a genetic distance closer to 1, such as the wild-derived laboratory strains, the collaborative cross mice, and the wild caught mice. Mice that are more genetically similar to the reference AA genotype (modified from a B6 mouse) have a genetic distance closest to 0, such as the BXD and classical laboratory strains. The highest genetic distance was 0.64601 for the F1 hybrid of CAST/EiJxPWK/PhJ (*Mus musculus castaneus* x *Mus musculus musculus*). The smallest genetic distance determined using this method was a pure B6 (*Mus musculus domesticus)* mouse with a distance of 0.00049 from the reference sample.

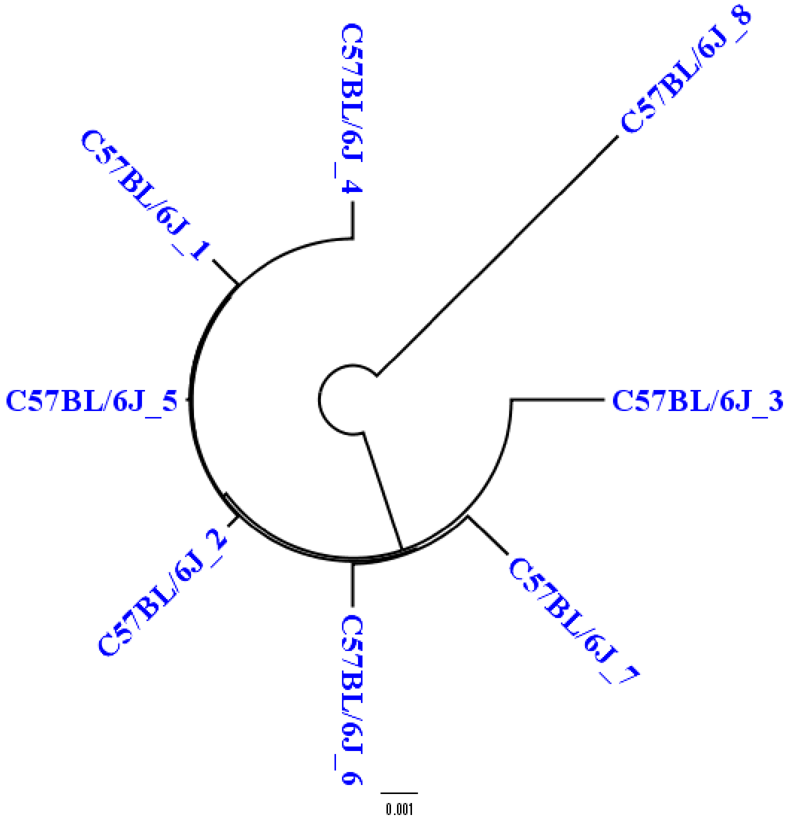
Figure 3-9: Genetic diversity of 336 samples from The Jackson Laboratory in comparison to a homozygous A reference. Colours in order from left to right across the graph represent BXD (light blue), collaborative cross (light purple), classical laboratory strains (dark blue), congenic (black), consomic (grey), F1 hybrid (dark purple), wild caught (pink), and wild-derived mouse strains (teal).


Genetic Distance
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
50
100
150
200
250
300
350
Mouse Sample

3.5.2 Genetic distance measures from pairwise comparisons can distinguish between distantly and closely related samples

Genetic distances were calculated to generate a genetic distance matrix for all 336 passing samples from The Jackson Laboratory database using the autosomal post-genotyping filtered SNP list (Table D-1). Samples were given a genetic distance value between 0 and 1. Samples with the highest genetic distance measures were from *Mus musculus musculus* and *Mus musculus castaneus* substrains and include wild-derived laboratory mice, CC, F1 hybrids, and wild caught mice. Samples that were most similar to the reference of all homozygous A genotypes had genetic distances closest to 0. Samples with genotypes closest to the reference include *Mus musculus domesticus* samples for classical laboratory strains, BXD strains, and consomic mice that are on a B6 genetic background with a single chromosome replacement from the PWD/PhJ-ForeJ mouse strain. Phylogenetic trees were created using this distance matrix and mouse strain types were colour coded to distinguish a pattern among the phylogeny based on strain type (Figure 3-8). Each strain type groups closely together within the phylogenetic tree. Genetic distance measures calculated for the eight B6 mice, could differentiate between each of the samples (Figure 3-10; Table D-2).

Figure 3-10: Genetic distance values calculated for the eight B6 samples using the autosomal post-genotyping filtered SNP list. Genetic distance values were calculated after genotyping with 336 samples. All samples are part of the classical laboratory strain set (blue).


C57BL/6J_8
C57BL/6J_4
C57BL/6J_1
C57BL/6J_5
C57BL/6J_3
C57BL/6J_2
C57BL/6J_6
C57BL/6J_7
0.001

3.5.3 Nearly all SNPs on the array can detect genetic variation

Of the 450,927 SNPs in the autosomal post-genotyping filtered SNP list that were used for genetic distance comparisons, a total of 448,174 SNPs were capable of detecting genetic variation between 336 samples. This is attributed to 99.4% of SNPs used in the analysis. A total of 2,753 SNPs did not detect any genotype differences between the 336 samples. The percentage of genetic variation detectable for each sample varies depending on the sample type (Table 3-9). B6 mice detect the least amount of genetic variation and wild-derived laboratory strains detect the most amount of genetic variation.

Table 3-9: Percentage of genetic variation detectable within each sample type using the Mouse Diversity Genotyping Array.

Sample Type[a]	Number of CEL files	% of SNPs that detect genetic variation
Classical Laboratory Strain	116	71.4
C57Bl/6J[b]	8	1.7
Congenic[c]	1	
Consomic	10	32.1
BXD	44	24.7
Wild-Derived Laboratory Strains	51	96.5
F1 Hybrid	55	91.9
CC-UNC G2:F1	40	87.7
Wild Caught	19	82.0

[a] Samples are ordered based on their potential to be divergent from the reference genome

[b] the samples included in C57BL/6J are part of the 116 samples in classical laboratory strains

[c] only contains one sample therefore could not detect the amount of genetic variation within group

3.5.4 Genetic variation can be detected among C57BL/6J mice

Overall, the eight B6 samples genotyped with a genetic background over 98% B6 (Table 3-10). The range of genetic distance values for B6 mice is between 0.0005 and 0.0103 (SEM 0.0011). A total of 7,793 SNPs detected genotype differences between the eight B6 mice. These differences include 2% of the genetic variation detectable on the array from the 448,174 informative SNPs that detect differences between 336 samples. When separated by sample, there was a significant difference in the total number of differences between the eight mice (average 1044 ± 504 genotype differences) ($p<0.001$; Fisher's Exact Test MCS) (Figure 3-11). When mouse eight is removed from analysis the average number of genetic differences is 557 ± 152. When looking at how differences accumulated across autosomes of the mouse genome, all eight mice were not significantly different from each other. The distribution of observed differences in genotype as an over- or under-representation of differences from the expected random distribution was compared by subtracting the expected (the number of differences in genotype calls calculated from the observed value in proportion to the number of SNPs on each chromosome) from the total number of differences observed on each chromosome. The distribution for all eight mice was not significantly different from a random distribution. The distribution of differences based on position along each of the chromosomes for all eight mice does not show any obvious regions that preferentially accumulate or lack differences for any of the mice (Figure 3-12). Of the differences detected between the eight mice, a total of 47 have two major genotype calls represented among the samples in ratios of 4:4, 4:3, and 5:3 genotype differences; where a minimum of three mice genotype with one call and three mice genotype with a second genotype call. The few differences between two major genotypes were divided among all but one of the mouse chromosomes, chromosome 16.

Table 3-10: Percentage of SNPs that genotyped as B6[a] for the eight B6 mice in The Jackson Laboratory dataset using the autosomal post-genotyping filtered SNP list.

Sample	B6 genotypes	non-B6 genotypes	% B6	Genetic distance value
B6 (1)	448124	2803	99.38	0.0015
B6 (2)	448386	2541	99.44	0.0009
B6 (3)	447340	3587	99.20	0.0033
B6 (4)	448143	2784	99.38	0.0014
B6 (5)	448555	2372	99.47	0.0005
B6 (6)	447853	3074	99.32	0.0017
B6 (7)	448017	2910	99.35	0.0021
B6 (8)	445230	5697	98.74	0.0103

[a] 450,927 SNPs from the autosomal post-genotyping filtered SNP list were capable of detecting a B6 genotype and were used to calculate the percentage of B6 genotypes for each sample

Figure 3-11: Differences attributed to each of the eight B6 mice from The Jackson Laboratory dataset. A) The total number of genotype differences associated with each of the eight samples (average 1044 ± 504 genotype differences). **B)** The distribution of observed differences in genotype graphed to show instances of over- and under-representation from a random occurrence based on the number of genotype calls per chromosome.

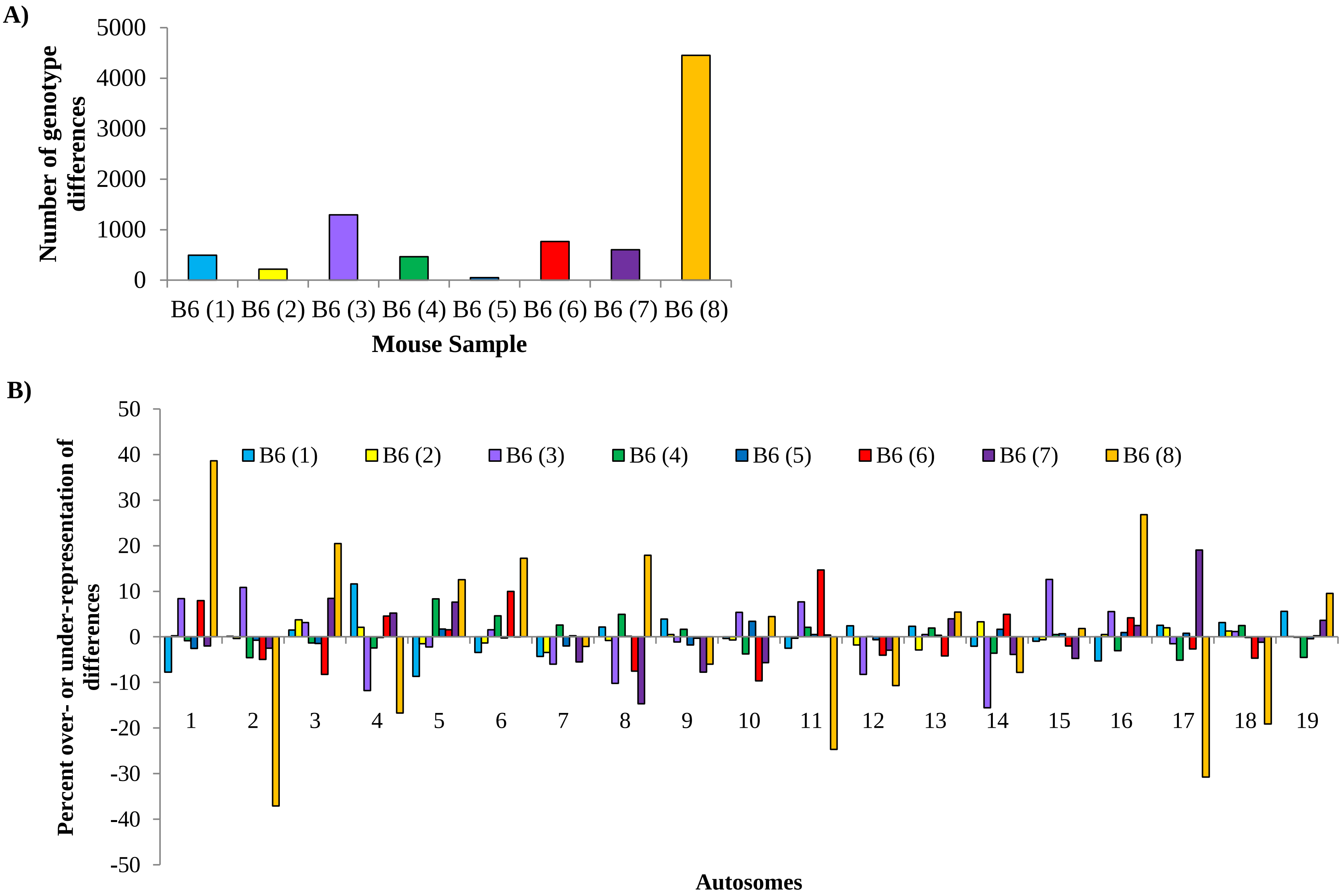
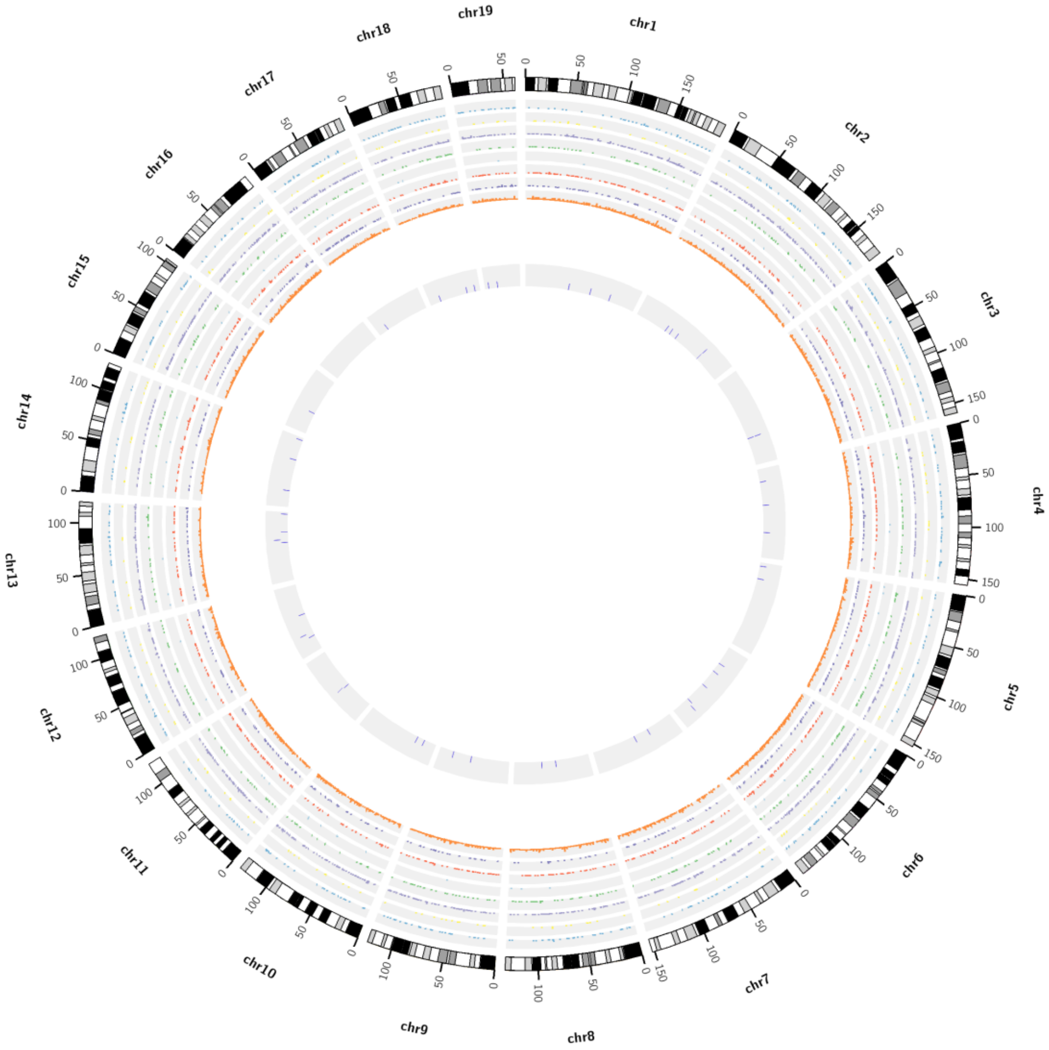

A)
Number of genotype differences
5000
4000
3000
2000
1000
0
B6 (1)
B6 (2)
B6 (3)
B6 (4)
B6 (5)
B6 (6)
B6 (7)
B6 (8)
Mouse Sample
B)
Percent over- or under-representation of differences
50
40
30
20
10
0
-10
-20
-30
-40
-50
B6 (1)
B6 (2)
B6 (3)
B6 (4)
B6 (5)
B6 (6)
B6 (7)
B6 (8)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Autosomes

Figure 3-12: The distribution of differences across the autosomes for each of the eight B6 mice from The Jackson Laboratory dataset. Each of the differences attributed to the eight mice is shown in its own track inside the ring of chromosomes. Mouse 1 in blue is closest to the karyotypes, and mouse 8 the second inner-most ring in orange. The remaining colours correspond in the following order, mouse 2 in yellow, mouse 3 in purple, mouse 4 in green, mouse 5 in light blue, mouse 6 in red, and mouse 7 in dark purple. The distribution of differences across the autosomes for the differences detected between the "majority" of B6 mice from The Jackson Laboratory dataset is the inner-most ring. The number of differences detected when B6 mice were divided evenly into two distinct genotype calls in ratios of (4:4, 4:3, and 5:3). The location of the SNP affected is plotted in purple on the inside track for each chromosome.


chr1
0
50
100
150
chr2
0
50
100
150
chr3
0
50
100
150
chr4
0
50
100
150
chr5
0
50
100
150
chr6
0
50
100
chr7
0
50
100
150
chr8
0
50
100
chr9
0
50
100
chr10
0
50
100
chr11
0
50
100
chr12
0
50
100
chr13
0
50
100
chr14
0
50
100
chr15
0
50
100
chr16
0
50
chr17
0
50
chr18
0
50
chr19
0
50

3.5.5 Genetic variation can be detected between wild caught mice

The total number of SNPs where differences were detected between 19 wild caught mice was 367,695 SNPs (Figure 3-13). These differences include 82% of the genetic variation detectable on the array from the 450,927 informative SNPs that detect differences between 336 samples. Of these 367,695 SNP differences, 6,671 SNP differences are shared between the wild caught mice and the eight B6 mice from the previous analysis (Figure 3-14). Of the 47 SNPs that were represented in the majority of samples genotyping differently, 41 of the SNPs were also present in the 6,671 SNPs detecting differences in both the B6 and wild caught mice. Of the 41 SNPs where differences occur between both wild caught and the eight B6 mice, 14 are located in regions affecting genes (Table 3-11).

3.6 *Apoptosis-inducing factor* genotyping

Mice that genotyped as WT were 911.17, 911.49, 911.50, 300.6, and 300.7. Mice that genotyped with the *hq* mutation were 904.9, 904.11, 911.143, and 911.148. Mouse 900.3 genotyped as a *hq* carrier.

3.7 DNA quality for DNA extractions

The concentration, 260/280 ratio, and 260/230 ratio for each of the 33 samples from the Hill laboratory are shown in Table 3-12. The DNA extracted for the 33 samples was of high molecular weight (Figure 3-15).

3.8 Samples failed genotyping because of poor DNA quality

A total of six samples from the Hill laboratory failed genotyping. DNA quality ratios indicate that the 260/280 ratios and 260/230 ratios were below the optimum 1.8 and 2.0, respectively (Table 3-13).

Figure 3-13: Total number of differences across each of the chromosomes detected between 19 wild caught mice. The total number of SNPs detected using the filtered SNP list is shown by the maximum height of the black bar. The total number of differences detected between the 19 wild caught mice is shown in grey bars with blue borders.

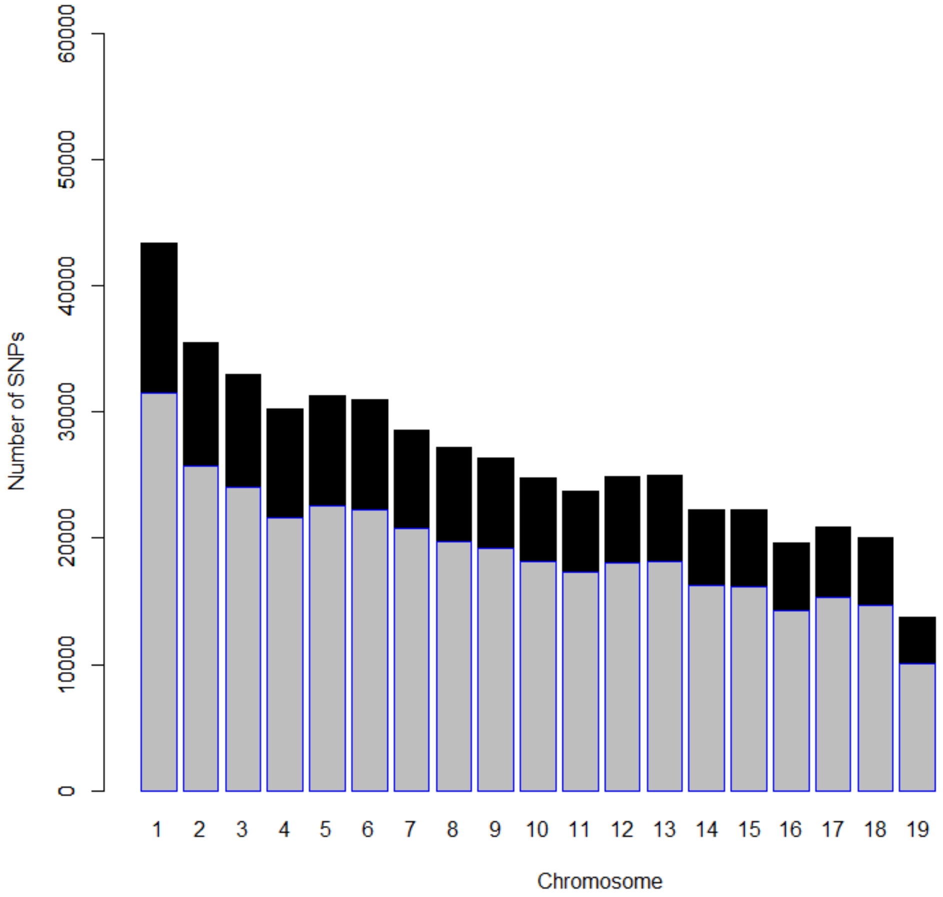

Number of SNPs
0
10000
20000
30000
40000
50000
60000
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Chromosome

Figure 3-14: The number of SNPs that detected differences between the eight B6 mice and between 19 wild caught mice from The Jackson Laboratory dataset. Differences identified between the B6 mice are indicated by the red circle. The purple circle indicates differences identified between the wild caught mice. The overlap between the circles is the number of SNPs in common between the two lists of differences. The numbers are associated with differences specific to the B6 mice, shared between the B6 and wild caught mice, and specific to the wild caught mice when moving from left to right across the figure.

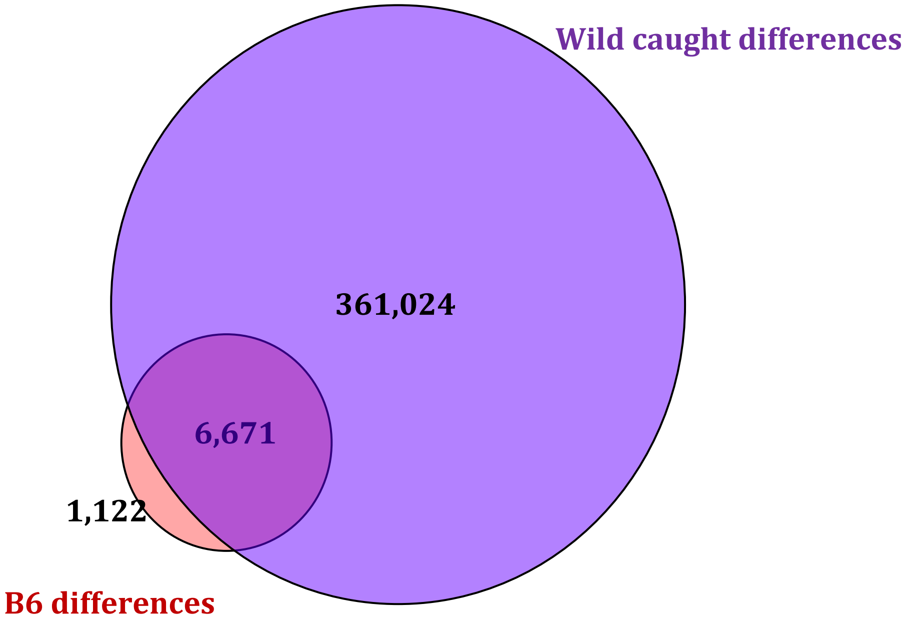

Wild caught differences
361,024
6,671
1,122
B6 differences

Table 3-11: SNPs affecting genes where differences occur in both B6 and wild caught mice

Chromosome	Chromosome position (bps)	Gene Name	Gene Symbol[a]
2	132520778	*RIKEN cDNA 1110034G24 gene*	*1110034G24Rik*
4	21780030	*Serine/arginine-rich splicing factor 18*	*Sfrs18*
5	28705949	*RNA binding motif protein 33*	*Rbm33*
6	32800650	*Coiled-coil-helix-coiled-coil-helix domain containing*	*Chchd3*
8	83952230	*Inositol polyphosphate-4-phosphatase, type II*	*Inpp4b*
9	67134475	*Talin 2*	*Tln2*
12	25588806	*Membrane bound O-acyltransferase domain containing 2*	*Mboat2*
12	67752382	*MAM domain containing glycosylphosphatidylinositol anchor 2*	*Mdga2*
13	71689784	*RIKEN cDNA 1700112M02 gene*	*AK007185*
14	120680516	*Muscleblind-like 2*	*Mbnl2*
14	28152097	*Rho guanine nucleotide exchange factor (GEF) 3*	*Arhgef3*
18	67392407	*Metallophosphoesterase 1*	*Mppe1*
19	6009147	*Calpain 1*	*Capn1*
19	21907126	*Transmembrane protein 2*	*Tmem2*

[a] Gene Symbol was taken from MGI

Table 3-12: Quality control data for DNA extractions

Mouse ID	Tissue	DNA Concentration (μg/μl)	Total DNA Yield (μg)	260/280 ratio	260/230 ratio	Electrophoretic assessment[a]	DNA Extraction location
904.9	Sp	1.35	134.6	1.85	2.16	√	JAX
904.11	Sp	1.69	169.4	1.82	2.23	√	JAX
911.49	Sp	1.84	184.3	1.84	2.23	√	JAX
911.50	Sp	1.53	153.4	1.83	2.28	√	JAX
904.9	Cl	0.27	13.6	1.74	1.31	√	JAX
904.11	Cl	0.37	18.46	1.76	1.34	√	JAX
911.49	Cl	0.28	13.81	1.73	1.27	√	JAX
911.50	Cl	0.30	14.97	1.76	1.33	√	JAX
911.143	Cl	0.19	9.27	1.73	1.16	√	JAX
911.148	Cl	0.19	9.46	1.72	1.13	√	JAX
300.6	Cl	0.10	9.96	1.81	1.43	√	LRGC
300.6 [b]	Sp	0.26	26.13	1.9	2.44	√	LRGC
900.3	Li	0.22	22.12	1.82	1.44	√	LRGC
900.3	Sp	0.48	48.00	1.87	2.42	√	LRGC
911.50	Cl-2	0.10	9.73	1.88	2.34	√	LRGC
911.50	Sp-2	0.25	24.83	1.88	2.33	√	LRGC
904.9[b]	Li	0.93	928.60	1.72	1.3	√	LRGC
904.11[b]	Li	1.97	1967.07	1.75	1.39	√	LRGC
904.12[b]	Li	1.28	1279.87	1.75	1.4	√	LRGC
911.49[b]	Li	2.82	2822.40	1.75	1.65	√	LRGC
911.47[b]	Li	1.93	1934.30	1.74	1.39	√	LRGC
911.50	Li	1.15	1151.46	1.8	1.93	√	LRGC
300.7	Cl-2	0.10	9.76	1.84	2.17	√	Hill Laboratory
300.7	Cl-3	0.10	9.60	1.81	1.96	√	Hill Laboratory
300.7	Sp-1	0.10	9.79	1.89	2.26	√	Hill Laboratory
300.7	Sp-2	0.17	17.41	1.88	2.24	√	Hill Laboratory
300.7	Sp-3	0.09	9.17	1.86	2.44	√	Hill Laboratory
911.17	Sp	0.10	10.07	1.9	2.23	√	Hill Laboratory

Mouse ID	**Tissue**	**DNA Concentration (μg/μl)**	**Total DNA Yield (μg)**	**260/280 ratio**	**260/230 ratio**	**Electrophoretic assessment[a]**	**DNA Extraction location**
911.17	Cl	0.08	8.44	1.86	2.24	√	Hill Laboratory
911.17	Li	0.25	24.68	1.87	1.94	√	Hill Laboratory
911.49	Li	0.10	9.72	1.88	2.13	√	Hill Laboratory
911.50	Li-2	0.19	19.462	1.89	2.11	√	Hill Laboratory

[a] indicates whether samples were electrophoresed through an agarose gel to determine if the DNA extraction was high molecular weight

[b] samples that failed to pass the first round of genotyping threshold of 97% overall genotype call

Figure 3-15: 1.5% agarose gel of high molecular weight DNA extractions. This gel shows the quality for samples isolated in the Hill Laboratory. This is a representation of all extractions for DNA quality. Lane 1 contains a 1 kb ladder. Lanes 2-4 contain DNA isolated from samples 300.7 Cl-1, Cl-2, and Cl-3 respectively. Lanes 5-7 contain DNA isolated from samples 300.7 Sp-1, Sp-2, and Sp-3 respectively. Lane 8-10 contain DNA isolated from 911.17 Sp, Cl, and Li respectively. Lane 11 contains 911.49 Li and lane 12 contains 911.50 Li-2. Samples were stained with SYBR Safe for visualization.

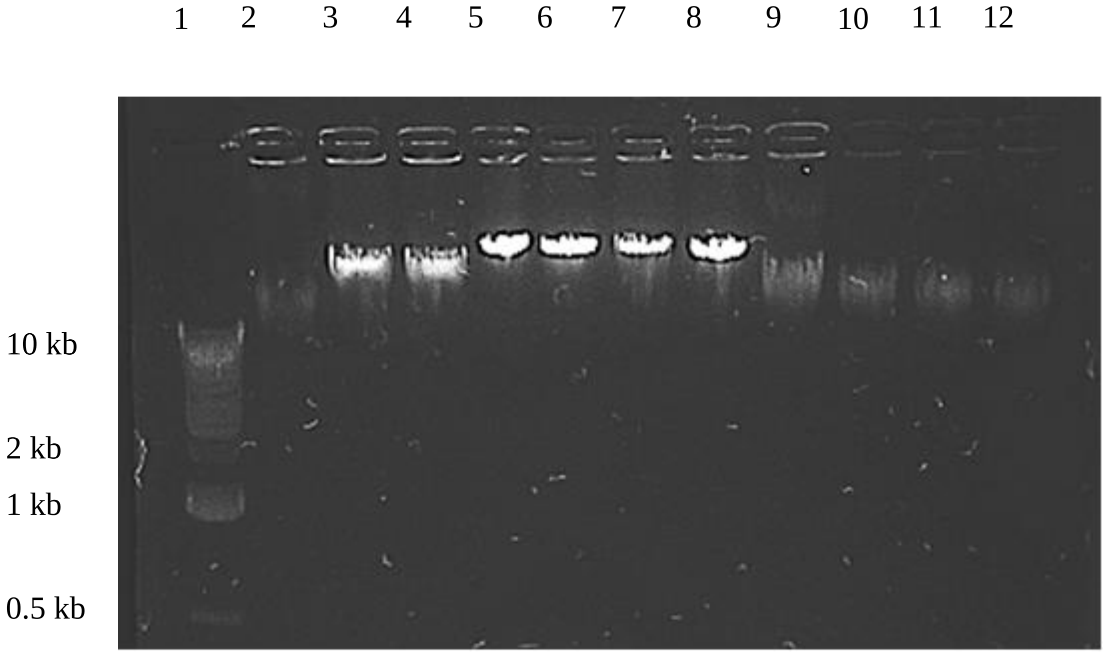

1
2
3
4
5
6
7
8
9
10
11
12
10 kb
2 kb
1 kb
0.5 kb

Table 3-13: DNA quality for failing Hill laboratory samples and their genotyping call rates

Mouse ID	Tissue[a]	260/280 ratio	260/230 ratio	Genotyping call rate
300.6	Sp	1.81	1.43	96.34
904.9	Li	1.72	1.30	84.75
904.11	Li	1.75	1.39	94.80
904.12	Li	1.75	1.40	78.23
911.49	Li	1.75	1.65	95.64
911.47	Li	1.74	1.39	95.00

[a] Tissue from which the DNA was extracted. Tissue types include the spleen (Sp) and liver (Li).

3.9 Genetic Background

3.9.1 Genetic distance measures can distinguish between samples when determined from a homozygous A reference

A total of 384 samples were genotyped in round one of genotyping using the filtered SNP list. Of these samples, 22 failed genotyping (Table 3-14). Sixteen of these 22 samples were from The Jackson Laboratory database, and six samples were from the Hill laboratory. Genotype results from the remaining 362 passing samples were obtained after a second round of genotyping using the filtered SNP list (Table C-2).

3.9.2 Genetic background comparisons can distinguish between samples that range in genetic background from pure C57BL/6J to pure CBA/CaJ

The percentage of B6, CBA/CaJ, and heterozygous genotypes was compared between the eight B6 mice from The Jackson Laboratory database, 27 passing samples from the Hill laboratory, and the CBA/CaJ sample from The Jackson Laboratory (Table 3-15). Samples can be distinguished based on the percentage of B6 genotype calls. Pure B6 samples genotype with the highest percentage of B6 genotypes. Pure CBA/CaJ samples genotyped with the highest percentage of CBA/CaJ genotypes calls. Samples that contain high levels of heterozygosity (50% B6:50% CBA/CaJ) genotyped with the highest number of heterozygous genotype calls. Overall, the B6 genotype decreases as the genetic background of a mouse becomes more CBA/CaJ. Mice will never genotype 100% CBA/CaJ because a B6 mouse and a CBA/CaJ mouse genotype the same for 369,597 SNPs. Multiple tissue samples from the same mice tend to group together.

Table 3-14: Samples that failed 1st round genotyping out of the 384 provided samples using the filtered SNP list

MDGA CEL file identifier	Mouse Gender	Overall call rate[a]	Sample Type	Mouse Strain
DNA3255[b]	Male	78.23	Classical laboratory strain	C57BL/6J/ CBA/CaJ
DNA3253[b]	Male	84.75	Classical laboratory strain	C57BL/6J/ CBA/CaJ
SNP_mDIV_A9-9_081308	Male	92.45	Wild-derived laboratory strain	CAST/EiJ
DNA3254[b]	Male	94.80	Classical laboratory strain	C57BL/6J/ CBA/CaJ
DNA3257[b]	Male	95.00	Classical laboratory strain	C57BL/6J/ CBA/CaJ
SNP_mDIV_D4-470_012209	Male	95.10	Wild-derived laboratory strain	MDGI
SNP_mDIV_B6-390_012709	Male	95.37	Classical laboratory strain	A/HeJ
DNA3256[b]	Male	95.64	Classical laboratory strain	C57BL/6J/ CBA/CaJ
SNP_mDIV_C10-121_090908	Male	95.85	Classical laboratory strain	ATEB/LeJ
SNP_mDIV_C11-35_081308	Male	96.19	Wild-derived laboratory strain	PWK/PhJ
SNP_mDIV_D6-133_090908	Male	96.20	Classical laboratory strain	MA/MyJ
SNP_mDIV_A6-54_082108	Male	96.25	Wild-derived laboratory strain	MSM/Ms
SNP_mDIV_B6-450_012209	Male	96.32	Wild-derived laboratory strain	PWK hybrid
DNA3159[b]	Male	96.34	Classical laboratory strain	C57BL/6J/CBA/CaJ
SNP_mDIV_A4-157_091708	Female	96.46	Wild caught	RDS10105
SNP_mDIV_A7-55_082108	Male	96.52	Wild-derived laboratory strain	SKIVE/EiJ
SNP_mDIV_B4-191_082108	Female	96.62	Wild caught	BAG94
SNP_mDIV_B10-491_022709	Female	96.64	Wild caught	Yu2120f
SNP_mDIV_C10-405_012709	Male	96.78	Classical laboratory strain	NONcNZO10/LtJ
SNP_mDIV_B6-188_103008_3	Female	96.86	Wild caught	BAG102
SNP_mDIV_B2-446_012209	Male	96.90	Wild-derived laboratory strain	MPB
SNP_mDIV_C9-464_012209	Male	96.97	Wild-derived laboratory strain	DDO

[a] overall call rate refers to the total number of genotypes determined as AA, AB, or BB for each sample

[b] are Hill laboratory samples

Table 3-15: Percentage of B6 and CBA genetic background[a] for all of the Hill laboratory plus B6 and CBA/CaJ samples from The Jackson Laboratory dataset[b]

Sample ID and tissue	Expected %B6 genotypes	% B6 genotypes	% CBA genotypes	% Heterozygous genotypes	% other
B6 (5)	100	99.49	0.39	0.00	0.13
B6 (2)	100	99.44	0.38	0.00	0.18
B6 (1)	100	99.40	0.38	0.00	0.21
B6 (7)	100	99.37	0.38	0.00	0.24
B6 (4)	100	99.37	0.38	0.01	0.24
300.7 Cl-3	100	99.37	0.40	0.01	0.22
B6 (6)	100	99.34	0.38	0.00	0.27
300.7 Sp-2	100	99.30	0.40	0.01	0.29
300.7 Sp-3	100	99.29	0.40	0.01	0.30
300.7 Sp-1	100	99.18	0.40	0.01	0.40
B6 (3)	100	99.18	0.38	0.01	0.43
300.7 Cl-1	100	99.14	0.40	0.01	0.45
300.7 Cl-2	100	99.13	0.40	0.02	0.45
B6 (8)	100	98.77	0.21	0.02	1.00
300.6 Cl	100	98.65	0.39	0.04	0.92
904.11 Cl	75	87.67	0.41	11.70	0.21
904.11 Sp	75	87.67	0.41	11.72	0.20
904.9 Sp	75	86.68	0.42	12.69	0.21
904.9 Cl	75	86.66	0.42	12.71	0.21
900.3 Sp	50	82.09	6.62	10.21	1.08
900.3 Li	50	82.06	6.59	10.19	1.16
911.143 Cl	7	79.67	13.14	7.01	0.18
911.148 Cl	7	79.66	12.97	7.19	0.18
911.50 Sp	7	79.25	14.36	6.23	0.16
911.49 Cl	7	79.24	12.93	7.63	0.20
911.49 Sp	7	79.24	12.93	7.64	0.19
CBA/CaJ	0	79.21	20.60	0.00	0.19
911.50 Cl	7	79.18	14.32	6.23	0.27
911.50 Sp-2	7	79.14	14.32	6.20	0.34
911.17 Cl	7	79.11	15.04	5.50	0.34
911.50 Cl-2	7	79.03	14.28	6.21	0.48
911.50 Li-2	7	79.01	14.29	6.21	0.48
911.17 Sp	7	79.01	15.00	5.48	0.51
911.17 Li	7	78.85	14.99	5.48	0.68
911.49 Li	7	78.56	12.80	7.40	1.24
911.50 Li	7	77.83	14.04	6.10	2.03

[a] Samples are ordered from the highest to the lowest percentage of B6 genotypes

[b] Samples highlighted in red are the pure B6 and CBA/CaJ samples from The Jackson Laboratory dataset

Comparison of each sample to a reference of all homozygous A genotypes determined the genetic distance for each of the 362 samples (Figure 3-16). Genetic distance values were compared between the eight B6 mice from The Jackson Laboratory database, 27 passing samples from the Hill laboratory, and the CBA/CaJ sample from The Jackson Laboratory (Table 3-16). Genetic distance values could be used to distinguish between mice that were pure B6 to pure CBA/CaJ genetic backgrounds. Ordering of values showed mice mostly B6 in genetic background had a genetic distance closest to 0 and mice more CBA/CaJ had a genetic distance further from 0. Samples that were half B6 and half CBA/CaJ had genetic distance values falling in the middle of samples that were primarily B6 and those primarily CBA/CaJ. When two tissue samples were available for the mouse, both tissue samples grouped together. Mice with three tissue samples did not have all three tissues group together. Liver samples consistently showed the higher genetic distance measure when there are multiple tissue types available for a single mouse.

Pairwise comparisons were used to calculate genetic distance measures between all 362 samples (Figure 3-17) (Table D-3, Table D-4). These genetic distance measures were used to generate a phylogenetic tree. Genetic distance measures between the eight B6 mice and one CBA/CaJ mouse from The Jackson Laboratory, and the passing 27 samples from the Hill laboratory can distinguish between a spectrum of B6 to CBA/CaJ genotypes (Figure 3-18). Tissue samples from the same mouse group together. Mice group together in the phylogenetic tree as expected based on breeding relationships.

Figure 3-16: Genetic distance from pairwise comparisons between 362 Mouse Diversity Genotyping Array samples. Colours in order from left to right across the graph represent BXD (light blue), collaborative cross (light purple), classical laboratory strains (dark blue), congenic (black), consomic (grey), F1 hybrid (dark purple), wild caught (pink), wild-derived mouse strains (teal), and Hill samples (mint).

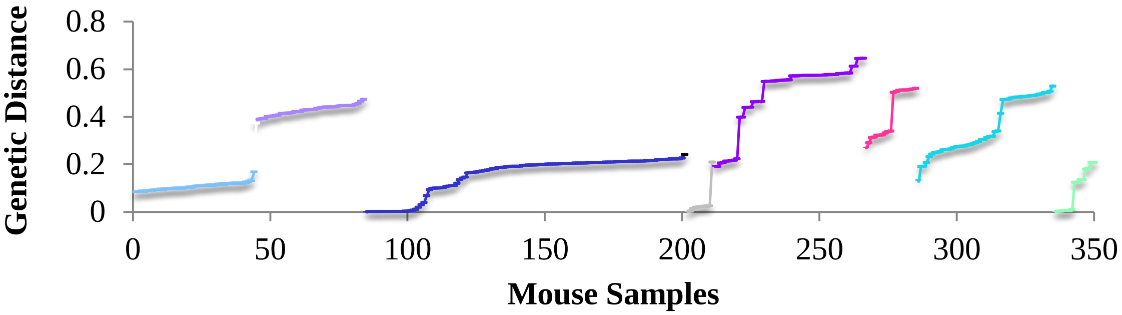

Genetic Distance
0.8
0.6
0.4
0.2
0
0
50
100
150
200
250
300
350
Mouse Samples

Table 3-16: Genetic distance measures[a] for the 27 Hill laboratory samples, eight B6 samples, and one CBA/CaJ sample from The Jackson Laboratory dataset.

Sample Identifier	Genetic Distance	Mouse Strain[b]
B6 (5)	0.0005	B6
B6 (2)	0.0008	B6
B6 (1)	0.0014	B6
B6 (4)	0.0015	B6
300.7 Cl-3	0.0017	B6
B6 (7)	0.0017	B6
B6 (6)	0.0021	B6
300.7 Sp-2	0.0024	B6
300.7 Sp-3	0.0025	B6
B6 (3)	0.0032	B6
300.7 Sp-1	0.0036	B6
300.7 Cl-1	0.0041	B6
300.7 Cl-2	0.0042	B6
300.6 Cl	0.0093	B6
B6 (8)	0.0101	B6
904.11 Cl	0.1239	<B6
904.11 Sp	0.1239	<B6
904.9 Sp	0.1342	<B6
904.9 Cl	0.1344	<B6
900.3 Sp	0.1793	B6/CBA
900.3 Li	0.1813	B6/CBA
911.50 Sp	0.2071	<CBA
911.143 Cl	0.2071	<CBA
911.148 Cl	0.2071	<CBA
911.49 Cl	0.2071	<CBA
911.49 Sp	0.2072	<CBA
CBA/CaJ	0.2075	CBA
911.50 Cl	0.2078	<CBA
911.50 Sp-2	0.2082	<CBA
911.17 Cl	0.2085	<CBA
911.50 Cl-2	0.2094	<CBA
911.50 Li-2	0.2096	<CBA
911.17 Sp	0.2096	<CBA
911.17 Li	0.2112	<CBA
911.49 Li	0.2142	<CBA
911.50 Li	0.2218	<CBA

[a] Samples are ordered from smallest to highest genetic distance

[b] Pure B6 and CBA samples are indicated in red. Samples more B6 than CBA are in blue. Samples half B6 and half CBA are in black. Samples more CBA than B6 are in purple.

Figure 3-17: Phylogenetic tree for 362 samples created using genetic distance values determined with the autosomal post-genotyping filtered SNP list. The colours for each of the mouse samples indicate which sample type they are according to their mouse strain (see legend).

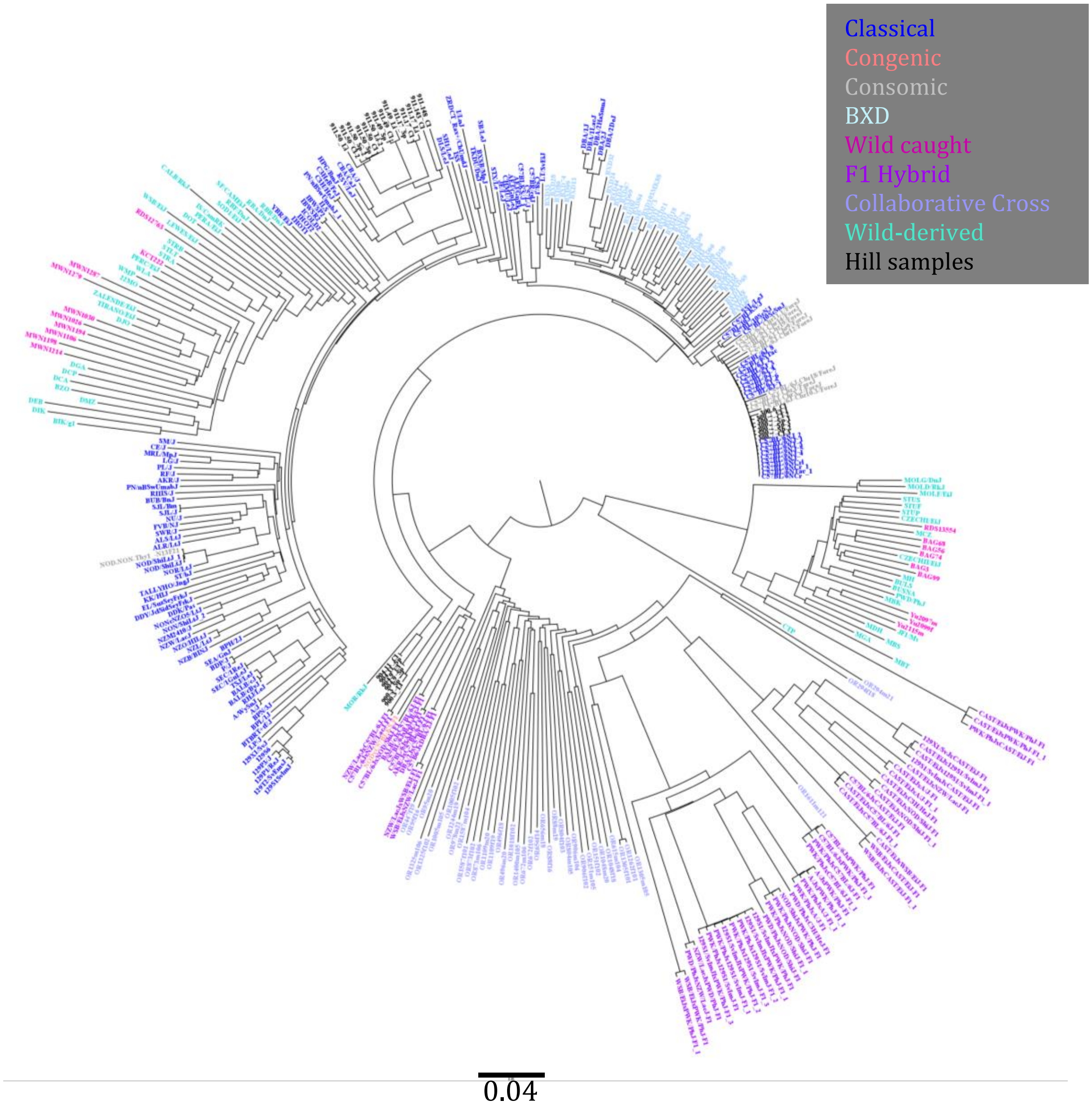
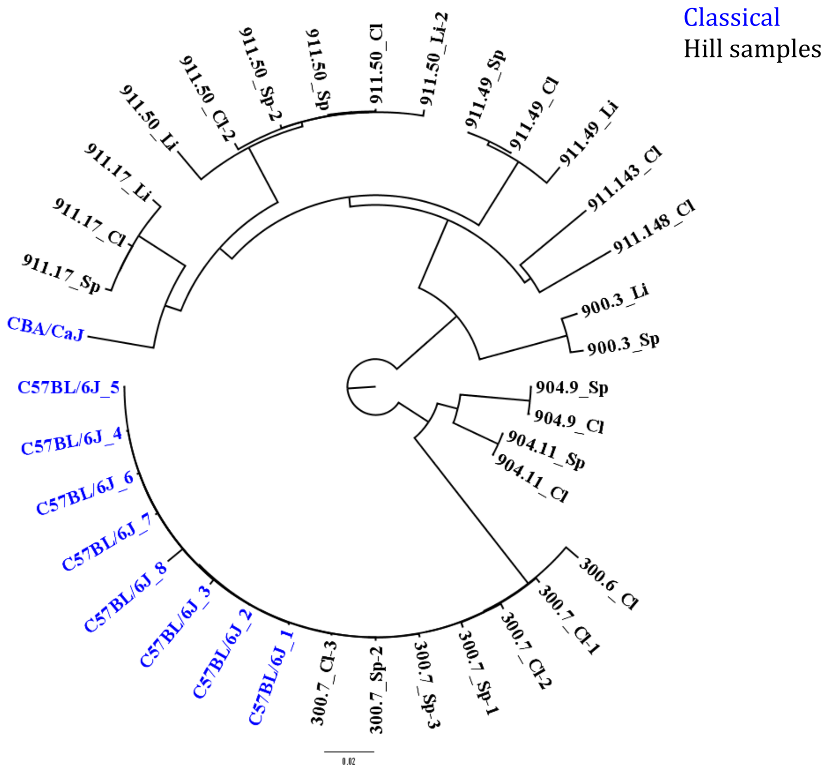

Classical
Congenic
Consomic
BXD
Wild caught
F1 Hybrid
Collaborative Cross
Wild-derived
Hill samples
0.04

Figure 3-18: Genetic distance measures for samples ranging from 100% B6 genetic background to 100% CBA/CaJ genetic background. Samples from the Hill laboratory are shown in black and those from The Jackson Laboratory dataset are classical laboratory strains in blue. The genetic distance values were determined using the autosomal post-genotyping filtered SNP list generated from genotyping of the 362 samples.


Classical
Hill samples
911.50_Li-2
911.50_Cl
911.50_Sp
911.50_Sp-2
911.50_Cl-2
911.50_Li
911.17_Li
911.17_Cl
911.17_Sp
CBA/CaJ
911.49_Sp
911.49_Cl
911.49_Li
911.143_Cl
911.148_Cl
900.3_Li
900.3_Sp
904.9_Sp
904.9_Cl
904.11_Sp
904.11_Cl
300.6_Cl
300.7_Cl-1
300.7_Cl-2
300.7_Sp-1
300.7_Sp-3
300.7_Sp-2
300.7_Cl-3
C57BL/6J_1
C57BL/6J_2
C57BL/6J_3
C57BL/6J_8
C57BL/6J_7
C57BL/6J_6
C57BL/6J_4
C57BL/6J_5
0.02

3.9.3 Tissue replicates for the spleen and the cerebellum of a C57BL/6J mouse show genotype differences within a tissue

The average number of genotype differences identified between replicates of the spleen was 959 ± 158 standard error of the mean (SEM) (Figure 3-19). Of the differences identified within each of the three replicates, a total of 0.60% of SNPs (2,780 SNPs) detected at least one genotype differences between the three tissues. The distribution of these genotype differences across the autosomes and the X chromosome was significantly different from random ($p<0.001$ for spleen, $p<0.01$ for cerebellum; Fisher's Exact Test MCS) and all three samples showed an under-representation of genotype differences on the X chromosome. When the X chromosome was removed from analysis, all samples had a random distribution of genotype differences across the autosomes.

The average number of genotype differences detected between three replicates of the cerebellum was 1,150 ± 287 (SEM). A total of 0.70% of SNPs (3,288 SNPs) detected at least one genotype difference between the three cerebellum replicates. The distribution of differences across the autosomes and the X chromosome was not significantly different between replicates. Two of the tissue replicates had a distribution of genotype differences that was significantly different from a random distribution ($p<0.001$, $p<0.01$; Fisher's Exact Test MCS). All replicates of the cerebellum showed an under-representation of genotype differences on the X chromosome. When the X chromosome was removed, all samples had a random distribution across the autosomes.

Of the probes that detected genotype differences between spleen replicates and between cerebellum replicates, 622 (0.16 % of genotypes) detected differences in both the spleen and cerebellum. The differences for all tissue replicates of the spleen and cerebellum appear to be uniformly distributed along each of the chromosomes (Figure 3-20).

Figure 3-19: The total number of genotype differences within each B6 spleen replicate and cerebellum replicate, and the over- or under-representation of differences across the genome. 300.7 is the specific mouse identifier. Sp refers to the spleen, and the replicate number is indicated after the dash. **A)** The total number of differences specific to each of the spleen replicates (959 ± 158 genotype differences). **B)** The total number of differences specific to each of the cerebellum replicates (1,150 ± 287 genotype differences). **C)** The distribution of observed differences in genotype graphed to show instances of over- and under-representation from a random occurrence based on the number of genotype calls per chromosome. * indicates samples with a distribution significantly different from the expected random distribution (p<0.005; Fisher's Exact Test MCS).

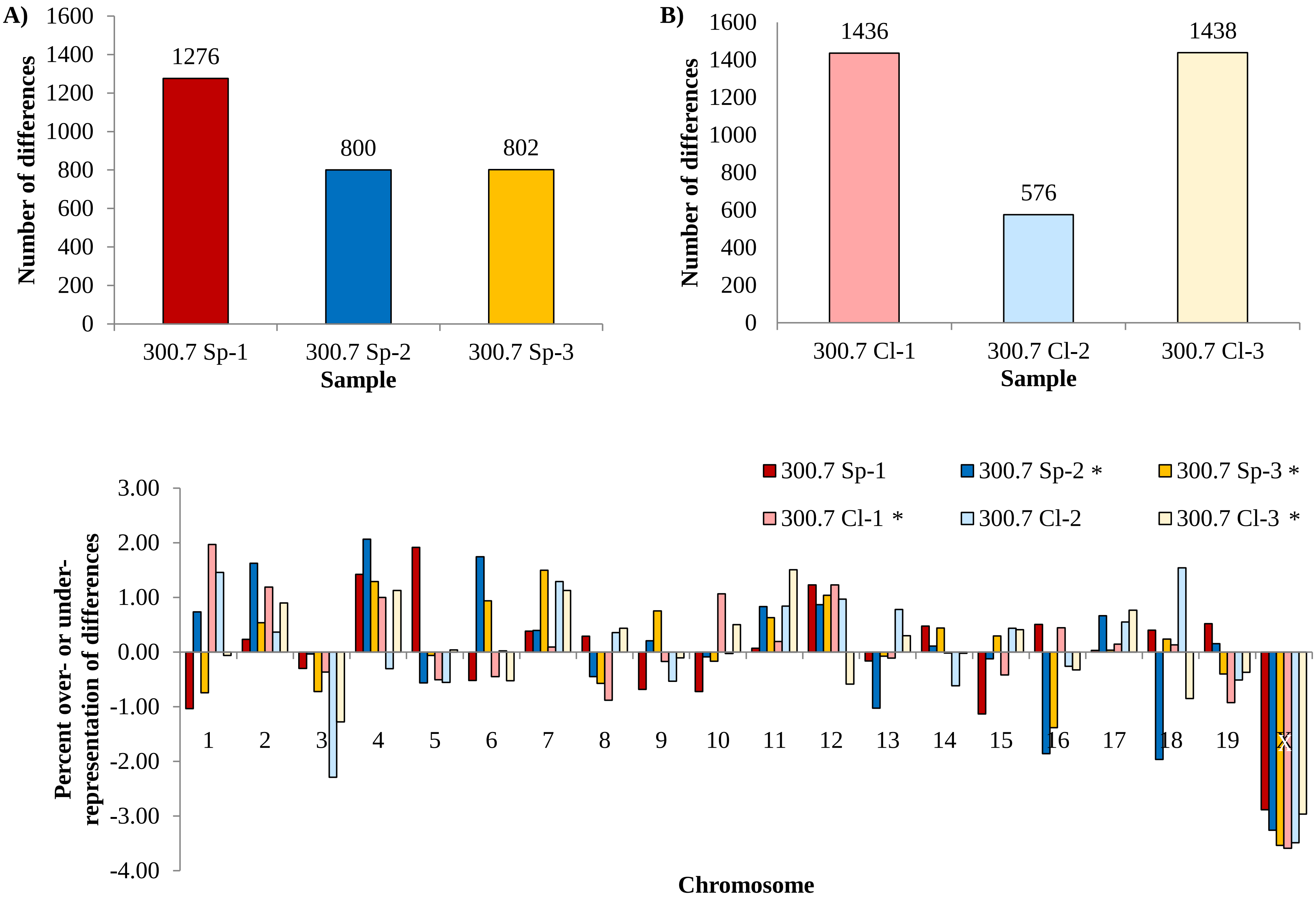
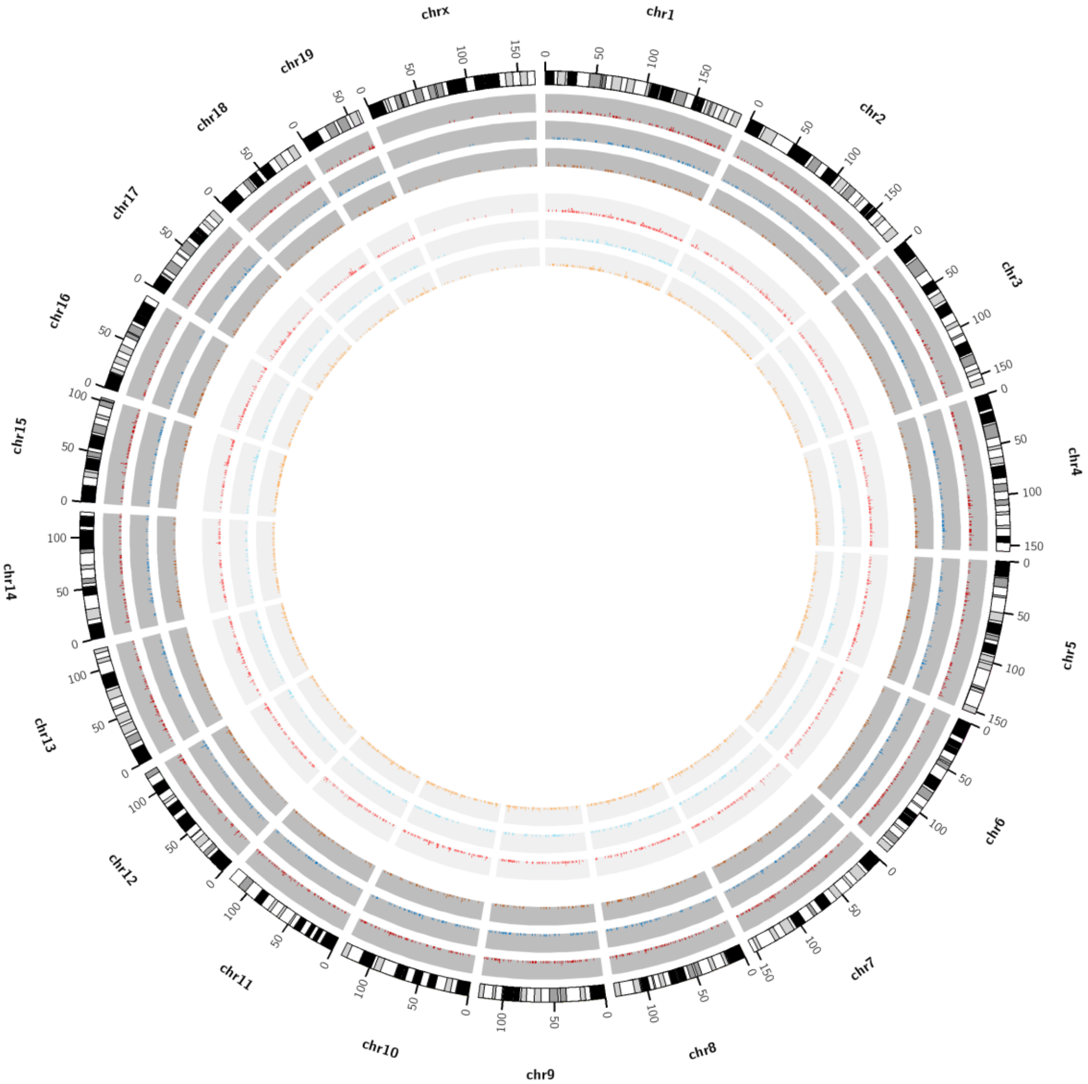

A)
Number of differences
1600
1400
1200
1000
800
600
400
200
0
1276
800
802
300.7 Sp-1
300.7 Sp-2
300.7 Sp-3
Sample
B)
Number of differences
1600
1400
1200
1000
800
600
400
200
0
1436
576
1438
300.7 Cl-1
300.7 Cl-2
300.7 Cl-3
Sample
300.7 Sp-1
300.7 Sp-2 *
300.7 Sp-3 *
300.7 Cl-1 *
300.7 Cl-2
300.7 Cl-3 *
Percent over- or under-representation of differences
3.00
2.00
1.00
0.00
-1.00
-2.00
-3.00
-4.00
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome

Figure 3-20: The distribution of genotype differences along the chromosomes for spleen and cerebellum replicates of mouse 300.7. The differences in genotype calls determined between the spleen replicates are the three outer rings (grey). The differences in genotype calls determined between the cerebellum replicates are the inner three rings (light grey). Each ring is ordered for the tissue types with the outer-most of the three representing replicate 1 (red for spleen, pink for cerebellum), the middle representing replicate 2 (dark blue for spleen, light blue for cerebellum), and the inner representing replicate 3 (dark orange for spleen, light orange for cerebellum).


chr1
0
50
100
150
chr2
0
50
100
150
chr3
0
50
100
150
chr4
0
50
100
150
chr5
0
50
100
150
chr6
0
50
100
chr7
0
50
100
150
chr8
0
50
100
chr9
0
50
100
chr10
0
50
100
chr11
0
50
100
chr12
0
50
100
chr13
0
50
100
chr14
0
50
100
chr15
0
50
100
chr16
0
50
chr17
0
50
chr18
0
50
chr19
0
50
chrx
0
50
100
150

3.9.4 Tissue comparisons between the spleen and cerebellum of a C57BL/6J mouse show genotype differences between tissues

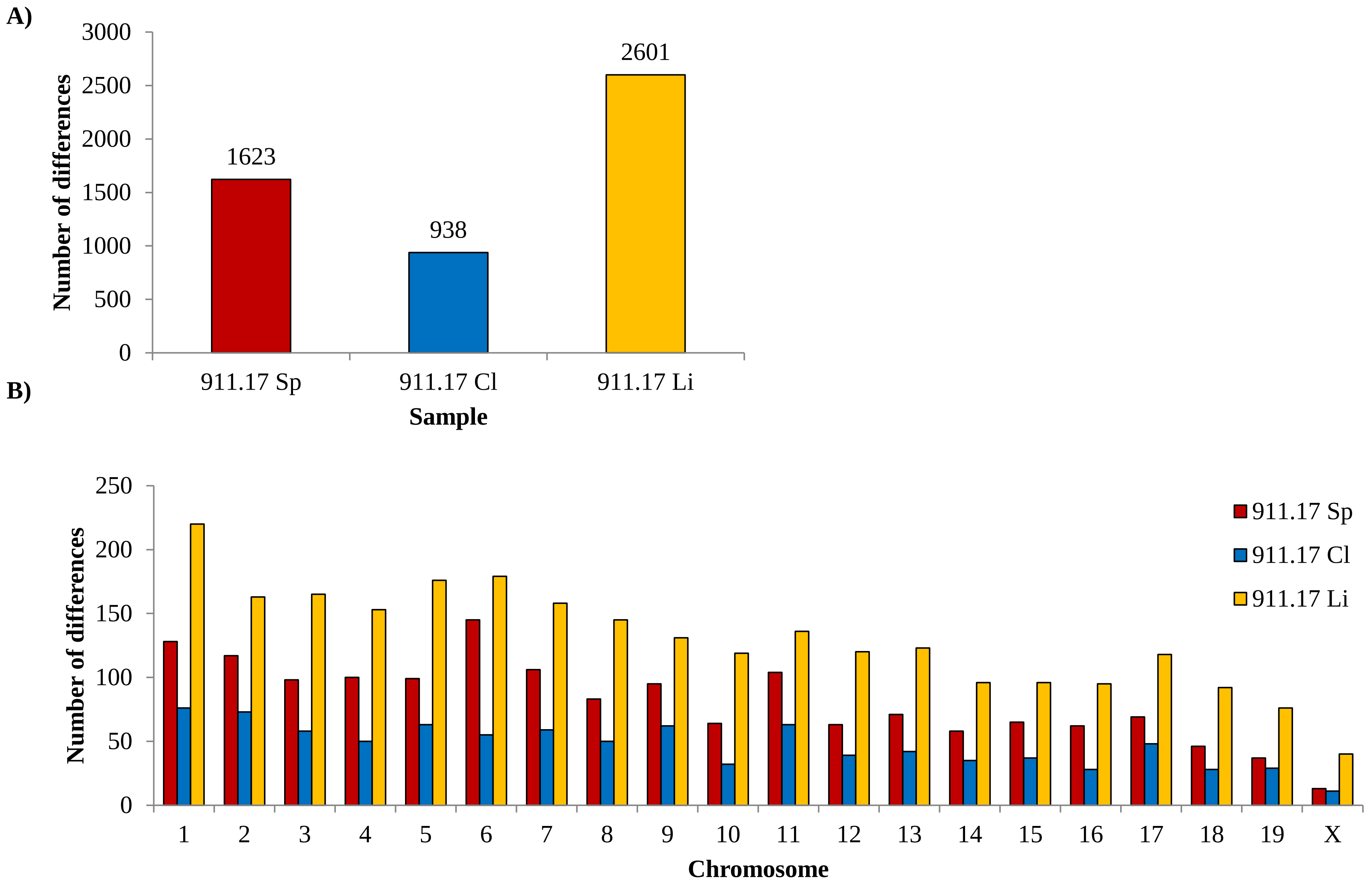
Both the spleen and cerebellum are similar in having a high degree of variation in the genotype differences between the three replicate samples (959 ±158 and 1,150 ± 287, respectively). The distribution of differences between the spleen and cerebellum of a B6 mouse was not significantly different (Figure 3-19). The distribution for each of these differences across the genome for the spleen and cerebellum was not significantly different in all nine possible pairwise comparisons. The distribution of these differences was significantly different from a random distribution in all comparisons where the 300.7 Sp-1, Sp-2, Cl-1, and Cl-2 were used ($p<0.05$; Fisher's Exact Test MCS).

3.10 Genotype differences between tissues

3.10.1 Tissue triads show genotype differences between the spleen, cerebellum, and liver of wild-type mice

The total number of genotype differences determined between the spleen, cerebellum, and liver for mouse 911.17 showed an average of 1,721 ± 483 genotype differences (Figure 3-21). The distribution of differences across the autosomes and X chromosome between these three tissues was not significantly different. The pattern of observed differences across different chromosomes for each of the three tissues was similar to the pattern of total genotype differences observed with cerebellum having the least number of differences consistently and the liver having the most.

Figure 3-21: Differences specific to the spleen, cerebellum, and liver of mouse 911.17. A) The total number of differences observed in the spleen (Sp), cerebellum (Cl) and liver (Li) of mouse 911.17 detected using the post-genotyping filtered SNP list (1,721 ± 483 genotype differences). **B)** Tissue-specific differences detected using the post-genotyping filtered SNP list.


A)
3000
2500
2000
1500
1000
500
0
Number of differences
1623
938
2601
911.17 Sp
911.17 Cl
911.17 Li
Sample
B)
250
200
150
100
50
0
Number of differences
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome
🟥 911.17 Sp
🟦 911.17 Cl
🟨 911.17 Li

The average number of genotype differences observed between the spleen, cerebellum, and liver for mouse 911.49 was 1,949 ± 1,718 (Figure 3-22). The distribution of tissue-specific genotype differences across the autosomes and the X chromosome was significantly different between the spleen, cerebellum, and liver ($p<0.001$; Fisher's Exact Test MCS). The significance observed between the three tissues was due to a difference in the distribution of genotype differences between the cerebellum and the liver ($p<0.001$; Fisher's Exact Test MCS). When the X chromosome was removed from analysis and the distribution across the autosomes was compared, tissues still showed a significant difference in the distribution of genotype differences across the autosomes ($p<0.001$; Fisher's Exact Test MCS).

For mouse 911.50, the average number of genotype differences between the spleen, cerebellum, and liver was 800 ± 469 (SEM) (Figure 3-23). The distribution of genotype differences across the autosomes and X chromosome between the three tissues was significantly different ($p<0.001$; Fisher's Exact Test MCS). Pairwise comparisons between each of the tissues showed that the spleen was significantly different from the cerebellum ($p<0.05$; Fisher's Exact Test MCS), the cerebellum was significantly different from the liver ($p<0.05$; Fisher's Exact Test MCS), and the spleen was significantly different from the liver ($p<0.05$; Fisher's Exact Test MCS). When the X chromosome was removed from analysis, the distribution of genotype differences between the three tissues was significantly different across the autosomes ($p<0.005$; Fisher's Exact Test MCS). Pairwise comparisons between the spleen and cerebellum, spleen and liver, and cerebellum and liver all showed a significant difference in the distribution of genotype differences across the autosomes ($p<0.05$; Fisher's Exact Test MCS).

Figure 3-22: Differences specific to the spleen, cerebellum, and liver of mouse 911.49. A) The total number of differences observed in the spleen (Sp), cerebellum (Cl) and liver (Li) of mouse 911.49 detected using the post-genotyping filtered SNP list (1,949 ± 1,718 genotype differences). **B)** Tissue-specific differences detected using the post-genotyping filtered SNP list. Distributions that are significantly different between tissues are ($p<0.05$; Fisher's Exact Test MCS) are indicated with an *.

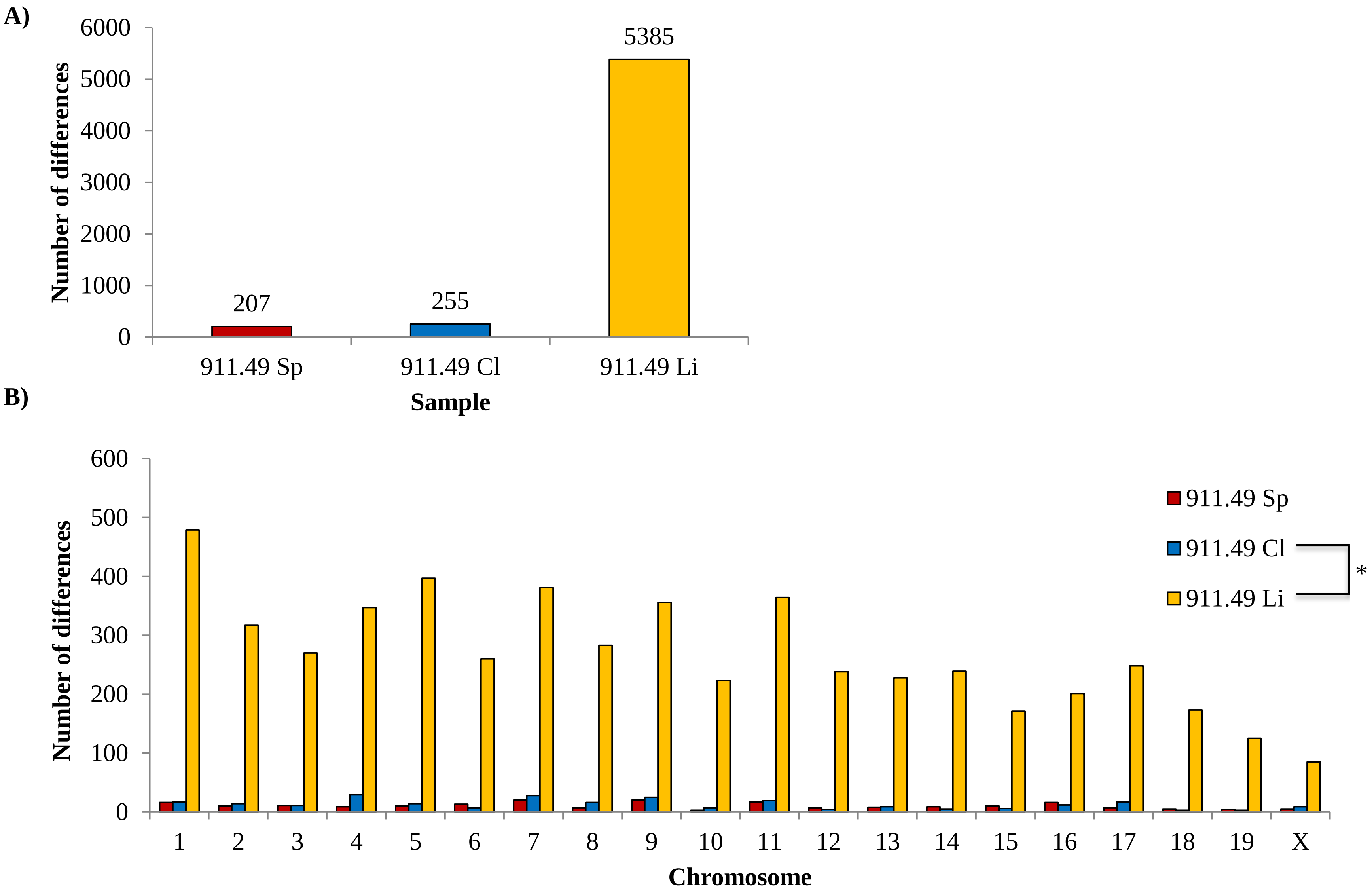
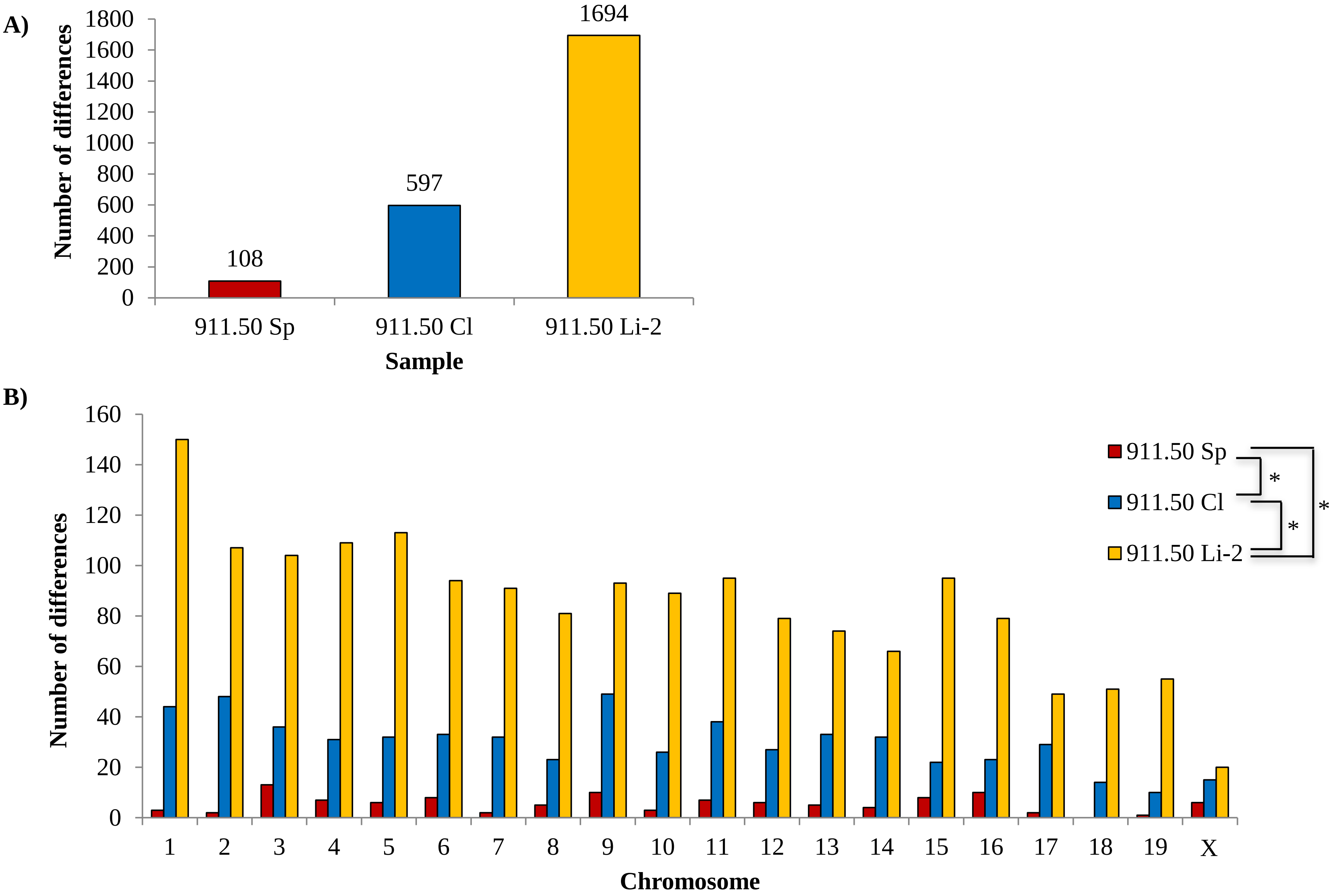

A)
Number of differences
6000
5000
4000
3000
2000
1000
0
207
255
5385
911.49 Sp
911.49 Cl
911.49 Li
Sample
B)
Number of differences
600
500
400
300
200
100
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome
🟥 911.49 Sp
🟦 911.49 Cl
🟨 911.49 Li
*

Figure 3-23: Differences specific to the spleen, cerebellum, and liver of mouse 911.50. A) The total number of differences observed in the spleen (Sp), cerebellum (Cl) and liver (Li) of mouse 911.50 detected using the post-genotyping filtered SNP list (800 ± 469 genotype differences). **B)** Tissue-specific differences detected using the post-genotyping filtered SNP list. Distributions that are significantly different between tissues are ($p<0.05$; Fisher's Exact Test MCS) are indicated with an *.


A)
Number of differences
1800
1600
1400
1200
1000
800
600
400
200
0
108
597
1694
911.50 Sp
911.50 Cl
911.50 Li-2
Sample
B)
Number of differences
160
140
120
100
80
60
40
20
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome
🟥 911.50 Sp
🟦 911.50 Cl
🟨 911.50 Li-2
*
*
*

The average number of genotype differences identified in the spleen (determined from comparisons against the cerebellum and liver) was 646 ± 489 (Figure 3-24). The distribution of these genotype differences compared between mouse 911.17, 911.49, and 911.50 was significantly different ($p<0.001$; Fisher's Exact Test MCS). The distribution of differences compared to a random distribution was significantly different for sample 911.17 ($p<0.001$; Fisher's Exact Test MCS), however, was randomly distributed for samples 911.49 and 911.50. The number of differences on the X chromosome was under-represented for mouse 911.17 and 911.49 and over-represented for mouse 911.50. When the X chromosome was removed from analysis, the distribution of genotype differences across the autosomes was significantly different between the three mice ($p<0.05$; Fisher's Exact Test MCS). The distribution of genotype differences across the autosomes was not significantly different from random for all mice.

The average number of cerebellum-specific differences detected for mouse 911.17, 911.49, and 911.50 in comparison to the spleen and liver was 597 ± 197 (Figure 3-25). The distribution of these genotype differences across the autosomes and X chromosome was significantly different between the three mice ($p<0.001$; Fisher's Exact Test MCS). Compared to a random distribution of genotype differences, sample 911.17 and 911.50 were not different. However, sample 911.49 was significantly different from a random distribution ($p<0.001$; Fisher's Exact Test MCS). The number of genotype differences on the X chromosome was over-represented for mouse 911.17, and under-represented for mouse 911.49 and 911.50. When the X chromosome was removed from analysis, the distribution of genotype differences across the autosomes was significantly different between the three mice ($p<0.05$; Fisher's Exact Test MCS). In comparison to a random distribution of genotype

Figure 3-24: Spleen-specific differences observed for the three wild-type mice. A) The total number of spleen-specific differences determined in comparison to the cerebellum and liver for each WT mouse. (646 ± 489 genotype differences) **B)** The distribution of observed spleen-specific differences in genotype graphed to show instances of over- and under-representation from a random occurrence based on the number of genotype calls per chromosome for three WT mice. * indicates samples with a distribution significantly different from the expected random distribution (p<0.005; Fisher's Exact Test MCS).

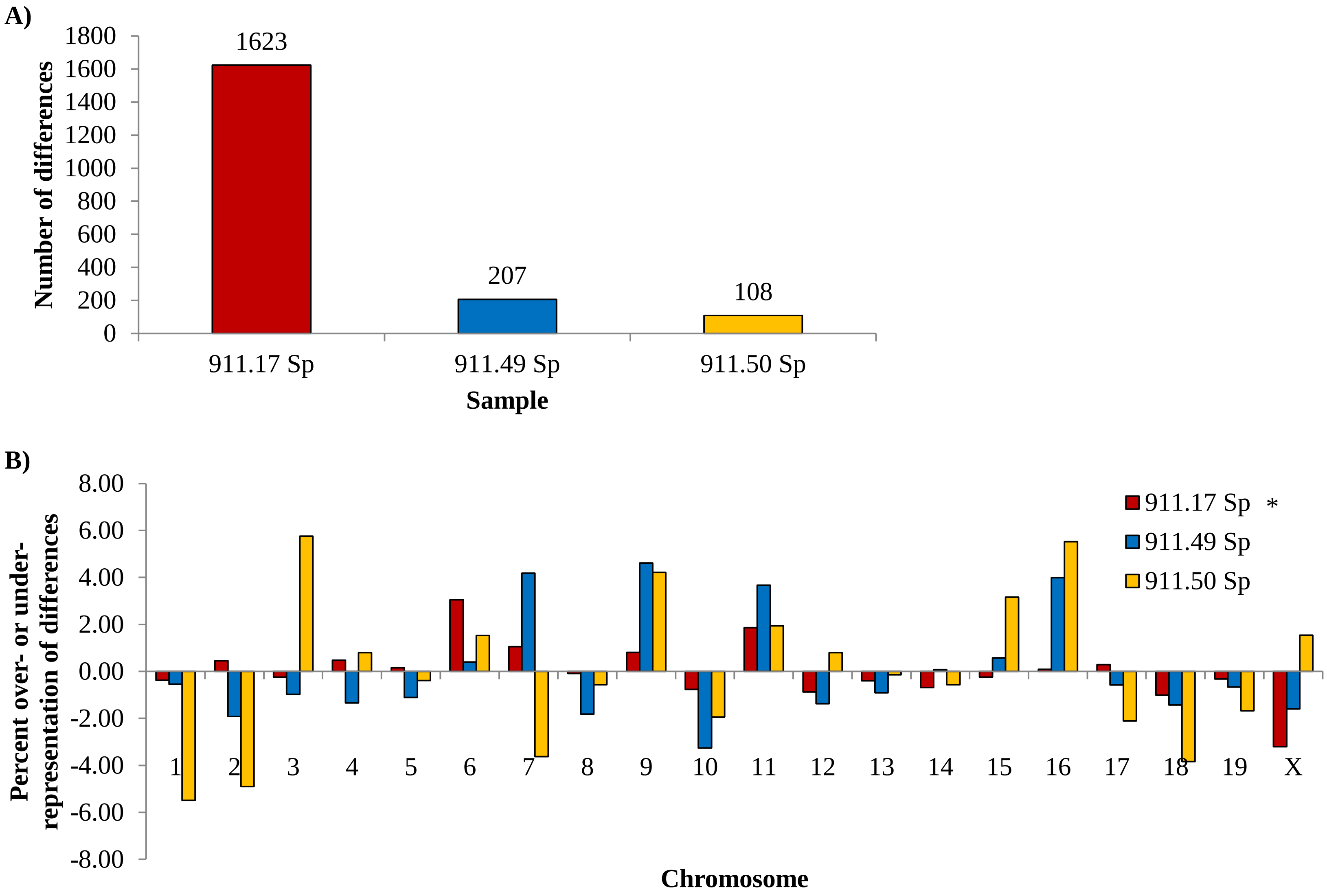
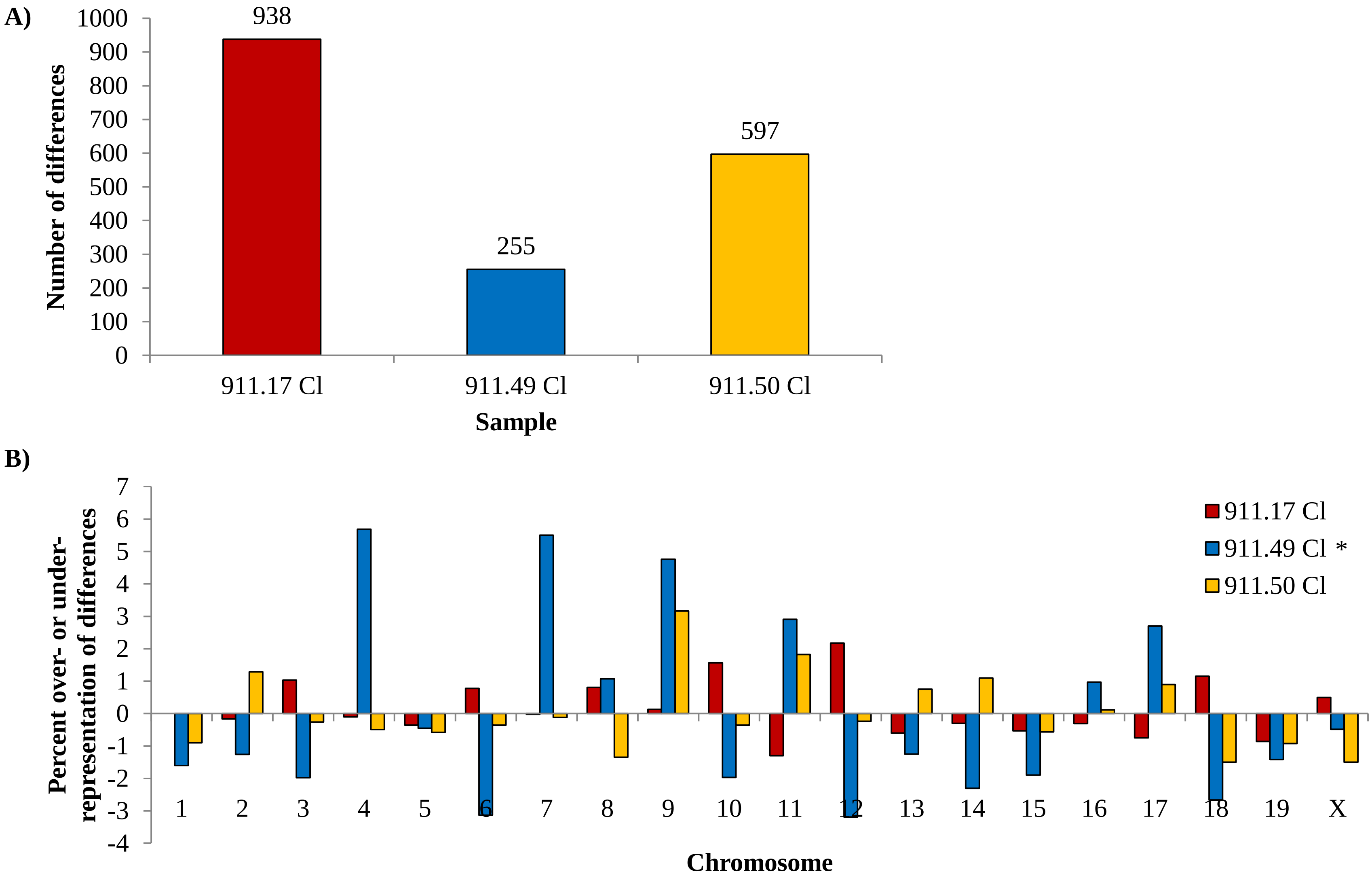

A)
Number of differences
1800
1600
1400
1200
1000
800
600
400
200
0
1623
207
108
911.17 Sp
911.49 Sp
911.50 Sp
Sample
B)
Percent over- or under-representation of differences
8.00
6.00
4.00
2.00
0.00
-2.00
-4.00
-6.00
-8.00
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome
🟥 911.17 Sp *
🟦 911.49 Sp
🟨 911.50 Sp

Figure 3-25: Cerebellum-specific differences between the three wild-type mice. A) The total number of cerebellum-specific differences determined in comparison to the spleen and the liver for each WT mouse (597 ± 197 genotype differences). **B)** The distribution of observed cerebellum-specific differences in genotype graphed to show instances of over- and under-representation from a random occurrence based on the number of genotype calls per chromosome for three WT mice. * indicates samples with a distribution significantly different from the expected random distribution ($p<0.005$; Fisher's Exact Test MCS).


A)
1000
900
800
700
600
500
400
300
200
100
0
Number of differences
938
255
597
911.17 Cl
911.49 Cl
911.50 Cl
Sample
B)
7
6
5
4
3
2
1
0
-1
-2
-3
-4
Percent over- or under-representation of differences
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome
🟥 911.17 Cl
🟦 911.49 Cl *
🟨 911.50 Cl

differences across the autosomes, only mouse 911.49 showed a significant difference ($p<0.05$ Fisher's Exact Test MCS).

In the liver, an average of 3,227 ± 1,110 genotype differences was detected for mouse 911.17, 911.49, and 911.50 when samples were compared to the spleen and cerebellum (Figure 3-26). The distribution of these genotype differences across the autosomes and X chromosome was significantly different between the three mice ($p<0.001$; Fisher's Exact Test MCS). All samples had a distribution of genotype differences that was significantly different from a random distribution of differences ($p<0.01$; Fisher's Exact Test MCS). The number of genotype differences on the X chromosome was over-represented for mouse 911.17, and under-represented for mouse 911.49 and 911.50. When the X chromosome was removed from analysis, the distribution of genotype differences across the autosomes was significantly different between the three mice ($p<0.001$; Fisher's Exact Test MCS). In comparison to a random distribution of genotype differences across the autosomes, only mouse 911.49 showed a significantly different distribution from random ($p<0.001$; Fisher's Exact Test MCS).

The distribution of genotype differences for each mouse was graphed for the spleen, cerebellum, and liver using Circos to determine the distribution of genotype differences along each of the chromosomes (Figure 3-27). The small number of differences for the spleen, cerebellum, and liver did not appear to cluster along the chromosomes.

Figure 3-26: Liver-specific differences between the three wild-type mice. A) The total number of liver-specific differences determined in comparison to the spleen and the cerebellum for each WT mouse (3,227 ± 1,110 genotype differences). **B)** The distribution of observed liver-specific differences in genotype graphed to show instances of over- and under-representation from a random occurrence based on the number of genotype calls per chromosome for three WT mice. * indicates samples with a distribution significantly different from the expected random distribution ($p<0.005$; Fisher's Exact Test MCS).

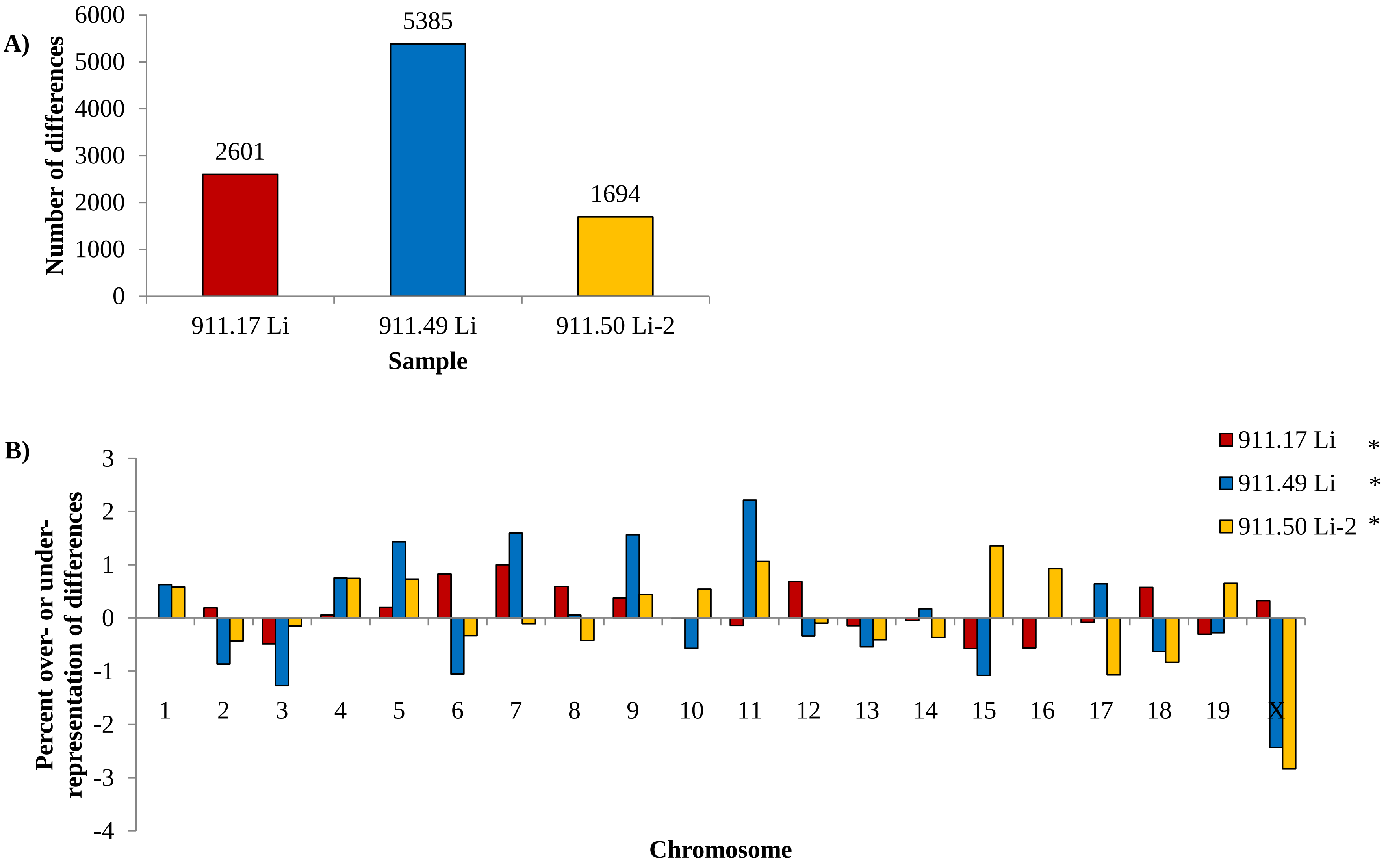
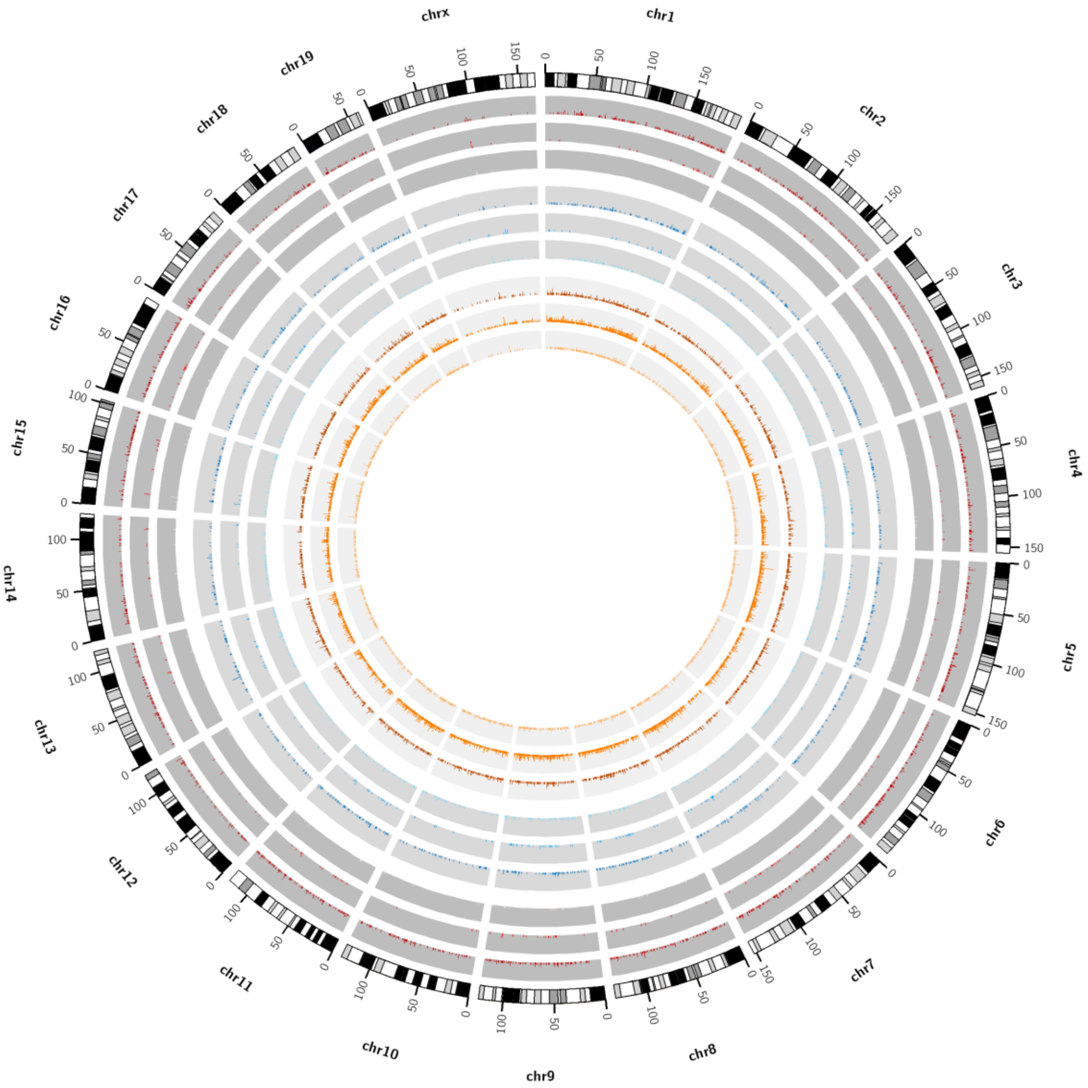

A)
Number of differences
6000
5000
4000
3000
2000
1000
0
2601
5385
1694
911.17 Li
911.49 Li
911.50 Li-2
Sample
B)
Percent over- or under-representation of differences
3
2
1
0
-1
-2
-3
-4
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome
🟥 911.17 Li *
🟦 911.49 Li *
🟨 911.50 Li-2 *

Figure 3-27: The distribution of differences along each of the chromosomes for the spleen, cerebellum, and liver of mouse 911.17, 911.49, and 911.50. The spleen-specific differences are indicated in the three outer-most rings (dark grey) in red, the cerebellum-specific differences are indicated in the middle three rings (light grey) in blue, and the liver-specific differences are indicated in the three inner-most rings (very light grey) in orange. The groups of three rings for the spleen, cerebellum, and liver samples each represent a different mouse and are ordered with the outer-most ring as mouse 911.17, middle ring as 911.49 and inner-most ring as 911.50.


chr1
0
50
100
150
chr2
0
50
100
150
chr3
0
50
100
150
chr4
0
50
100
150
chr5
0
50
100
150
chr6
0
50
100
chr7
0
50
100
150
chr8
0
50
100
chr9
0
50
100
chr10
0
50
100
chr11
0
50
100
chr12
0
50
100
chr13
0
50
100
chr14
0
50
100
chr15
0
50
100
chr16
0
50
chr17
0
50
chr18
0
50
chr19
0
50
chrx
0
50
100
150

3.10.2 Replicates of tissue triads in a wild-type mouse show high amounts of genetic variation within and between tissues

The replicate of tissue triads for mouse 911.50 showed an average of 12,255 ± 10,862 (SEM) genotype differences between the spleen, cerebellum, and liver (Figure 3-28). The distribution of these genotype differences was significantly different across the autosomes and X chromosome between the spleen, cerebellum, and liver ($p<0.001$; Fisher's Exact Test MCS). There was a significant difference in the distribution of genotype differences between the spleen and liver, and cerebellum and liver ($p<0.001$, $p<0.001$ respectively; Fisher's Exact Test MCS); however, there was no significant difference between the spleen and cerebellum. The distribution of differences in the spleen, cerebellum, and liver were all significantly different from a random distribution ($p<0.001$, $p<0.001$, $p<0.001$ respectively; Fisher's Exact Test MCS). On the X chromosome, all three tissues showed an under-representation of genotype differences. When the X chromosome was removed from analysis, the distribution of genotype differences between the three tissues was significantly different ($p<0.001$; Fisher's Exact Test MCS). In comparison to a random distribution of genotype differences across the autosomes, only the liver was significantly different from a random distribution of differences. The distribution of differences for each of the tissues is shown along each of the chromosomes (Figure 3-29).

Figure 3-28: Tissue-specific differences for the spleen, cerebellum, and liver of the replicates for mouse 911.50. A) The total number of tissue-specific differences in the spleen, cerebellum, and liver for mouse 911.50. **B)** The distribution of observed differences in genotype graphed to show instances of over- and under-representation from a random occurrence based on the number of genotype calls per chromosome. * indicates samples with a distribution significantly different from the expected random distribution and when tissues have a different distribution from each other (p<0.005; Fisher's Exact Test MCS).

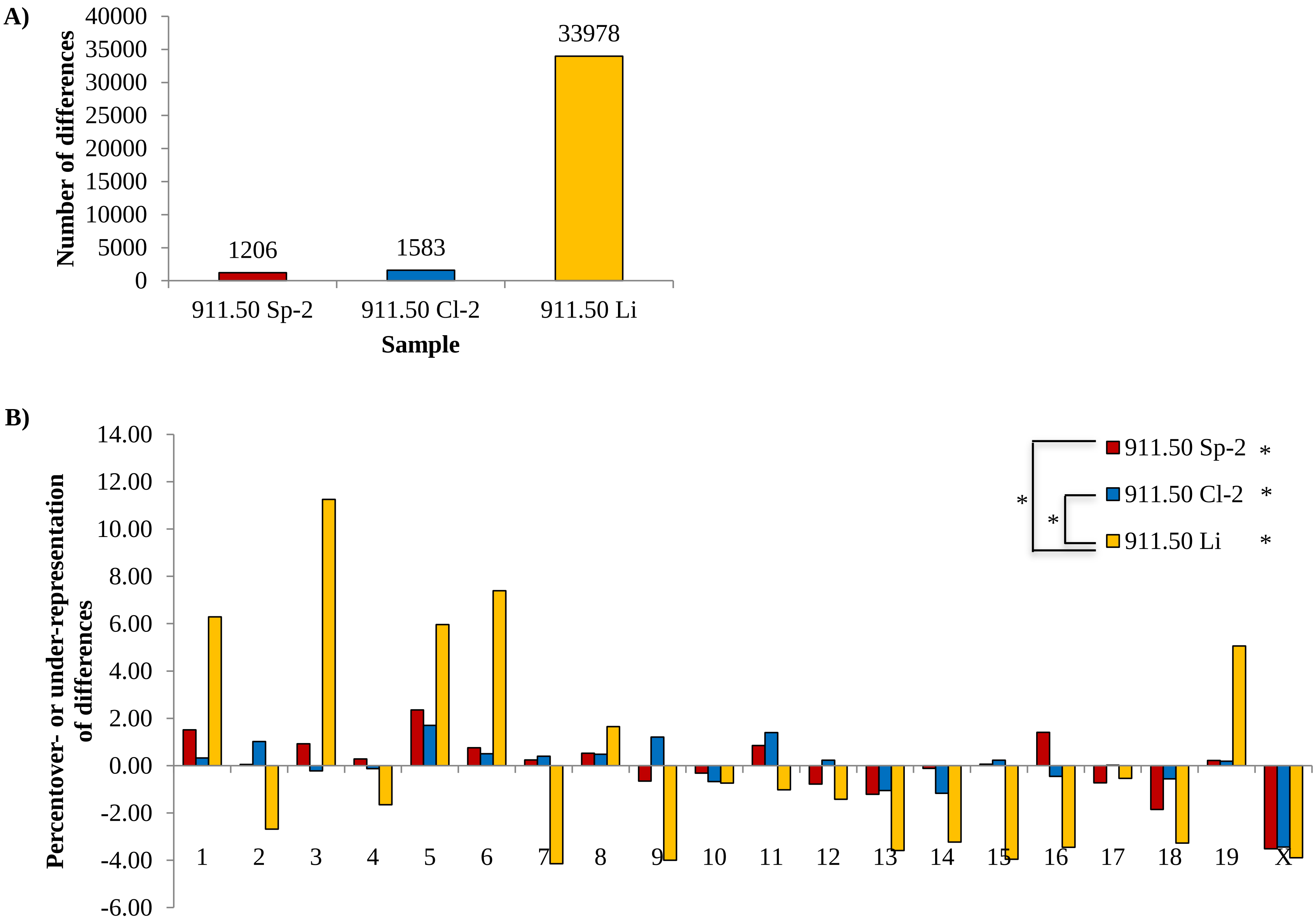
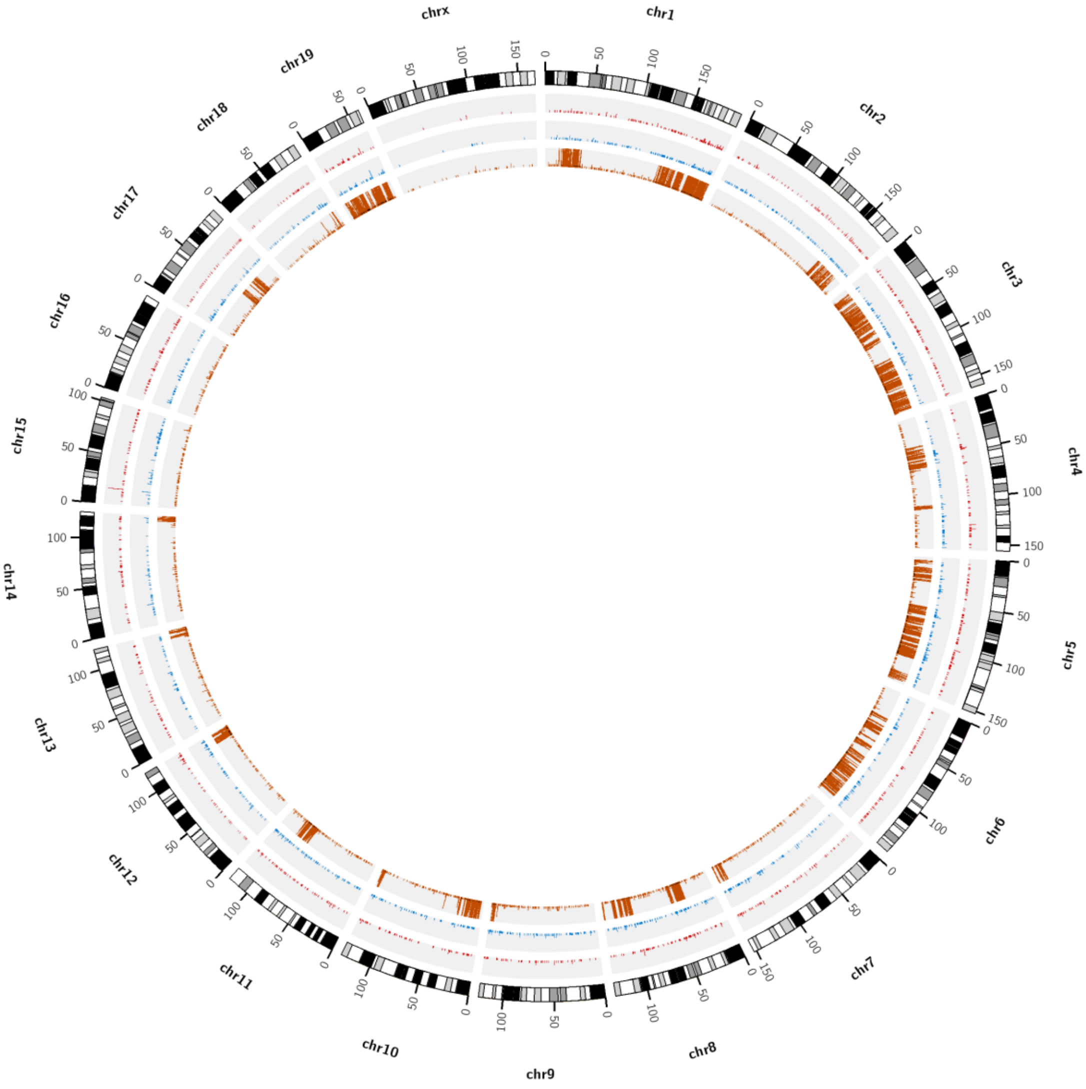
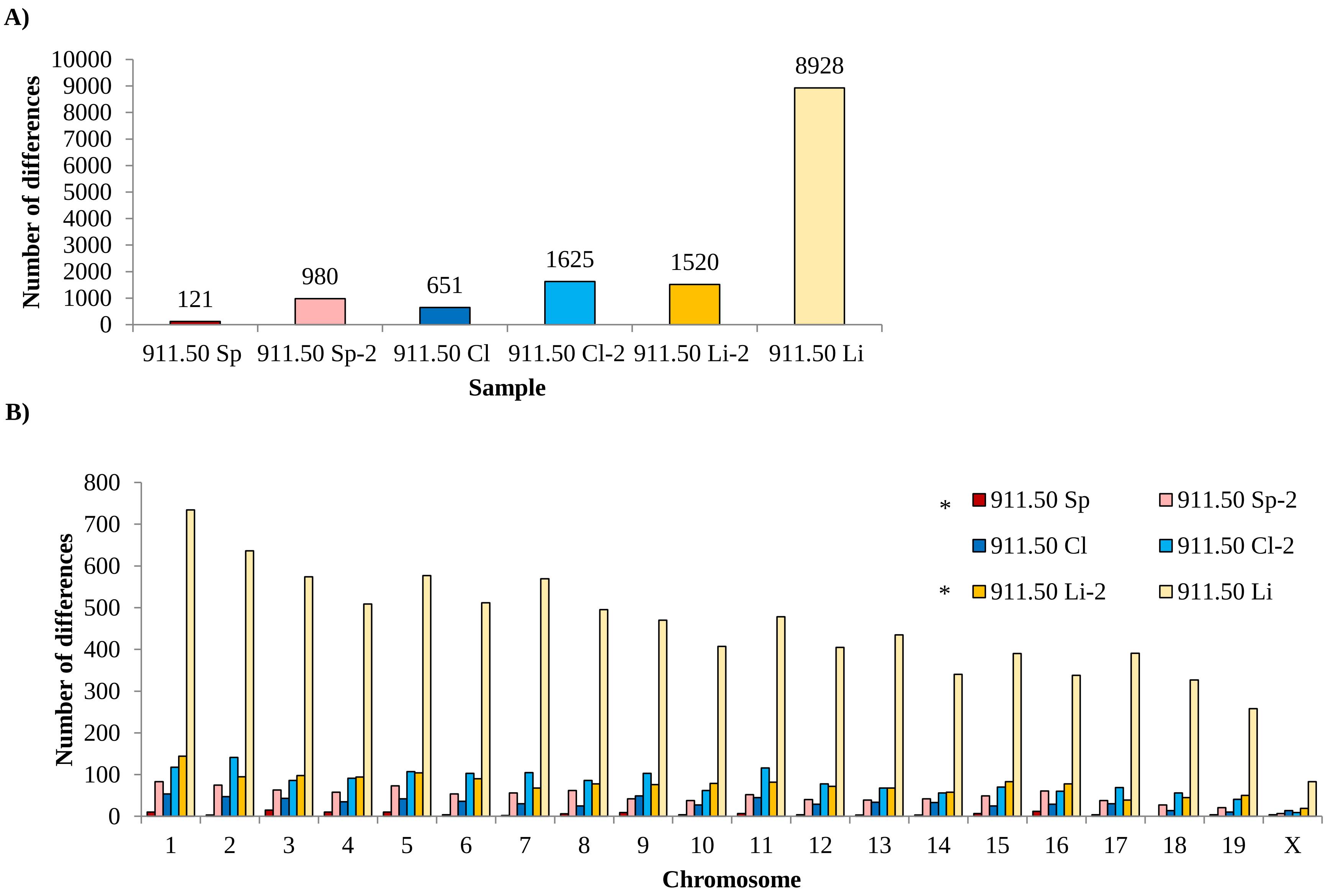

A)
Number of differences
40000
35000
30000
25000
20000
15000
10000
5000
0
1206
1583
33978
911.50 Sp-2
911.50 Cl-2
911.50 Li
Sample
B)
Percentover- or under-representation of differences
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
-2.00
-4.00
-6.00
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
🟥 911.50 Sp-2 *
🟦 911.50 Cl-2 *
🟨 911.50 Li *
*
*

Figure 3-29: The distribution of differences in the spleen, cerebellum, and liver replicates for mouse 911.50. The outer-most circle shows spleen-specific differences in red, the middle displays cerebellum-specific differences in blue, and the inner-most circle shows the liver-specific differences in yellow. The replicate liver for mouse 911.50 is an outlier in the sample set. Clusters of colour indicate regions that are prone to the detection of differences.


chr1
0
50
100
150
chr2
0
50
100
150
chr3
0
50
100
150
chr4
0
50
100
150
chr5
0
50
100
150
chr6
0
50
100
chr7
0
50
100
150
chr8
0
50
100
chr9
0
50
100
chr10
0
50
100
chr11
0
50
100
chr12
0
50
100
chr13
0
50
100
chr14
0
50
100
chr15
0
50
100
chr16
0
50
chr17
0
50
chr18
0
50
chr19
0
50
chrx
0
50
100
150

The total number of genotype differences varied between the spleen, cerebellum, and liver replicates (Figure 3-30). The distribution of genotype differences across the autosomes and X chromosome observed between replicates was significantly different for the spleen and the liver, but not the cerebellum ($p<0.05$, $p<0.001$ respectively; Fisher's Exact Test MCS). When the X chromosome was removed from analysis, the distribution of genotype differences across the autosomes between tissue replicates was only significantly different for the liver ($p<0.001$; Fisher's Exact Test MCS). The replicate for the liver had far more genotype differences in comparison to any other tissue sample from mouse 911.50. The distribution of these differences when graphed along each of the chromosomes clustered to specific regions on all chromosomes except for chromosome 15, 16, and X (Figure 3-29).

Figure 3-30: Differences specific to the spleen, cerebellum, and liver between replicates of mouse 911.50. A) The total number of differences observed in the spleen (Sp), cerebellum (Cl) and liver (Li) of mouse 911.50 detected using the post-genotyping filtered SNP list. **B)** Tissue-specific differences detected using the post-genotyping filtered SNP list. * indicates a significant difference between replicates of a tissue type (p<0.05).


A)
Number of differences
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
121
980
651
1625
1520
8928
911.50 Sp
911.50 Sp-2
911.50 Cl
911.50 Cl-2
911.50 Li-2
911.50 Li
Sample
B)
Number of differences
800
700
600
500
400
300
200
100
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome
* 911.50 Sp
911.50 Sp-2
911.50 Cl
911.50 Cl-2
* 911.50 Li-2
911.50 Li

3.10.3 Tissue pairs show genotype differences between tissues of a wild-type mouse and tissues of a *harlequin* mouse

The total number of genotype differences in the spleen and the liver of mouse 900.3 was 3,859 and 3,982 respectively (Figure 3-31). The distribution of genotype differences across the autosomes and X chromosome for the spleen and the liver was significantly different (p<0.001; Fisher's Exact Test MCS). In comparison to a random distribution of genotype differences across the autosomes and X chromosome, both the spleen and the liver were significantly different (p<0.001, p<0.001 respectively; Fisher's Exact Test MCS). Both the spleen and the liver showed an over-representation of differences on the X-chromosome (Figure 3-32). When the X chromosome was removed from analysis, there was no significant difference in the distribution of genotype differences across the autosomes. When compared to a random distribution of genotype differences across the autosomes, neither the spleen nor the liver was significantly different from a random distribution.

The total number of differences in the spleen and cerebellum for mouse 904.9 was 246 and 244 respectively (Figure 3-33). The distribution of differences between the spleen and cerebellum across the autosomes and the X chromosome was significantly different (p<0.001; Fisher's Exact Test MCS). When the X chromosome was removed from analysis, the distribution of genotype differences across the autosomes was significantly different between the spleen and the cerebellum (p<0.001; Fisher's Exact Test MCS).

For mouse 904.11, a total of 225 genotype differences were detected in the spleen and 299 genotype differences were detected in the cerebellum (Figure 3-34). The distribution of these differences across the autosomes and X chromosome was not significantly different.

Figure 3-31: Tissue-specific differences detected between the spleen and liver of mouse 900.3. A) The total number of differences between the spleen and the liver. **B)** The distribution of observed differences in genotype graphed to show instances of over- and under-representation from a random occurrence based on the number of genotype calls per chromosome. * indicates samples with a distribution significantly different from expected ($p<0.005$; Fisher's Exact Test MCS).

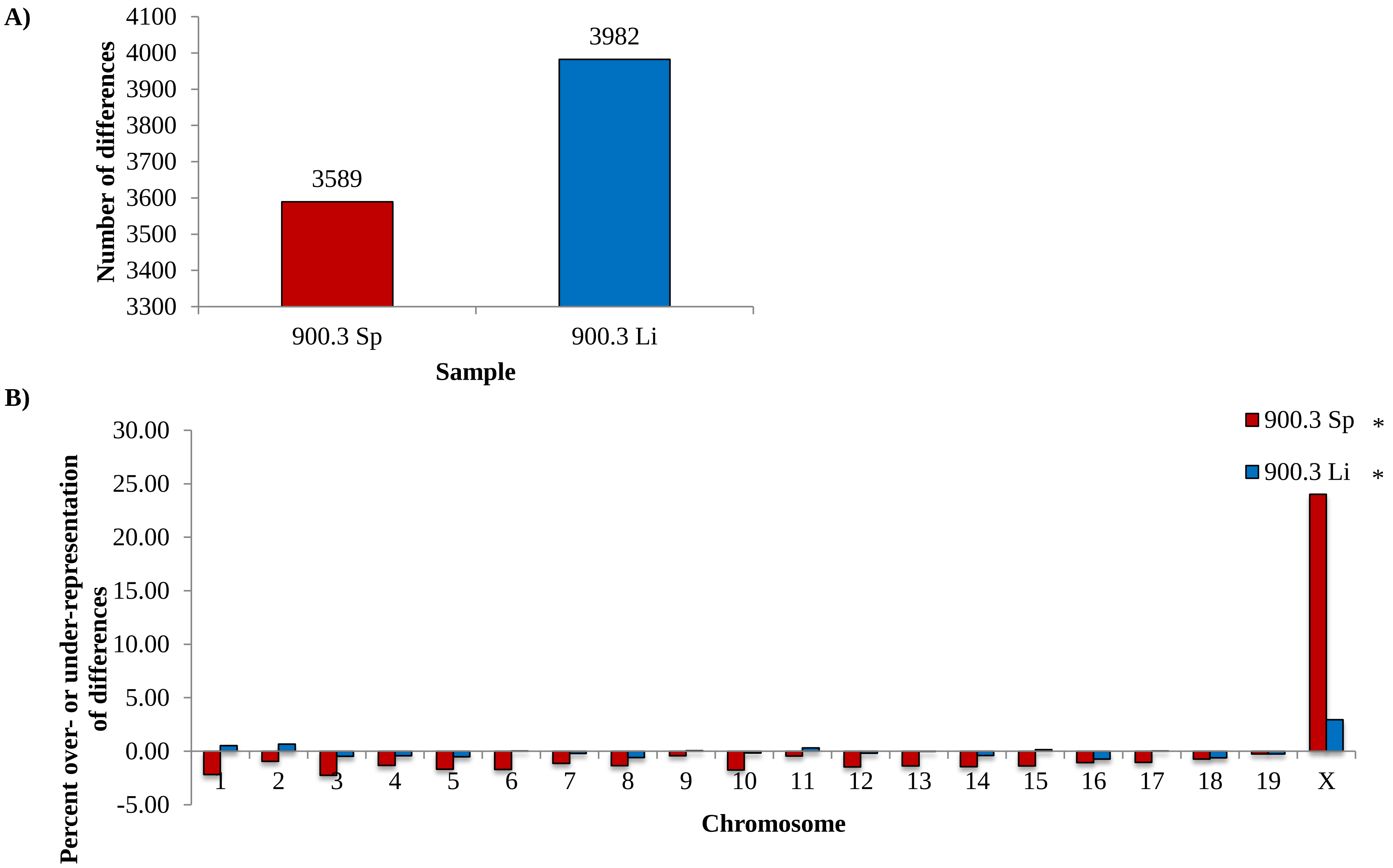

A)
Number of differences
4100
4000
3900
3800
3700
3600
3500
3400
3300
3589
3982
900.3 Sp
900.3 Li
Sample
B)
Percent over- or under-representation of differences
30.00
25.00
20.00
15.00
10.00
5.00
0.00
-5.00
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome
🟥 900.3 Sp *
🟦 900.3 Li *

Figure 3-32: The distribution of tissue-specific differences across the chromosomes for mouse 900.3. Spleen specific differences are plotted in red and liver specific differences are plotted in blue. The distribution of differences does not cluster along any of the chromosomes. Chromosome X does show a high proportion of differences in comparison to the remainder of the genome.

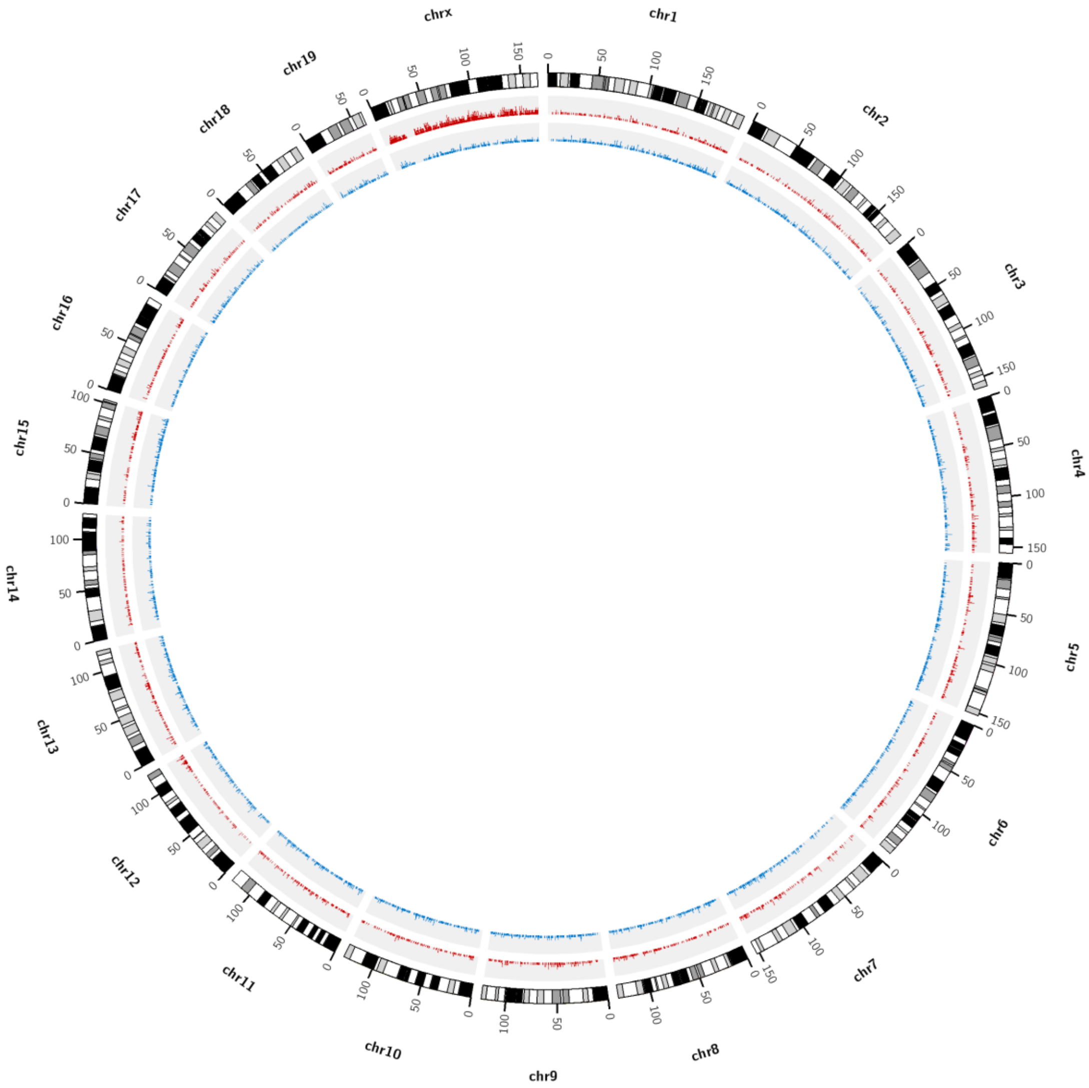

chr1
0
50
100
150
chr2
0
50
100
150
chr3
0
50
100
150
chr4
0
50
100
150
chr5
0
50
100
150
chr6
0
50
100
chr7
0
50
100
150
chr8
0
50
100
chr9
0
50
100
chr10
0
50
100
chr11
0
50
100
chr12
0
50
100
chr13
0
50
100
chr14
0
50
100
chr15
0
50
100
chr16
0
50
chr17
0
50
chr18
0
50
chr19
0
50
chrx
0
50
100
150

Figure 3-33: Tissue-specific differences detected between the spleen and cerebellum of mouse 904.9. A) The total number of differences between the spleen and the cerebellum. **B)** Tissue-specific differences detected using the post-genotyping filtered SNP list. The distribution of differences in the spleen was significantly different from the distribution of differences in the cerebellum across chromosomes (p<0.001; Fisher's Exact Test MCS).

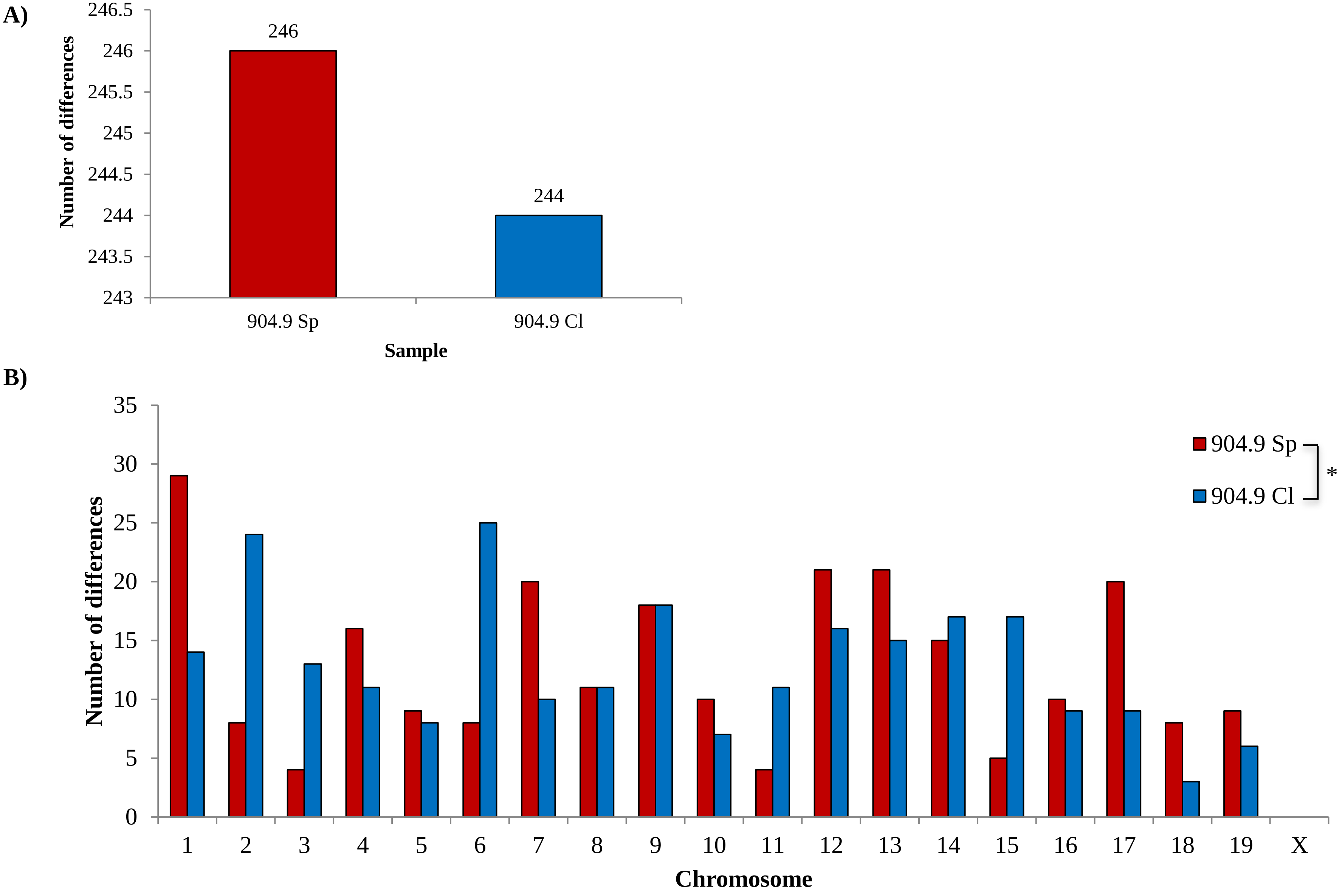
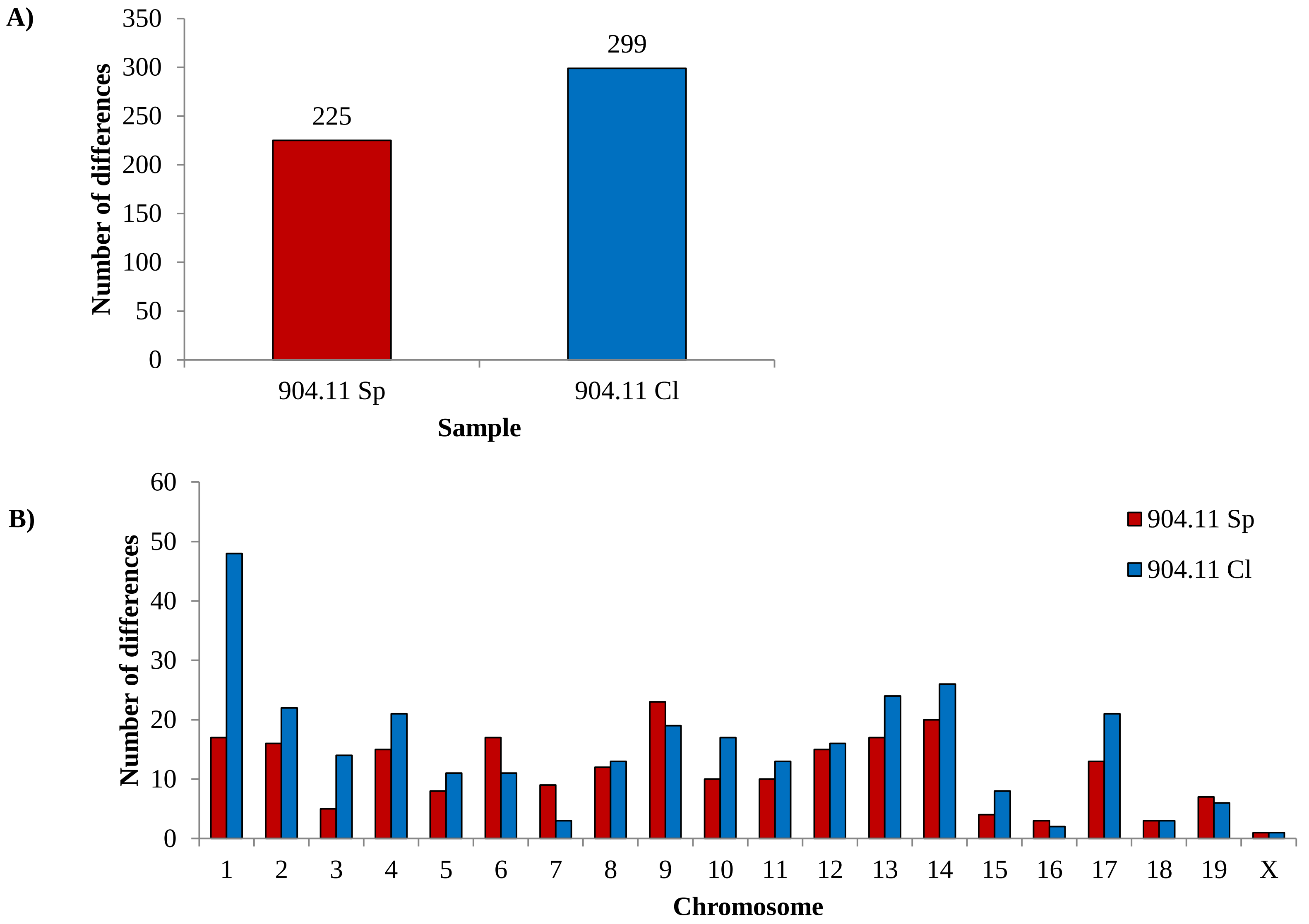

A)
246.5
246
245.5
245
244.5
244
243.5
243
Number of differences
246
244
904.9 Sp
904.9 Cl
Sample
B)
35
30
25
20
15
10
5
0
Number of differences
🟥 904.9 Sp
🟦 904.9 Cl
*
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome

Figure 3-34: Tissue-specific differences detected between the spleen and cerebellum of mouse 904.11. A) The total number of differences between the spleen and the cerebellum. **B)** Tissue-specific differences detected using the post-genotyping filtered SNP list. The distribution of differences across chromosomes for the spleen and the cerebellum was not significantly different.


A)
Number of differences
350
300
250
200
150
100
50
0
225
299
904.11 Sp
904.11 Cl
Sample
B)
Number of differences
60
50
40
30
20
10
0
🟥 904.11 Sp
🟦 904.11 Cl
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome

The total number of spleen-specific differences, in comparison to the cerebellum, between the two *hq* mice 904.9 and 904.11 was 246 and 225 respectively (Figure 3-35). The distribution of differences in each of the spleen samples was not significantly different across the autosomes and X chromosome. The distribution of differences across the autosomes and X chromosome was significantly different from a random distribution of genotype differences for both mice (p<0.001, p<0.05; Fisher's Exact Test MCS). Both mice showed an under-representation of genotype differences on the X chromosome. When the X chromosome was removed from analysis, the distribution of genotype differences across the autosomes was not significantly different between mice. The distribution of genotype differences across the autosomes in comparison to a random distribution of genotype differences was significantly different from random for mouse 904.9 and mouse 904.11 (p<0.001, p<0.01 respectively; Fisher's Exact Test MCS).

The total number of cerebellum-specific differences, detected in comparison to the spleen, between mouse 904.9 and 904.11 was 244 and 299 respectively (Figure 3-36). The distribution of genotype differences between the two cerebellum samples across the autosomes and X chromosome was significantly different (p<0.001; Fisher's Exact Test MCS). When compared to a random distribution across the autosomes and X chromosome, the distribution of cerebellum-specific differences in mouse 904.9 was not different from random; however, mouse 904.11 did have a significantly different distribution from random (p<0.001; Fisher's Exact Test MCS). Both mice showed an under-representation of genotype differences on the X chromosome. When the X chromosome was removed from analysis, the distribution of genotype differences across the autosomes was significantly different between mice (p<0.001; Fisher's Exact Test MCS). The distribution of genotype differences across the autosomes in comparison to a random distribution of genotype differences was

Figure 3-35: Spleen-specific differences between the two *harlequin* mice. A) The total number of spleen-specific differences in the cerebellum for the two *hq* mice. **B)** The distribution of observed spleen-specific differences in genotype graphed to show instances of over- and under-representation from a random occurrence based on the number of genotype calls per chromosome in two *hq* mice. * indicates samples with a distribution significantly different from the expected random distribution (p<0.005; Fisher's Exact Test MCS).

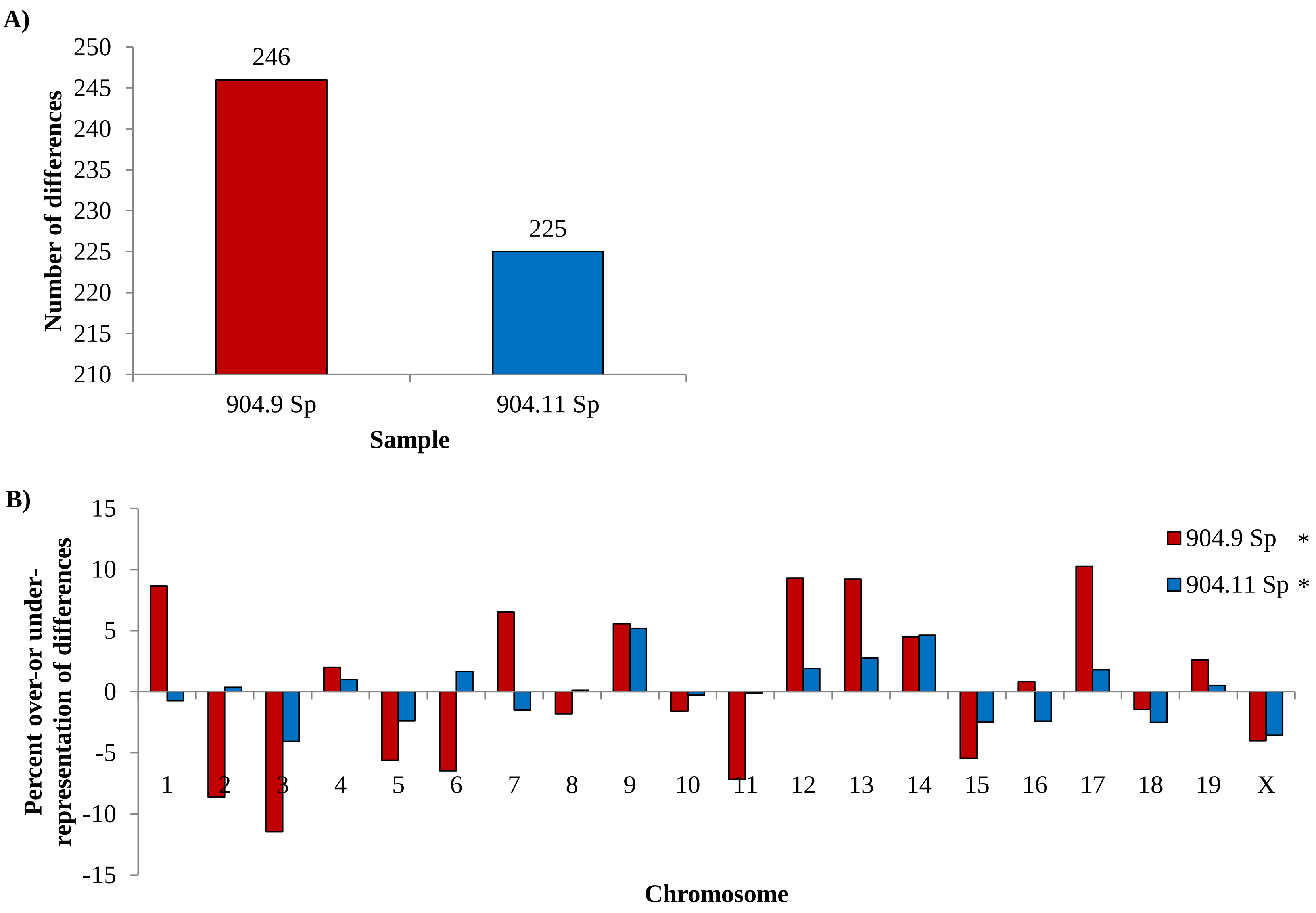
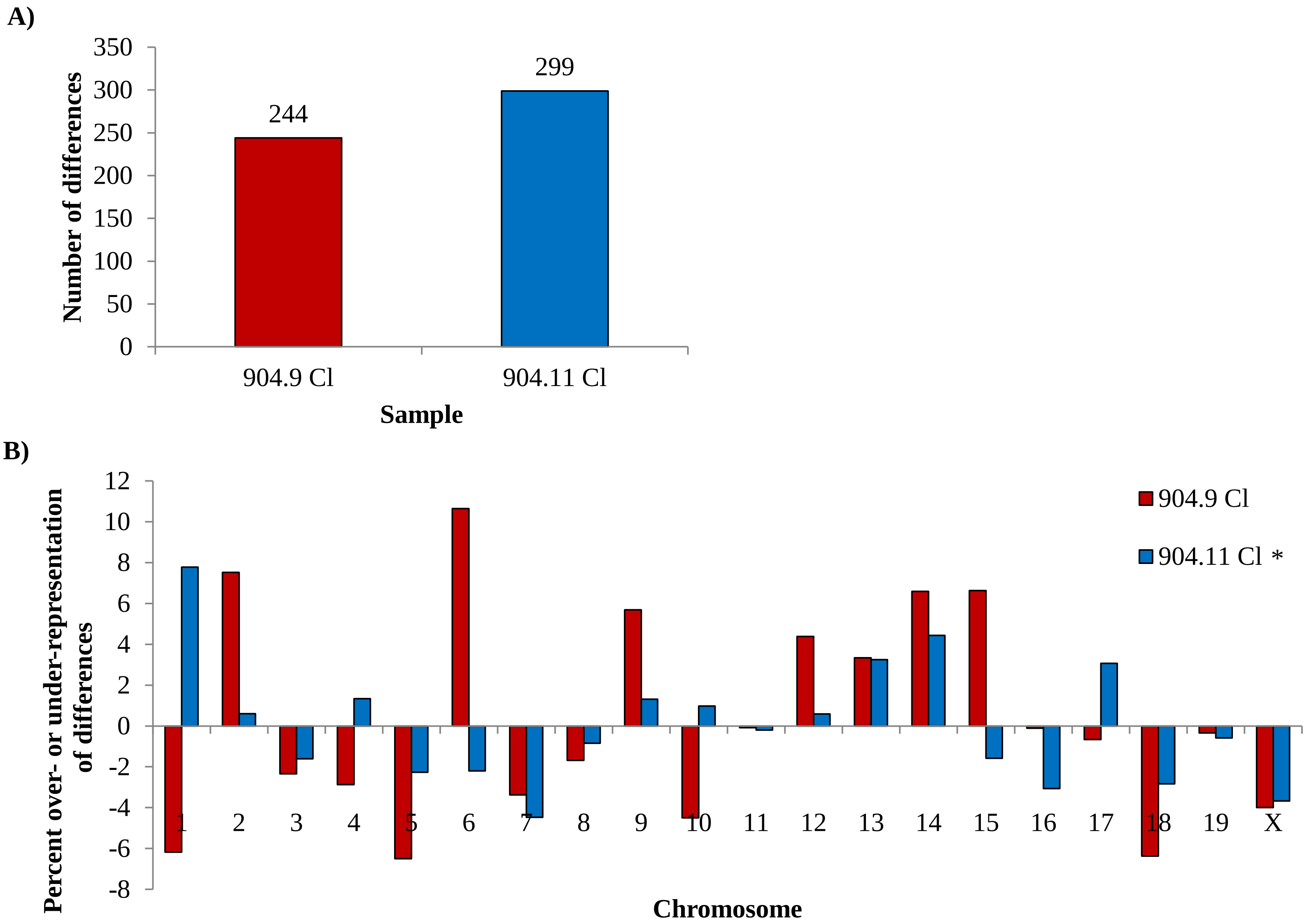
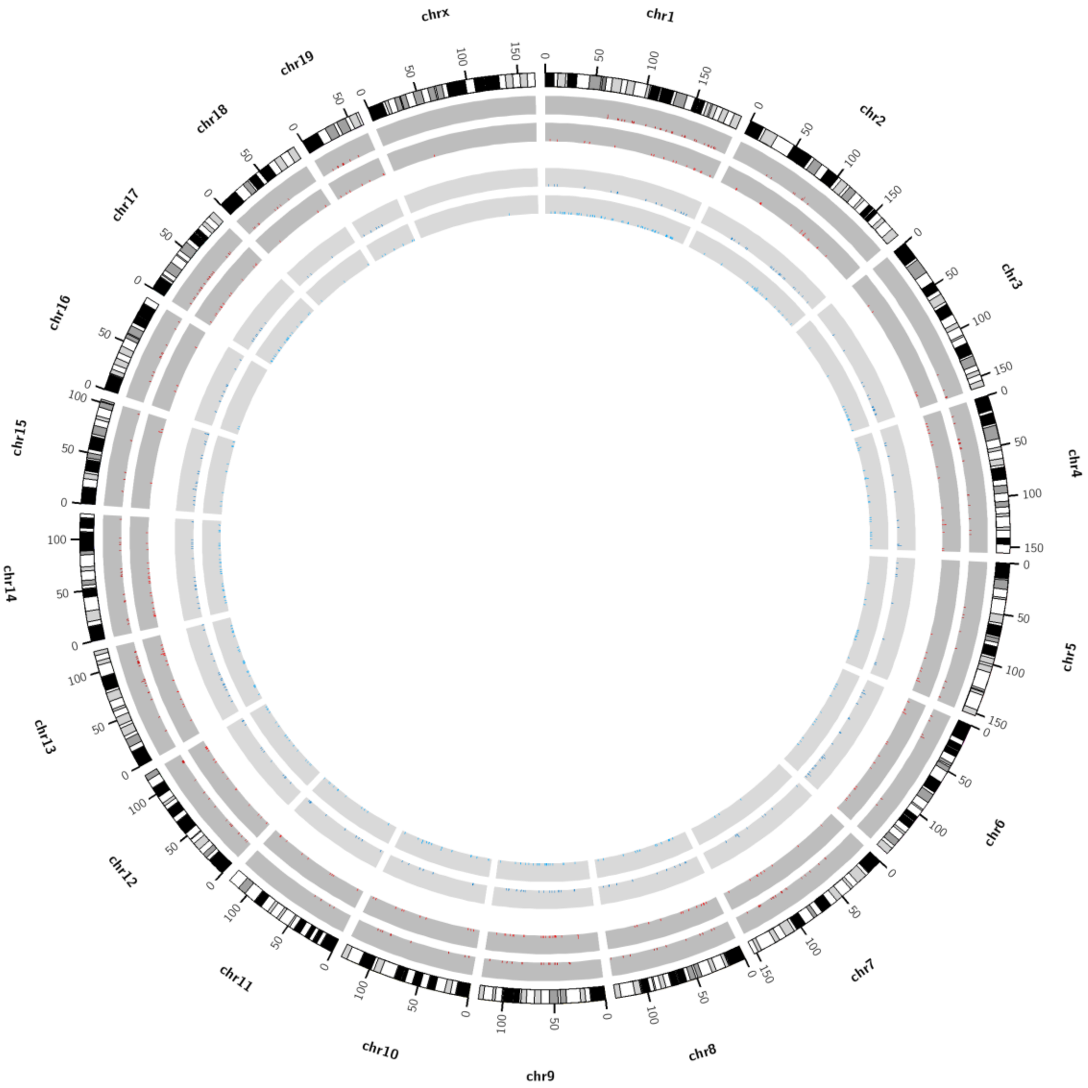

A)
Number of differences
250
245
240
235
230
225
220
215
210
246
225
904.9 Sp
904.11 Sp
Sample
B)
Percent over-or under-representation of differences
15
10
5
0
-5
-10
-15
🟥 904.9 Sp *
🟦 904.11 Sp *
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome

Figure 3-36: Cerebellum-specific differences between the two *harlequin* mice. A) The total number of cerebellum-specific differences in the cerebellum for the two *hq* mice. **B)** The distribution of observed cerebellum-specific differences in genotype graphed to show instances of over- and under-representation from a random occurrence based on the number of genotype calls per chromosome for two *hq* mice. * indicates samples with a distribution significantly different from the expected random distribution (p<0.005).


A)
Number of differences
350
300
250
200
150
100
50
0
244
299
904.9 Cl
904.11 Cl
Sample
B)
Percent over- or under-representation of differences
12
10
8
6
4
2
0
-2
-4
-6
-8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome
904.9 Cl
904.11 Cl *

significantly different from random for mouse 904.9 and mouse 904.11 ($p<0.001$, $p<0.01$ respectively; Fisher's Exact Test MCS). There are too few spleen-specific and cerebellum-specific differences for mouse 904.9 and 904.11 to identify any regions prone to accumulate genotype differences (Figure 3-37).

Figure 3-37: The distribution of spleen-specific and cerebellum-specific differences in the *harlequin* mouse across the chromosomes. Spleen-specific differences are shown in the outer-most rings (grey) and cerebellum-specific differences are shown in the inner-most rings (light grey). Mouse 904.9 is shown in dark red (spleen) and dark blue (cerebellum). Mouse 904.11 is shown in light red (spleen) and light blue (cerebellum).


chr1
0
50
100
150
chr2
0
50
100
150
chr3
0
50
100
150
chr4
0
50
100
150
chr5
0
50
100
150
chr6
0
50
100
chr7
0
50
100
150
chr8
0
50
100
chr9
0
50
100
chr10
0
50
100
chr11
0
50
100
chr12
0
50
100
chr13
0
50
100
chr14
0
50
100
chr15
0
50
100
chr16
0
50
chr17
0
50
chr18
0
50
chr19
0
50
chrx
0
50
100
150

Chapter 4 : Discussion

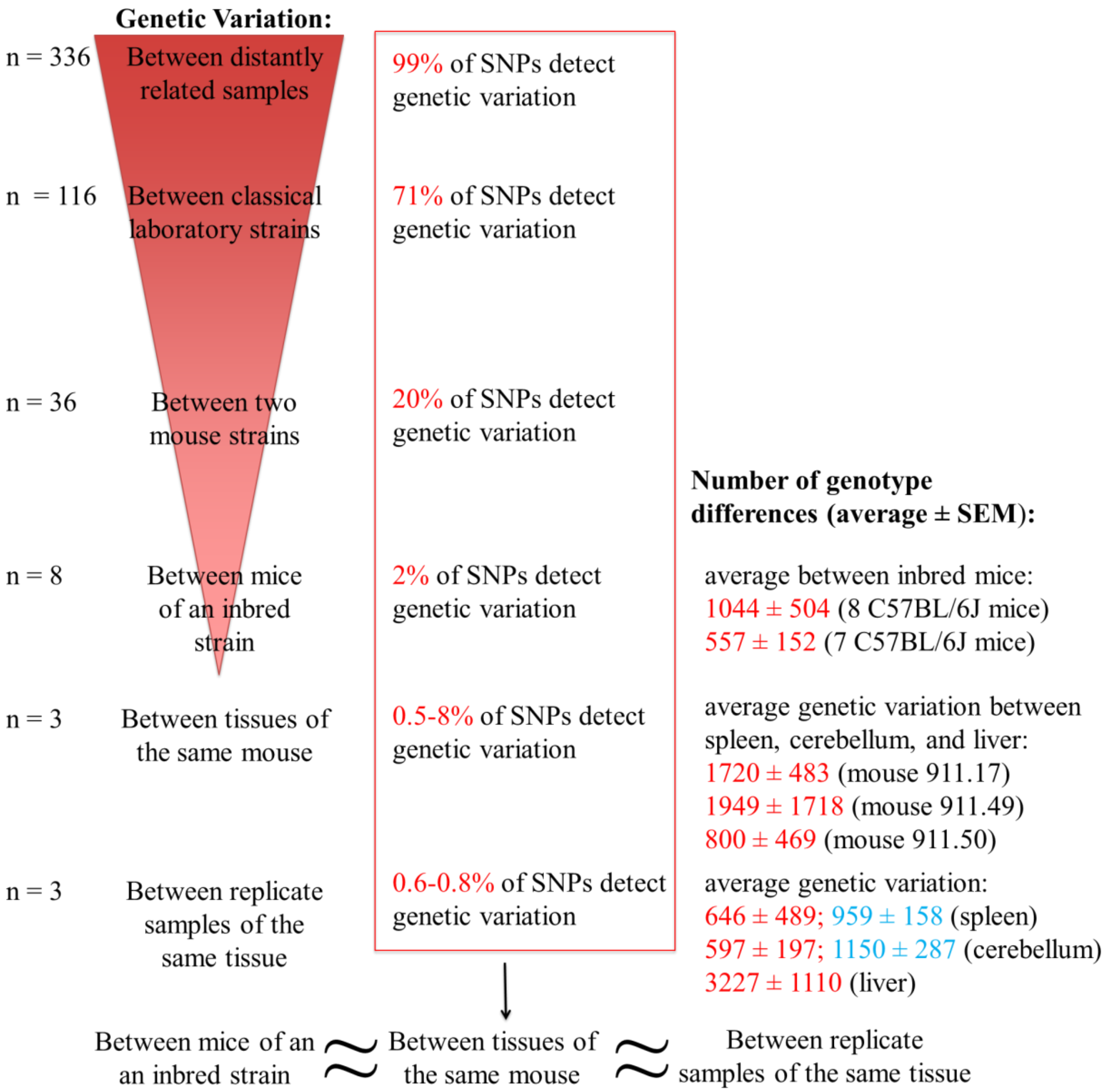
4.1 General Discussion

The MDGA can detect genetic variation from the global perspective of known diversity existing between different *Mus musculus* subspecies, to the fine scale detection of genetic variation within a single tissue (Figure 4-1). The MDGA can accurately distinguish between samples containing high levels of genetic diversity. However, when applied to samples that contain little genetic variation, the MDGA falls short. Currently, the small sample sizes and high variation in SNP differences between samples result in low power for interpreting the extent of variability within tissues of the same mouse. To determine the sensitivity of the array, and identify to what scale genetic variation can accurately be detected, the error rate for the array must be empirically determined to identify all sources of variation.

4.2 The 99,802 probe sets with design flaws in the original SNP list caused technical errors in genotyping calls

Through the filtering process, a total of 99,802 SNPs were removed that either performed poorly as identified by The Jackson Laboratory or Genotyping Console, or that fail to meet basic design criteria. The removal of probe sets that did not meet basic design criteria increased the overall genotyping accuracy (number of samples passing a specified genotyping threshold, such as 97%) of the microarray. Probe sets that are not designed according to specific criteria introduce technical error into the genotyping results. By removing probes that do not meet optimal design criteria, technical errors (specifically false positives) can be reduced. False positive genotypes occur when a SNP is assigned a genotyping call that is different from its true genotype. For example, a false positive occurs

Figure 4-1: The amount of genetic variation detectable using the Mouse Diversity Genotyping Array between distantly related samples to replicates within a single tissue. The amount of genetic variation detectable using the MDGA in terms of the total number of SNPs from distantly to closely related samples. Genetic variation is detected between samples ranging in genetic diversity between distantly related samples to ranging in genetic diversity within a single mouse as between tissues and within a single tissue. The average number of genotype differences ± the standard error of the mean is indicated on the far right. The average number of genetic differences between mice, between tissues, and within a tissue is shown in red. Numbers in blue indicate the average number of genetic differences determined from tissue replicates of a B6 mouse.



Genetic Variation:

n = 336 Between distantly related samples — 99% of SNPs detect genetic variation

n = 116 Between classical laboratory strains — 71% of SNPs detect genetic variation

n = 36 Between two mouse strains — 20% of SNPs detect genetic variation

n = 8 Between mice of an inbred strain — 2% of SNPs detect genetic variation

n = 3 Between tissues of the same mouse — 0.5-8% of SNPs detect genetic variation

n = 3 Between replicate samples of the same tissue — 0.6-0.8% of SNPs detect genetic variation

Number of genotype differences (average ± SEM):

average between inbred mice:
1044 ± 504 (8 C57BL/6J mice)
557 ± 152 (7 C57BL/6J mice)

average genetic variation between spleen, cerebellum, and liver:
1720 ± 483 (mouse 911.17)
1949 ± 1718 (mouse 911.49)
800 ± 469 (mouse 911.50)

average genetic variation:
646 ± 489; 959 ± 158 (spleen)
597 ± 197; 1150 ± 287 (cerebellum)
3227 ± 1110 (liver)

Between mice of an an inbred strain $\approx$ Between tissues of the same mouse $\approx$ Between replicate samples of the same tissue when a sample has a homozygous A genotype and the array determines the genotype of that samples is a NoCall.

4.3 Accuracy of the filtered SNP list causes more samples to pass the genotyping threshold of 97%

The number of samples that pass the genotyping threshold of 97% increased with the use of the filtered SNP list. An additional 35 samples that failed using the original SNP list can be used for further analysis when genotyping is performed with the filtered SNP list. This increase in sample size indicates that the majority of samples failing when genotyped with the original SNP list did so as a result of technical error on the array rather than true biological variation between samples. Samples fail when three percent or more of the genotyping calls are NoCalls. Probe sets that were designed poorly caused samples to accumulate more NoCalls. NoCalls caused by poor probe design resulted in the detection of false positives. Therefore, removing probes that do not meet design standards reduces the number of false positives detected using the MDGA. The removal of false positive genotype calls reduces the noise on the microarray and improves the genotyping ability.

The 35 additional samples that failed when genotyped with the original SNP list did so as a result of technical error introduced from probe design flaws. To test if probe design resulted in these additional failures between the original and filtered SNP lists, twenty randomly filtered SNP lists were used to genotype the dataset. Using these randomly filtered SNP lists, more than 50 samples failed the first round of genotyping. Because the same number of samples failed when the dataset was genotyped with the original and randomly filtered SNP lists, the reduced number of failures when using the filtered SNP list was not due to a reduction in the number of SNPs used when genotyping. The increase in the number of passing samples was caused by the removal of over 99,000 probe sets that were flawed in their design. The SNPs removed contributed to false positives that were detected when genotyping with the original SNP list and resulted in the failure of the additional 35 samples.

4.4 High genetic variation caused samples to fail when using the filtered SNP list

Samples that failed using the filtered SNP list included samples from classical laboratory strains, wild-derived strains, and wild caught mice. The majority of samples failing were those that are very distantly related to a B6 mouse. These strains include alleles from different subspecies of *Mus musculus*. Since alleles from subspecies other than *Mus musculus domesticus* are poorly represented on the array, it is likely that the alleles within each of the failed samples were not detectable. When genotyping calls are made, a minimum of 97% of the calls must be an AA, AB, or BB genotype. However, when the alleles that are present in a sample cannot be detected by either allele that is represented on the MDGA, this minimum of 97% genotype calling may not be reached. When few alleles present in a sample are detectable on the array, the sample is too genetically diverse from a B6 mouse to be included in analyses when the genotyping cutoff has been set so high. Because the samples that likely failed due to high levels of genetic diversity had genetic distance values greater than 90%, when studying mouse strains that are very distant from a B6 mouse, the minimum cutoff should be lowered to roughly 90% in order to incorporate the variation of these diverse samples. The cutoff threshold will be subject to change depending on the genetic diversity within a sample set.

The classical laboratory strains that failed to reach the minimum threshold did not fail because of errors in design. These samples represented strains from some of the most genetically diverse classical laboratory strains available and include the NONcNZO10/LtJ, Ma/MyJ, ATEB/LeJ, and A/HeJ strains. These strains are all derived from primarily *Mus musculus domesticus* subspecies; however, they are distant from a B6 mouse. The

NONcNZO10/LtJ mouse is actually a congenic mouse on a NON/J genetic background. Mice within this particular strain contain alleles from mice originally obtained in New Zealand and therefore were very geographically distant from the North American B6 mouse. MA/MyJ is a inbred mouse strain that contains a mutation in the *hepatic fusion* gene and has previously genotyped with many SNP differences from a B6 mouse.[49] ATEB/LeJ is also a spontaneous mutant strain homozygous for a mutation in the *Glutamine receptor interacting protein.* If containing a mutation in a gene causes an increase in mutation frequency elsewhere in the mouse genome, it is likely to impact how these samples genotype. A/HeJ are albino mice that were founded in North America and they are susceptible to a variety of diseases including cancers.[45] As a result, mutations may accumulate at a higher rate within the A/HeJ mouse. The accumulation of mutations across the genome, may affect areas where SNPs are detectable with the MDGA. As a result, when samples that have mutator phenotypes are genotyped, samples may appear to have a higher proportion of NoCalls due to off target mutations causing a lack of DNA hybridization to the MDGA. The increased proportion of NoCalls in these samples will make mice appear more genetically distant than would be expected considering their genealogy. Therefore, when genotyping mutant mice, or mice that are highly susceptible to mutations, the call rate threshold may also need to be lowered.

4.5 Most of the SNPs in the filtered SNP list are informative when detecting genetic variation within *Mus musculus*

The genetic variation that is detectable with the array refers to the total number of informative SNPs that are capable of detecting variation within *Mus musculus*. Of the 336 samples genotyped from The Jackson Laboratory dataset, over 99% of SNPs in the filtered list of SNPs detected one or more genetic variants between genotype calls across the autosomes. Therefore, the vast majority of SNPs are contributing to the analysis of overall genotypic differences observed between distantly and closely related mice.

4.6 Future directions for SNP list filtering

Further SNP list filtering should be completed to remove probe sets that contain any overlap between probe sequences, as there may be competition between digested DNA fragments. My initial probe filtering failed to remove probes that overlapped by one nucleotide at the 3' end. Although this slight overlap is unlikely to have an effect on genotyping because probe overlap is most concerning when it occurs in the center of the probe sequence, I recommend that an additional 948 SNPs be removed from future analyses. Furthermore, I only removed one of the two SNPs that were affected by the overlap. To increase the accuracy of the genotyping calls even more, both the SNPs affected by the SNP overlap should be removed prior to genotyping. At the very least, the SNPs still containing one nucleotide overlap with another probe set on the array, as well as those that were capable of detecting the SNP that was removed based on overlap should be flagged and watched for genotyping inconsistencies (between samples). Further filtering can be completed based on probe proximity and template availability by removing probe sets that are competing for template. This can be accomplished by determining the minimum distance (bps) that probe sets should be spaced in order for a DNA fragment to bind to a single probe set. To do this, the average fragment size of the DNA prior to hybridization to the MDGA must be determined empirically. This average fragment size can then be compared to the spacing between probes in the original SNP list to identify consecutive probe sets that have a spacing smaller than the average fragment size. For filtering, the full list of SNP probes must be used because although the SNPs are removed from data analysis, they are still present on the array and will compete for DNA template.

4.7 Genetic distance measures distinguish between 364 distantly and closely related mouse strains

4.7.1 Genetic distance measures between 364 samples represent known mouse phylogenies

Genetic distance measures calculated using the MDGA could differentiate between mice of different subspecies and of different strains. The MDGA was designed with the intent of genotyping mice that contain more genetic diversity that traditional laboratory mouse strains; therefore we can use known phylogenetic relationships to assess the accuracy of MDGA genotyping. Alleles detectable on the MDGA favour those present in the B6 inbred mouse strain. Therefore, mice that are distant from a B6 mouse strain are going to appear more genetically distant when genotyped with the MDGA.

The ability of the MDGA to distinguish between distantly and closely related mouse strains reconstructs known phylogenetic relationships of the mouse and allows for the distinction between different mouse strains.[44,45,129] Distantly related samples, including those within classical laboratory strains, that are distant from a B6 mouse, are expected to genotype with a lower overall genotype call rate. As samples become more distantly related from classical laboratory strains they begin to incorporate alleles from different mouse subspecies including *Mus musculus musculus* and *Mus musculus castaneus.* As a result, genetic distance measures for these samples increases. The increase in genetic distance results from an increased representation of alleles (or increased genetic diversity) derived from mouse subspecies that are not well represented on the MDGA. The lack of representation for these alleles causes an increase in the number of NoCalls when genotyping and results in overall lower genotyping call rates for genetically diverse samples.

4.7.2 Globally, the original and filtered SNP lists have equal ability to distinguish between mouse strains

Genetic distance matrices created from pairwise comparisons for the original and filtered SNP list are very similar from a global perspective. This indicates that the overall outcome of the phylogenetic tree does not change when the probes were removed. With no significant differences between the genetic distance values, it can be concluded that the MDGA has the capability of detecting differences in genotype when samples are compared between subspecies and between strains. The MDGA can be used for its original purpose of distinguishing between mice that contain different genetic backgrounds using either the original or filtered SNP lists.

4.7.3 Genetic distance measures calculated in comparison to a homozygous A reference strain can distinguish between distantly and closely related strains

Genetic distance measures calculated from a homozygous A reference strain distinguish between distantly and closely related mice. Genetic distance in comparison to a homozygous A reference determines how similar a sample is to a B6 mouse strain. Using these genetic distance measures, mice within each of the sample types were ordered based on their known phylogenetic and genealogical relationships to a B6 mouse. Mice within each of the sample types that are genetically distant from a B6 mouse had a genetic distance measure that was closest to 1. Mice within each of the sample types that were known to be genetically similar to a B6 mouse genotyped with a genetic distance measure closest to 0. Therefore, when comparing mice to a reference strain, the genetic distance measures can be used to identify samples that are closely and distantly related.

4.7.4 Genetic distance values calculated with the filtered SNP list refine the genetic distance values to better reflect known relationships to a B6 mouse

Genetic distance measures calculated from a reference homozygous A strain were similar between the original and filtered SNP lists. When the range of genetic distance values was compared between the SNP lists for each of the sample types, the differences in genetic distance values observed between the SNPs for each of the sample types reflected known relationships to the B6 mouse.[45] Using the filtered SNP list, mice that are known to be distantly related to a B6 mouse have a genetic distance that is closer to 1 when compared with the genetic distance value determined using the original SNP list. Likewise, mice that are known to be genetically similar to a B6 mouse have a genetic distance closer to 0 when compared to values calculated using the original SNP list. Therefore, based on known phylogenetic and genealogical relationships for each of the strain types used in The Jackson Laboratory dataset, it can be concluded that the change in genetic distance values better match what would be expected.

The increase in genetic distance values observed for the minimum, maximum, and average genetic distances in wild-derived laboratory mice (with an exception for the minimum genetic distance), F1 hybrids, CC, and wild caught mice reflect the increased genetic diversity of these samples. Mice within these sample types contain more genetic variation than classical laboratory strains. This is because wild-derived, F1, CC, and wild caught mice contain alleles that are derived from the *Mus musculus musculus* and *Mus musculus castaneus* subspecies. These subspecies are not well represented in classical laboratory mice, and as a result have poor representation on the MDGA.[107] Therefore, when genotyping samples containing many alleles from the *Mus musculus musculus* and *Mus musculus castaneus* subspecies, there will be a higher proportion of NoCalls. This higher proportion of NoCalls caused by the lack of representation of these subspecies on the MDGA ultimately causes samples to have a genetic distance closer to 1. As mice get more genetically distant from a B6 mouse (closer to 1), the constantly increasing number of NoCalls makes it increasingly difficult to differentiate between samples.

The minimum genetic distance for wild-derived laboratory strains decreased. This decrease in genetic distance value can be explained by identifying the mouse strain at the lowest genetic distance range of this sample type. The mouse strain that contained the smallest genetic distance value in wild-derived laboratory strains was the CALB/RkJ mouse. This strain was primarily derived from the *Mus musculus domesticus* subspecies, indicating that in comparison to other wild-derived strains, the CALB/RkJ mouse strain is more genetically similar to a B6 mouse. Because many of the SNPs on the array distinguish between alleles that are present in the *Mus musculus domesticus* subspecies, the genotyping call rate for samples containing alleles derived from this subspecies should be high. When a sample contains a high proportion of its genetic background from the *Mus musculus domesticus* subspecies, the genetic distance should be closer to 0. The removal of probe sets that contained design flaws reduced the number of false positives that were contributing to an increase in genetic distance. The removal of these probe sets resulted in an increased genetic distance value for CALB/RkJ mice as would be expected based on genetic background.

The genetic distance values calculated with the filtered SNP list for the BXD mice moved closer towards 0 for both the minimum and maximum values. This shift of both values towards 0 is expected, as BXD is a recombinant inbred mouse strain that contains the genetic background of the classical laboratory strains B6 and DBA/2J. Therefore, BXD mice are genetically similar to a B6 mouse and should genotype with genetic distances very close to the reference.

Consomic mice contain genetic variation from other mouse strains and would therefore be expected to have genetic diversity higher than that of the isogenic founder strain. In this study, all consomic mice are on a B6 genetic background and contain a single chromosome from the wild-derived mouse strain PWD/PhJ-ForeJ that belongs to the *Mus musculus musculus* subspecies. By introducing genetic variation from a different subspecies of mouse, consomic mice are now less related to the reference in comparison to a pure B6 mouse. Therefore, the shift in genetic distance values for consomic mice was expected to shift towards 1, which was observed for the maximum genetic distance.

Classical laboratory mouse strains encompass the genetic variation found within the *Mus musculus domesticus* subspecies for the majority of laboratory strains. Mice within the classical laboratory strain type will therefore have genetic distance values very close to 0, as they are the most related to the reference. However, classical laboratory mice will also have genetic distance values approaching 1 because of the variation between the different mouse subspecies that are found within some laboratory strains. The shift in the minimum genetic distance value closer to 0 therefore matches what would be expected based on genetic background for each of the strains that are derived from *Mus musculus domesticus*.

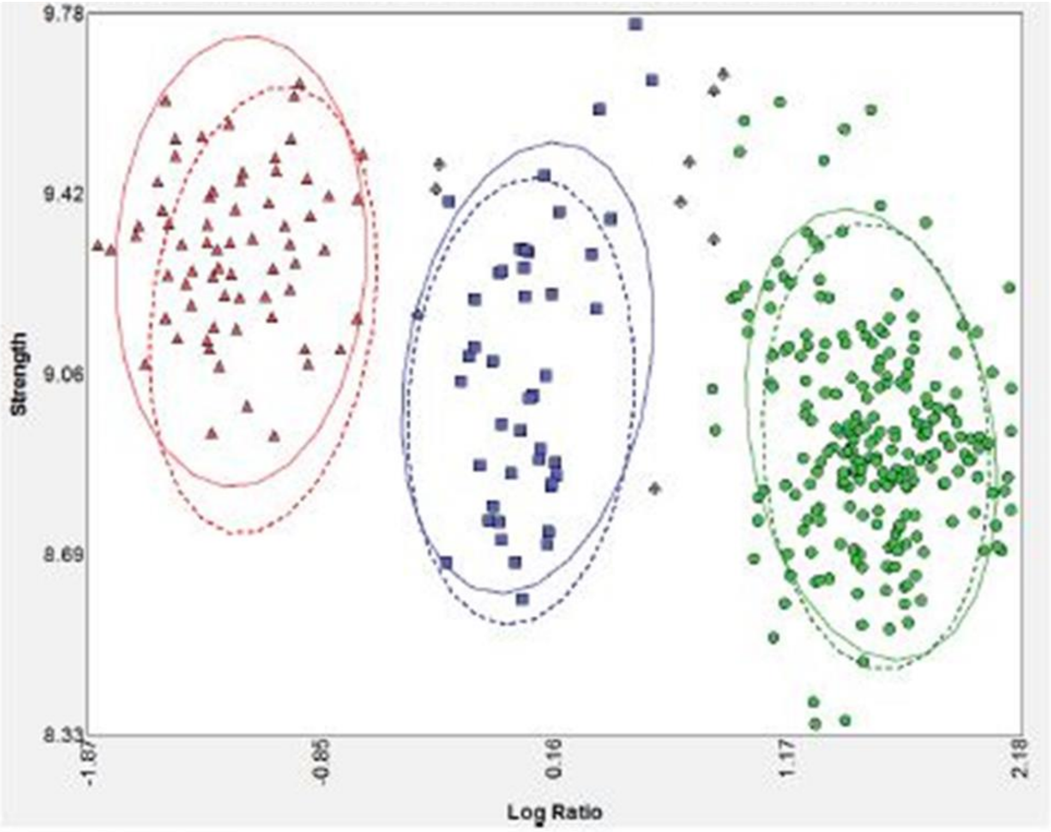
4.7.4.1 The filtered SNP list was better able to distinguish between genetically diverse samples using genetic distance measures in comparison to the original SNP list

When genetic distance measures were compared between the original and the filtered SNP lists, the genetic distance values calculated for the most diverse samples in the sample set, the F1 and CC mice were significantly different. This means that the genetic distance values calculated with the filtered SNP list were more sensitive at detecting genetic distance values within highly diverse mice as compared to the original SNP list. Genetic distance values for the F1 and the CC mice had the maximum and average genetic distance values increased. The increase in genetic distance values reflects the increased genetic diversity within these samples.[44,65] By increasing the genetic variation within a mouse, it becomes more difficult to genotype accurately. This becomes exacerbated when probes do not perform well as a result of their design. This further reduces the accuracy of genotyping calls. However, when the probes are removed that do not meet design criteria, a significant improvement to the genotyping accuracy of the MDGA is made. Samples containing high amounts of genetic variability are genotyped more accurately, which increases the confidence of genotyping calls for genetically diverse samples.

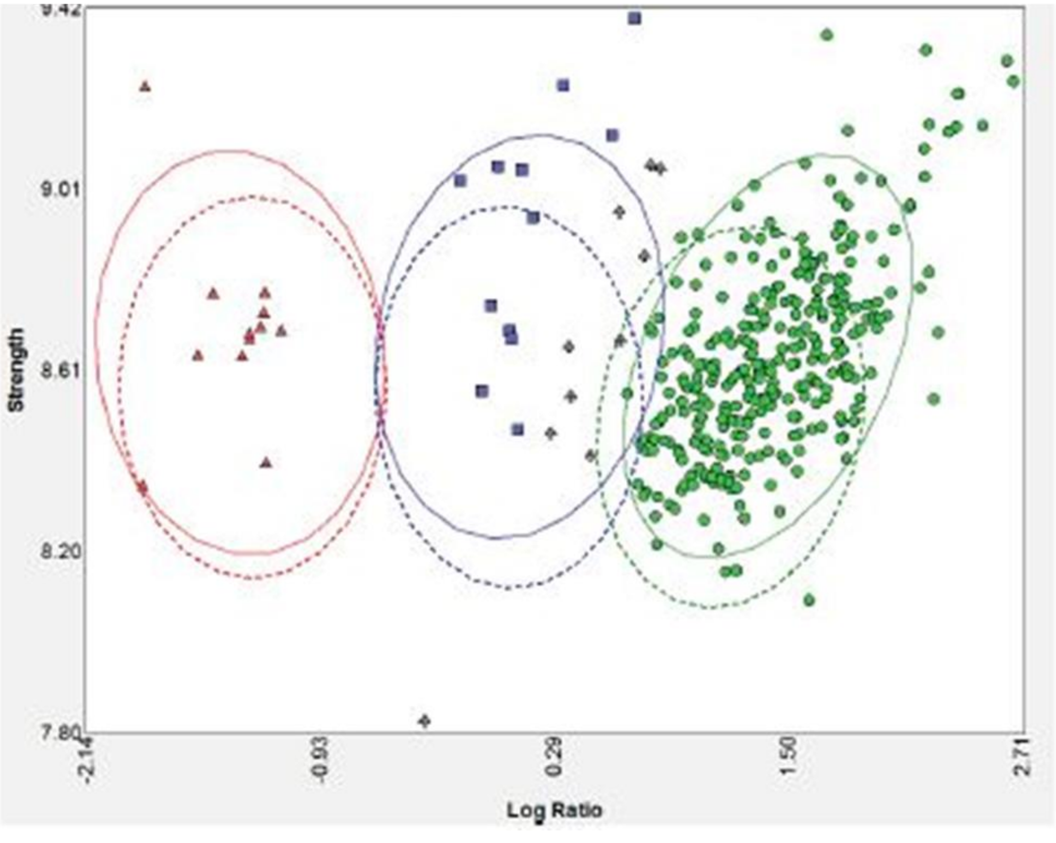
Inaccuracy in genotype calling is particularly high for samples that contain high levels of heterozygosity.[108,117] When assigning a SNP to a specific genotype, the fluorescence intensities are used to determine whether a sample is likely homozygous or heterozygous. This information is based on predefined clusters that predict the intensities for each of the three genotype possibilities (Figure 4-2). The more samples that genotype within a specific cluster increases the confidence of genotyping calls within that particular group. Therefore, when datasets, such as The Jackson Laboratory dataset contain a high number of isogenic strains, genotyping heterozygosity is less accurate.[108,130] When genotyping clusters do not contain a high number of samples, the cluster range increases to accommodate those samples that are most likely to be that genotype. As a result genotype calls associated with this cluster become less reliable. Because poorly designed probes have poor hybridization and altered fluorescence intensities, it makes it more difficult to assign a genotype to the SNP. Therefore, by removing probes that contain design flaws, genotyping becomes more accurate, especially for those samples that contain high levels of heterozygosity.

Figure 4-2: Predefined clusters for genotyping each of the SNPs. There are three predefined clusters that associate with the three genotype calls, AA (green), AB (blue), BB (red). NoCalls in each of the graphs are represented in grey. These three clusters each have a predefined circle (dotted lines) that is based off the expected area where the fluorescence intensities will be graphed. As samples are genotyped, each of the clusters adjusts to better fit the data set (solid lines). **A)** Clustering where the clusters are well separated and most of the genotype calls fall within the range of the adjusted circle. **B)** Clustering where the clusters are poorly separated and the pre-defined clusters overlap and genotyping is difficult to perform.

A)


9.78
9.42
9.06
8.69
8.33
Strength
-1.87
-0.85
0.16
1.17
2.18
Log Ratio

B)


9.42
9.01
8.61
8.20
7.80
Strength
-2.14
-0.93
0.29
1.50
2.71
Log Ratio

F1 hybrids contain the highest amount of heterozygosity within the Jackson Laboratory dataset. F1 hybrid mice contain one half of their genetic background from one mouse strain and the second half of their genetic background from a different mouse strain. The genetic distance values generated for each of the F1 hybrid mice are indicative of the parental strains for each of the samples. Meaning that the genetic distance values for F1 mice directly reflect the strains from which the parents belong. When an F1 mouse has a parent that is from a wild-derived strain, the genetic distance value increases and when both parents of an F1 mouse are wild-derived laboratory strains the genetic distance value increases even more. This is because wild-derived strains are primarily members of the *Mus musculus musculus* and *Mus musculus castaneus* subspecies. These subspecies have poor representation on the array therefore, the genetic distance values would be expected to increase when poorly performing SNPs were removed. Similarly, when both the parents were classical laboratory strains the genetic distance values decreased. The genetic distance value for F1 hybrids is expected to be near the upper range of values calculated for classical laboratory strains because classical laboratory strains contain alleles derived primarily from the *Mus musculus domesticus* subspecies. Therefore, most of the alleles present within the F1 mouse should be detectable on the array when both parents are classical laboratory strains.

Collaborative cross mice, although primarily homozygous across the genome within an individual line, contain very high levels of genetic variation across CC lines because they incorporate alleles from eight different founder populations.[67] The genetic distance values for CC mice from different CC lines would therefore be expected to be higher than classical laboratory strains because there is a lot of genetic diversity across the different lines. Of the eight founder strain used to create the CC mouse lines, three were wild-derived laboratory mice. This means that the MDGA is unlikely to detect alleles that were contributed by these strains because of the lack of representation for the *Mus musculus musculus* and *Mus musculus castaneus* subspecies on the microarray. The remaining mice were from the classical laboratory strains, only one of which was very closely related to the reference, the B6 inbred strain. Because the majority of alleles contributing to the CC mouse were not from a B6 genetic background, the genetic distance values for CC mice are expected to be at the higher end of the genetic distance measures. The increase in genetic distance measures when using the filtered SNP list indicates that there is an improvement in detecting genotypes for mice that contains high levels of genetic variation.

Although classical laboratory strains, consomic, congenic, BXD, wild caught, and wild-derived laboratory strains did not show any significant difference in the genetic distance values calculated in comparison to a reference strain of homozygous A genotypes, the removal of SNPs from genotyping analysis increased the accuracy of genotyping. This ability to better distinguish between samples that contain the most genetic variation in the sample set refines the ability of the MDGA to distinguish between strains that contain high levels of genetic variation. Therefore, when samples begin to accumulate heterozygosity within the genome, the filtered SNP list will perform better. This allows for more accurate genotype calls when using mouse samples that mimic the diverse genetic backgrounds of the human population.

4.7.5 Genetic distance can be used to distinguish between mice with different percent admixture between C57BL/6J and CBA/CaJ strains

The MDGA is capable of distinguishing between mice that range from pure B6 to pure CBA/CaJ genetic backgrounds using a variety of different ordering methods. Genetic distance measures can be used to differentiate between samples and compared to breeding records. For genetic distance measures determined between paired samples and in comparison to an AA reference correctly grouped each sample based on the percentage of B6 and CBA/CaJ genetic backgrounds that were determined from their genealogical records. Mice genotyping with the highest genetic distance values were those of a pure CBA/CaJ genetic background. As mice are continually bred from a B6 to a CBA/CaJ genetic background, the number of CBA/CaJ alleles within a sample increases. Therefore, as mice become more CBA/CaJ, the genetic distance values increase. Although these two mouse strains are both classical laboratory strains the genetic distance value increases when a sample is bred to a CBA/CaJ genetic background because the MDGA is biased to detecting the genetic background of a B6 mouse.[107] Therefore, it was expected that the samples with the highest percentage B6 genetic background would genotype with a genetic distance value closest to 0. Because the MDGA was able to differentiate between samples containing varying amounts of B6 and CBA/CaJ genetic backgrounds, the MDGA can be used to distinguish between closely related mice.[45] The MDGA can even be used to differentiate between mice within a small pedigree.

4.7.6 The percentage of C57BL/6J and CBA/CaJ genotypes can be used to differentiate between mice within a family pedigree

The MDGA can also be used to detect genetic variation existing between an admixture of B6 and CBA/CaJ genetic backgrounds. By determining the total number of genotype calls detecting the B6 and CBA/CaJ mouse strains, mice can be correctly identified based on their admixture and known percentage of B6 genetic background. Mice on a pure CBA/CaJ genetic background will never genotype as 100% CBA/CaJ. This is because the CBA/CaJ and B6 mouse strains share the same allele at over 369,000 SNP locations (more than 70% of SNPs on the array), and will by default be genotyped as a B6 allele due to the MDGA bias towards this strain. The MDGA has sufficient SNPs on the array to distinguish between varying degrees of admixture between the closely related B6 and CBA/CaJ inbred mouse strains.

4.8 Genetic distance measures between C57BL/6J mice are consistent with reports of genetic variation within an isogenic mouse strain

Genetic distance values calculated for B6 mice were the lowest out of the 362 samples. More than 99% of SNPs are capable of detecting a B6 allele in the filtered SNP list. Therefore, B6 mice would be expected to genotype with the lowest genetic distance values because the array was specifically designed to detect genetic variation in comparison to a B6 mouse. The genetic distance values calculated for the B6 mouse were the values closest to 0, indicating that the MDGA is capable of differentiating between samples that were pure B6 and those that were not. However, differences in genetic distance values between mice within the B6 mouse strain, indicates that the MDGA can also distinguish between mice of a single strain.

The variation that was observed between the eight B6 mice shows that there is enough variation between the genotyping calls to cause a significant difference between each of the samples. This can mean one of two things, technical error or biological variation is causing the B6 mice to genotype differently. Samples within a single mouse strain are believed to be isogenic. Because mice of an isogenic strain are genetically identical, when genotyping multiple mice from the same isogenic mouse strain, different mouse samples should be indistinguishable as all genotypes would be the same at each location. However, because differences were observed between the samples there is either variation within an isogenic mouse strain, there is technical error causing these differences, or there is a combination of both phenomena.

4.8.1 Observed variation within the C57BL/6J mouse may be due to variation within the mouse strain.

The variation detected within a mouse strain has not been investigated fully; however, studies suggest that genetic variation does exist within a mouse strain.[30,131] One such study highlights how large scale genetic variants, such as CNVs, are variable between mice of a pure B6 genetic background.[131] This study showed that there is genetic variation between mice of the same isogenic mouse strain. A CNV affecting a 112 kb region located on chromosome 19 in B6 mice was observed between individuals within the B6 inbred strain. This CNV appeared to be inherited as it affected many individuals within a B6 population and was detected in multiple stock populations of frozen embryos in The Jackson Laboratory. Although there is limited research looking at the variation within an isogenic mouse strain, studies on somatic mosaicism indicates that variation does exist between what were once thought to be identical tissues or individuals.[61,73] As a result, mouse strains should not be presumed to be isogenic due to the genetic variation found between samples.

Genetic variation can accumulate within a mouse strain due to mutations. These mutations cause genetic variation within an "isogenic" population and can be inherited (as it affects the germline) or occur as *de novo* mutations and result in somatic mosaicism. Somatic mosaicism is the most studied form of genetic variation affecting inbred mouse strains. Somatic mosaicism refers to the accumulation of genetic differences between cell types or tissue types of an individual. As a result, different cell/tissue types within these individuals no longer contain identical genetic sequences. These differences begin to accumulate shortly after conception, and result in genetic variation between tissues.[60]

In humans, somatic mosaicism has been identified between identical twins. Identical twins were once thought to have genomes that were genetically identical because they develop from the same zygote. However, monozygotic twins have been found to contain genetic variation, particularly in terms of copy number.[96,132,133] Genetic variation begins to accumulate between monozygotic twins very early in development. The most well studied type of variation between twins is CNV. CNVs between twins have been linked to many diseases including multiple sclerosis, schizophrenia, and diabetes. Using twin studies, researchers have identified CNVs associated with each disease in identical twins that are discordant for the diseased phenotype. The difference in copy number between discordant twins reflects the genetic variation that accumulated during development and those differences that were the result of environmental factors.

In mice, somatic mosaicism has traditionally been studied using single gene approaches. These studies have shown that mutations accumulate in a tissue-specific manner, and contribute to the genetic complexity of an individual.[60,61,78] Other variation previously found naturally within the population is associated with immunity and the *Major histocompatibility complex* (*Mhc*).[134] The *Mhc* accumulates mutations over the lifespan of an individual because it is highly plastic and susceptible to modification. The *Mhc* has been shown to contain variation in the expression patterns and copy number between tissues of a B6 mouse.[135] This variation was observed as a duplication in the liver of the B6 mouse in comparison to the cerebellum in adult tissue. This observed variation within an individual makes it impossible for all mice of the same inbred mouse strain to contain identical genetic sequences. These studies have reported the genetic variation is present within a single mouse, and as a result, it is unlikely that the isogeneity of inbred mouse strains is true. Rather, mouse strains may contain genetic variation within and between individuals, while sharing localized regions of isogeneity.

4.8.2 Observed genetic variation within the C57BL/6J mouse strain may be due to technical error

4.8.2.1 Technical error can occur during DNA preparation for microarray hybridization

Technical error can occur as a result of poor microarray design or inconsistencies in the microarray platform. These errors inflate the number of genotype differences detected between samples. Technical errors can also be introduced at a number of different stages during DNA extraction, processing, hybridization, and genotyping. Technical error introduced prior to genotyping can be caused by protocols that are not optimized to extract or prepare the DNA for hybridization. When protocols are not optimized, the DNA can be contaminated with reagents, such as salts and proteins.[136] When DNA contains these contaminants, downstream DNA preparation procedures can be affected, such as digestion, ligation, and PCR amplification, and the resulting hybridization and genotyping calls can be inaccurate. For example, the type of tissue from which DNA is being extracted is important when selecting a DNA extraction procedure. Tissue containing high fat contents such as the cerebellum, or high protein, such as the liver, may need to be modified in order to reduce the amount of protein contamination.[137] Similarly, the conformation of DNA within each of these tissues with respect to open or closed chromatin may affect genotype calls when DNA extraction does not remove all protein. If too much of the DNA is bound up in the histones, and not enough of the proteins are removed from a sample, the DNA that is still wrapped around the histones that are present may not be accessible for genotyping. As a result, SNPs within these inaccessible regions are not going to have any genotype calls associated with them; therefore, resulting in technical errors during genotyping. Because the conformation of

DNA is dependent on gene expression (which is tissue specific), technical errors caused by DNA conformation may occur in a tissue-specific manner.

Technical error can also be introduced due to contamination with salt.[138] High salt concentrations reduce the optimal hybridization conditions for an array as it affects how DNA binds to the probe sequence. Contamination with salt is most likely to occur during DNA extraction but can also be introduced in buffers added during steps after DNA extraction. When too much salt is present in samples ready for microarray hybridization, it can cause non-specific binding of the DNA to the microarray. When DNA binds non-specifically to probe sequences, it causes the genotypes to be unreliable because the probe no longer is detecting one specific sequence of the genome.

During the processing of DNA, errors can again be introduced if the digestion is not completed. When the DNA digestion is not a full digestion, the fragment sizes will be too long for PCR amplification. This will cause no DNA template to be available during hybridization because it could not be amplified. DNA amplification can also introduce errors because mutations can occur from DNA polymerase. DNA polymerase adds errors at a rate of 10^{-5} errors/pb/duplication.[139] These errors can affect hybridization and genotyping calls. Finally, after hybridization, errors can be introduced from the genotype calling algorithm.[108] Because genotyping is based on fluorescence intensities, errors in clustering these intensities can result in an incorrect genotype call. When many samples contain the same genotype call, the cluster for that particular genotype call becomes more stringent. However, when few samples contain a specific genotype call, the cluster for that genotype is more flexible. Genotype clusters are calculated based on the number of samples in the proximity of a pre-defined area.[108,130] When few samples are within this area, the cluster becomes less defined and therefore, sample that are close to the pre-defined cluster but not within the area have a tendency to be called as that genotype.

4.8.2.2 Technical error can result from imperfections on the microarray

Technical error can also be caused by an imperfection in the design of the microarray. In the case of microarray technologies, these imperfections can manifest as inconsistencies in probe design. When probes are not designed to the same standards on microarrays, I have demonstrated that the genotyping accuracy for that microarray decreases. Because microarrays are designed to detect genotype calls at over hundreds of thousands of locations across the genome, probes detecting these locations are designed using a strict set of criteria. By designing probes with such strict criteria, they function optimally within the same temperature ranges. Therefore, probes that have flaws in their design impact the overall genotyping accuracy of the array because they have a tendency to produce failing genotype calls. Additionally, when manufacturing microarrays, artifacts occurring during preparation, such as scratches, can affect the hybridization of probes and result in poor genotyping calls.

4.8.3 The distribution of differences detected between C57BL/6J mice may in part be the result of technical error

Microarrays have high rates of false positive and false negative errors, typically in the range of 9-20%.[140,141] Hybridization technologies also have poor reproducibility, with some studies indicating as little as 30-40% correlation within inbred strains.[142] The high error rates and poor reproducibility of microarray technologies may be associated with the random failure of probes on the microarray.[142,143] Therefore, it is reasonable to expect genotype differences to occur at random across the genome in proportion to the number of SNPs representing each chromosome. The random distribution of genotype differences detected between B6 mice, lends itself to the hypothesis that the observed genotype differences between the samples may, in part, be due to the random failure of probe sets on the MDGA.

The accumulation of genotype differences across the genome would not be expected to be uniform based on sequence context. There are known sequence contexts, such as CpG dinucleotides and tandem repeats that accumulate mutations.[60,73,77] Previous research using transgenic mutation detection approaches discovered that mutations occur at a frequency of 10^{-6} to 10^{-5} mutations per base pair per cell generation in somatic genomes,[76,144,145] translating to a mutation rate of roughly 10^{-8} mutations per base pair per year.[59,146,147] Although this mutation rate is conventionally applied across the genome, it only represents selectively neutral (containing no genes) regions across the genome. Mutations that lead to phenotypes with lower fitness would be expected to be lost, due to negative selection pressures. Selection can therefore lead to a non-random distribution of mutations across the genome. Genomic plasticity has been observed and associated with certain genes. [7,8] *Mhc* and olfactory genes are examples of higher mutations due to positive selective pressures. Mutations on the X chromosome reflect strong negative selective pressures. Technical errors in genotypes may mask the true landscape of mutations across the genome.

4.8.4 True biological variants likely contribute to genotype differences detected in multiple C57BL/6J mice

The majority of genotype differences among B6 mice were the result of one sample having a different genotyping call from the remaining seven. There is greater confidence in calling the variant biologically relevant if more than one of the samples contains the alternate genotyping call. When the eight B6 mice are evenly split between two genotype calls, it is likely to represent variation within the B6 mouse strain. This is because, it is less likely for three or more sample to have the same technical error detected. Whereas, when only one or two samples contain a different genotype from the remainder of the B6 samples, it is more likely that these were just one off technical errors in each of the samples. Although variation could be detected between eight B6 mice, further research needs to be completed to increase the sample size and sequencing needs to be completed to identify where true genetic variation is present within the B6 mouse strain.

4.8.5 Genetic differences detected within inbred and wild caught mouse strains may be true biological variation

When wild caught mice were compared to each other, the majority of SNPs used in the comparison showed differences in genotype calls between samples. This would be expected because wild caught mice contain genetic variation from mouse subspecies that are different from the subspecies predominant in classical laboratory strains. However, there are regions of the genome that showed variation within wild caught mice and within B6 mice. Such areas are likely accumulating genetic differences due to *de novo* mutational events. The accumulation of new mutations within each of these mouse types is likely explanation because of the geographic separation and mouse subspecies of the wild caught and B6 samples. Since samples belong to different subspecies and are separated by entire continents (due to the comparisons with the North American B6 mouse), mutations that appear to be shared among these samples are likely the result of mutational hotspots within the *Mus musculus* species.

SNPs that share genetic variation within B6 mice and within wild caught mice affect a total of 14 genes, all of which have previously been shown to have genetic variation (MGI). These genes are involved in general functions within the mouse, primarily associated with binding to nucleic acids or proteins (MGI). Several have more specific activities such as calcium binding of *Capn1*, insulin receptor binding of *Tln2*, and metal ion binding for *Mbnl2* and *Mppel* (MGI). These processes are not obviously indicative of any mutational hotspots that affect genes; as would be found with specific genes that require a lot of genomic plasticity such as those involved in the immune response and the *Major histocompatibility complex* (*Mhc*) or olfaction in wild mice.[135,148]

4.8.6 Interpretations made using inbred mouse studies must now take into account genome-wide variation within a mouse strain.

Inbred mice are used to mimic complex phenotypes in humans. However, with the identification that genetic variation exists within an inbred mouse strain, this variation must now be taken into account. Inbred mice are not as isogenic as they were once thought to be. Although specific genes may not contain variation within a mouse, the remainder of the genome may have variation that affects the outcome of these complex interactions. Therefore, when conducting studies on complex traits researchers must also consider the genome-wide effects of somatic mosaicism. As a result, by using high-resolution technologies such as the MDGA to study complex phenotypes, researchers can gain a more accurate understanding of complex phenotypes and how they are affected by genomic variation.

4.9 The Mouse Diversity Genotyping Array can detect variation between tissues of the C57BL/6J mouse.

The MDGA can detect genotype differences between the spleen and cerebellum of a B6 mouse. This is indicative of somatic mosaicism occurring early in development causing genetic differences between these tissues. Previous research has shown that the spleen and the cerebellum differ in the total number of mutations using single gene mutation detection systems.[60,73,74] The accumulation of mutations within these tissues is specific to their developmental histories. However, because the known differences observed between these two tissues were initially based on single gene mutation detection systems, the genomic variation that exists between these tissues is still unknown. Therefore, the MDGA can be used as the first genome-wide mutation detection system to identify tissue-specific differences between the spleen and the cerebellum of a B6 mouse.

4.9.1 Comparable levels of genotype differences were observed for the spleen and cerebellum of a C57BL/6J mouse.

Replicate samples from a pure B6 mouse tissue show significant variation in the number of genotype differences. Although the total number of differences detected between each of the replicates differs, the distribution of these differences is similar. Each of the replicates within a tissue contains roughly the same number of differences on each of the chromosomes of the mouse genome. These observed differences may be a combination of technical error and biological variation. Because the reproducibility of genotyping results can be as low as 60%,[142] it would not be surprising to find a great deal of technical error contributing to the observed differences. Additionally, a total of 622 SNPs accounting for 0.16% of the genotyping calls for the array detected differences within the spleen replicates and within the cerebellum replicates. Because these differences were detected as a one off genotype differences between two different tissues they are likely genotyping errors and provide a very rough estimate for an error rate associated with the array. Therefore, it is possible that the differences observed between the replicates are in part the result of variation introduced by the inherent limitations of hybridization technologies.

Although there is a strong case for technical error introducing the differences detected between tissue replicates, the biological differences cannot be discounted. Tissues contain many different cell types. When tissue replicates do not contain the same composition of cells, differences may be detected based on these cell types. Each of these cell types will accumulate genetic variation at their own rate.[60,61,73,77] It is the genetic variation between all of the cell types that contribute to the genetic variation of a tissue. Therefore, certain replicates may appear to have more variation if the cell type that was most predominant within the replicate had higher variation in comparison to the predominant cell types of a different replicate.

Differences detected between replicates may be due to large clonal mutations occurring within different regions of the same tissue. For example, cancer is an extreme example of the extent of genetic variation that can arise *de novo* within a tissue. The *de novo* genetic variants that arise in tumor tissues cause the clonal expansion of mutations resulting in greater heterogeneity within tumor samples. Therefore, tissue replicates taken from areas affected by these clonal mutations, will appear to have a higher number of genetic differences in comparison to a tissue replicate that was not affected by the clonal mutation.[58]

A limited sample size makes conclusions about tissue-specific mutation accumulation unreasonable. However, the consistent under-representation of genotype differences on the X chromosome reflects the known negative selection that occurs for this chromosome.[149] In male mice, mutations occurring on the X chromosome are selected against the resulting phenotype because of its hemizygosity in male mice. Without a second copy of the X chromosome, mutations cannot be masked, and as a result are selected against. The biological evidence supporting an under-representation of genotype differences on this chromosome indicates that the MDGA is capable of detecting biologically relevant genotype differences.

4.10 Somatic mosaicism could be detected using the Mouse Diversity Genotyping Array between tissues that have different cell turnover rates

Somatic mosaicism analysis shows that for WT mice the total number of tissue-specific differences in the spleen, cerebellum, and liver is different. This is consistent with analyses of mutation accumulation during development which indicates that tissues differ in mutation frequency.[60,61,73,78,95] Tissues accumulate mutations in a manner that is specific to their developmental timeline. One such factor that contributes to this accumulation is the cell turnover rate for each of the respective tissues.

Cell turnover rate directly impacts the proportion of cells containing a mutated DNA sequence. These mutations arise from DNA damage that was failed to be repaired during DNA replication and cell division. As DNA damage occurs and DNA replication takes place, these changes to the genetic sequence, caused by DNA damage, become fixed within the cell populations of these tissues. Tissues with high replication rates contain mutations at clonal levels since mutations occurring are continually passed on to subsequent cell generations. The highly replicative nature of these tissues makes it easier to detect mutations, as a higher proportion of cells will contain the mutated DNA sequences. Therefore, tissues that contain high cell turnover rates, such as the spleen, are going to have a higher number of mutations as compared to tissues that are post-mitotic.

Post-mitotic tissues, like the cerebellum, have minimal replication. The DNA damage occurring within primarily non-replicative tissues is not passed onto daughter cells. As a result there is no clonal expansion of the mutated DNA sequence, which makes the detection of damage difficult with microarray technologies. Microarrays detect genotypes after PCR amplification of genomic DNA. This means, that the total number of cells containing the mutated DNA sequence has a large impact on the ability to detect these differences. When one in a million cells contains a difference in the nucleotide sequence, it will be out competed by the wild-type sequence, and come time for hybridization the mutant allele will be completely masked.

4.10.1 Variation detected in the liver consistently showed the highest number of genotype differences

The liver being a highly replicative tissue is also involved in the filtering of toxins from the body. Filtering of toxic materials exposes the liver to a high level of mutagens. High exposure to mutagens would result in a high level of DNA damage.[92] As the liver replicates, this DNA damage is fixed within the cell population and over time becomes present in clonal amounts. The clonal expansion of mutations within the liver allows for the detection of the mutation using microarray technologies. This is also consistent with previous mutation studies of exogenous genes which show that the liver has a higher accumulation of mutations in comparison to the spleen and the cerebellum.[95]

Liver-specific differences are also distributed non-randomly across the genome. This non-random distribution indicates that there is some biological basis to the observed number of differences. The liver has one of the highest mutation rates detected using single gene mutation studies. Therefore it is expected that the liver would also display the highest number of differences in each sample. The distribution of these differences across the genome reflects a tissue-specific mutation accumulation pattern. Because mutations are not distributed at random, there may be specific regions across the genome prone to mutation accumulation in the liver. The over- or under-representation of differences on specific chromosomes across the genome is indicative of liver-specific mutation hotspots. This may be due to the function of the liver and a result of specific genomic regions being accessible during certain time points across development and replication.

The replicate liver from mouse 911.50 is an outlier due to a large number of genotype differences detected within this sample. Although this tissue replicate is consistent with the liver containing more genotyping differences, many may be attributed to errors that are associated with this particular sample. The distribution of genotype differences detected for this sample, cluster at specific regions along each of the chromosomes. Although a biological reason may explain this specific clustering, without confirmation no conclusions can be made about the biological relevance of the observations. Histological assays can be conducted to determine if there are pathological differences in areas across the liver that may be associated with the large variation seen between the replicates. However, there are several technical factors associated with this sample that may explain part of the variation. The liver sample was extracted with the Wizard® Genomic DNA Purification Kit, whereas the spleen and cerebellum samples, to which it was compared, were extracted using the Gentra® Puregene® Kit. The liver also had a call rate of 97.50%, which is near the cut off threshold, and was accompanied by lower 260/280 and 260/230 ratios (1.8 and 1.93 respectively) meaning that there may have been contamination prior to DNA hybridization. The concentration immediately prior to DNA hybridization for this sample alone was also not checked meaning that the DNA hybridizes to the microarray may not have been at the proper concentrations for optimal genotyping. The probes that detected genotype differences within this sample could be checked for sequence context, such as GC content to determine if hybridization conditions or poor DNA concentration could result in the reduced accuracy of genotyping.

4.10.2 Biological reasons, other than SNP differences, that may account for genotype differences between samples

Genome-wide mutation detection also allows for the identification of large-scale genomic modifications. Insertions, deletions, and genomic rearrangements that cannot be detected using single gene mutation detection approaches, are now detected using genomic technologies. With microarray technologies, many of these modifications rely on the implementation of copy number variant calling algorithms; however, there is one type of variant that can be detected using genotype calls alone. Copy number deletions that result in the loss of two copies of a segment of DNA will result in the failure of one or multiple probe sets to determine a genotype. When the loss of the sequence does not occur in every tissue studied, this copy number deletion will be picked up as a genotype difference when tissues are compared. A genotype difference will be detected because one tissue will contain a genotype and the other tissue (containing the deletion) a NoCall. Therefore, not all the differences identified by the genotype calls are associated with point mutations. Rather, these large-scale deletions can impact the total number of differences observed between tissues.

4.11 Equivalent levels of genotype differences between the spleen and the cerebellum reflect recent literature on mutation accumulation in the *harlequin* mouse

The limited sample size for the *hq* mouse prevents any statistically significant conclusions about the accumulation of mutations in the spleen and the cerebellum. However, the observed genotype differences detected between the spleen and the cerebellum is consistent with recent reports of mutation accumulation within these tissues. Originally, the cerebellum of the *hq* mouse was thought to have an increase in mutations in comparison to other tissues as a result of an increase in ROS.[51] However, recent research contradicts these reports as mutation accumulation does not appear to be elevated in the cerebellum as a result of an increase in ROS associated mutation accumulation.[54] In this study, the total number of spleen-specific and cerebellum-specific genotype differences was not consistently elevated in any one tissue. This may be indicative of a comparable level of mutation accumulation between these two tissues.

4.12 Future directions and recommendations

Future experiments conducted using the MDGA must first address technical variability. Currently, the MDGA does not have an optimized workflow for data analysis from DNA isolation to genotype calling as the pipeline for data analysis was directly taken from the Human 6.0 array without any modifications. Therefore, the first step in accomplishing this is to optimize each step involved in the processing of DNA prior to hybridization to the MDGA for the new array design. Guidelines must be set outlining the ideal DNA extraction protocols to be used for each tissue type, as variation in tissue composition can result in varying amounts of DNA contamination. A minimum cut off value for 260/280 and 260/230 ratios of DNA immediately prior to restriction enzyme digestion should be provided to limit the amount of contamination that may affect downstream processing. The microarray also needs to be optimized so that the hybridization temperature is 5 degrees above the average melting temperature for the probes.[109] Ultimately, a whole new framework specific to the MDGA will be created that should yield the most accurate genotyping results when combined with the filtered SNP list that I have provided.

I have determined to what degree (within a tissue, between tissues, between mice, of within an inbred strain, and between closely and distantly related strains) genetic variation can reliably be detected using genotype differences determined with the MDGA. The MDGA can detect genetic variation consistently and accurately between distantly and closely related mice, between an admixture of two genetic backgrounds, and within an inbred mouse strain. However, there is low power to draw conclusions about detecting variability between and within tissues of the same mouse. Future experiments must therefore address the issue of small sample size by increasing the number of sample within each analysis and sequencing must be completed for the observed genotype differences to identify a false positive rate for the MDGA.

The genotype differences detected using the MDGA currently are a combination of true biological variation and an unknown false positive rate. Therefore, no conclusions can be made about the accumulation of genetic differences between tissues, within tissues, or about the *hq* genotype. To make any strong conclusions about specific genotype differences, the false positive rate for the MDGA must first be determined. Apart from sequencing the genotype differences detected in each of the samples, to determine false positive rate I recommend three experimental designs. Reconstruction experiments, which contain known amounts and locations of genomic variation can be analysed using the MDGA to determine the sensitivity of the array to identify the location of variation within the mouse genome. Known mutagen testing using a range of doses will determine the minimum amount of genetic variation required for the MDGA to detect significant differences between samples. Testing for known tissue-specific differences in mutation frequencies associated with age will help determine the threshold of mutation accumulation that is required for a significant differences to be detected. Further statistical analyses can also be conducted to not only determine differences in the total number and distribution of differences across different chromosomes, but also along individual chromosomes to identify regions prone to accumulate genotype differences.

4.13 Conclusion

The MDGA has the capability to be used to determine the overall genetic diversity between mouse samples. This is the first application demonstrating the range of genetic distances that the MDGA can be used to detect. The MDGA can distinguish between mice of different subspecies, between samples ranging in percentage of pure B6 to pure CBA/CaJ genetic backgrounds, and between mice within a B6 “isogenic” strain. However, when it comes to the fine scale analysis and application as a mutation detection system the MDGA requires further study with larger sample sizes. The inherent errors associated with hybridization technologies make it difficult to distinguish between technical errors and biologically relevant data. Once the above has been accomplished, the MDGA does have the potential to be used for the analysis of genome-wide mutation accumulation at a better cost than whole genome sequencing technologies.

After a framework has been optimized for the MDGA, the future directions will be to increase the number of tissue replicates in order to better understand intra-tissue variation. Increasing to the number of tissues that are analyzed will also give a better understanding as to how somatic mosaicism is accumulating throughout an individual. An exciting future application for the MDGA may be the elucidation of important mechanisms that are responsible for mutation accumulation during the developmental histories of each tissue type.

References

1. Hartwell Hood, L., Goldberg, M.L., Reynolds, A.E., Silver, L.M., and Veres, R.C., L. H. *Genetics from genes to genomes*. **Third edit**, 827 (McGraw-Hill: New York, 2008).

2. Piotrowski, A. *et al.* Somatic mosaicism for copy number variation in differentiated human tissues. *Human mutation* **29**, 1118–24 (2008).

3. Mkrtchyan, H. *et al.* The human genome puzzle - the role of copy number variation in somatic mosaicism. *Current genomics* **11**, 426–31 (2010).

4. Zhang, F. *et al.* The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nature genetics* **41**, 849–53 (2009).

5. Valverde, P., Healy, E., Jackson, I., Rees, J. L. & Thody, A. J. Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans. *Nature genetics* **11**, 328–30 (1995).

6. De, S. Somatic mosaicism in healthy human tissues. *Trends in Genetics* **27**, 217–223 (2011).

7. Smagulova, F. *et al.* Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature* **472**, 375–8 (2011).

8. Galtier, N., Enard, D., Radondy, Y., Bazin, E. & Belkhir, K. Mutation hot spots in mammalian mitochondrial DNA. *Genome Research* **16**, 215–222 (2006).

9. Russell, P. J., Wolfe, S. L., Hertz, P. E., Starr, C. & McMillan, B. *Biology the dynamic science*. (Thompson Brooks/Cole: Belmont, CA, 2008).

10. McKusick, V. A., Hostetler, J. A. & Egeland, J. A. Genetic studies of the amish, background and potentialities. *Bulletin of the Johns Hopkins Hospital* **115**, 203–22 (1964).

11. Diaz, G. A. *et al.* Gaucher disease: the origins of the Ashkenazi Jewish N370S and 84GG acid beta-glucosidase mutations. *American journal of human genetics* **66**, 1821–32 (2000).

12. Carmeli, D. B. Prevalence of Jews as subjects in genetic research: figures, explanation, and potential implications. *American journal of medical genetics. Part A* **130A**, 76–83 (2004).

13. Rosner, G., Rosner, S. & Orr-Urtreger, A. Genetic testing in Israel: an overview. *Annual review of genomics and human genetics* **10**, 175–92 (2009).

14. Moreau, C. *et al.* Genetic heterogeneity in regional populations of Quebec--parental lineages in the Gaspe Peninsula. *American journal of physical anthropology* **139**, 512–22 (2009).

15. Roy-Gagnon, M.-H. *et al.* Genomic and genealogical investigation of the French Canadian founder population structure. *Human genetics* **129**, 521–31 (2011).

16. Hou, L. *et al.* Amish revisited: next-generation sequencing studies of psychiatric disorders among the Plain people. *Trends in genetics : Trends in genetics* **29**, 412–418 (2013).

17. Cockayne, E. A. Hereditary Blue Sclerotics and Brittle Bones. *Proceedings of the Royal Society of Medicine* **7**, 101–2 (1914).

18. Landsteiner, K. & Levine, P. On individual differences in human blood. *The Journal of experimental medicine* **47**, 757–75 (1928).

19. Thomas, J. C. Blood Grouping Test in Disputed Paternity. *Journal of Criminal Law* **1**, 598 (1937).

20. Landsteiner, K. & Wiener, A. S. An Agglutinable Factor in Human Blood Recognized by Immune Sera for Rhesus Blood. *Experimental Biology and Medicine* **43**, 223 (1940).

21. Collins, F. S., Green, E. D., Guttmacher, A. E. & Guyer, M. S. A vision for the future of genomics research. **422**, 835–847 (2003).

22. Thorisson, G. A., Smith, A. V, Krishnan, L. & Stein, L. D. The International HapMap Project Web site. *Genome research* **15**, 1592–3 (2005).

23. Chee, M. *et al.* Accessing genetic information with high-density DNA arrays. *Science* **274**, 610–4 (1996).

24. Wang, D. G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–82 (1998).

25. Metzker, M. L. Emerging technologies in DNA sequencing. *Genome research* **15**, 1767–76 (2005).

26. Gräf, S. *et al.* Optimized design and assessment of whole genome tiling arrays. *Bioinformatics (Oxford, England)* **23**, i195–204 (2007).

27. Solinas-Toldo, S. *et al.* Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes chromosomes cancer* **20**, 399–407 (1997).

28. Holcomb, I. N. & Trask, B. J. Comparative genomic hybridization to detect variation in the copy number of large DNA segments. *Cold Spring Harbor protocols* **2011**, 1323–33 (2011).

29. Saccone, S. F. *et al.* Supplementing high-density SNP microarrays for additional coverage of disease-related genes: addiction as a paradigm. *PloS one* **4**, e5225 (2009).

30. Ameur, A. *et al.* Ultra-Deep Sequencing of Mouse Mitochondrial DNA: Mutational Patterns and Their Origins. *PLoS Genetics* **7**, 9 (2011).

31. Pang, A. W. *et al.* Towards a comprehensive structural variation map of an individual human genome. *Genome biology* **11**, R52 (2010).

32. Metzker, M. L. Sequencing technologies - the next generation. *Nature reviews. Genetics* **11**, 31–46 (2010).

33. Lizardi, P. M. Next-generation sequencing-by-hybridization. *Nature biotechnology* **26**, 649–50 (2008).

34. Schneider, G. F. & Dekker, C. DNA sequencing with nanopores. *Nature biotechnology* **30**, 326–8 (2012).

35. Fuller, C. W. *et al.* The challenges of sequencing by synthesis. *Nature biotechnology* **27**, 1013–23 (2009).

36. McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature genetics* **40**, 1166–74 (2008).

37. Vogler, C. *et al.* Microarray-based maps of copy-number variant regions in European and sub-Saharan populations. *PloS one* **5**, e15246 (2010).

38. Nishida, N. *et al.* Evaluating the performance of Affymetrix SNP Array 6.0 platform with 400 Japanese individuals. *BMC genomics* **9**, 431 (2008).

39. Wineinger, N. E. *et al.* Genome-wide joint SNP and CNV analysis of aortic root diameter in African Americans: the HyperGEN study. *BMC medical genomics* **4**, 4 (2011).

40. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *The American Journal of Human Genetics* **90**, 7–24 (2012).

41. Weissman, I. L. Translating Stem and Progenitor Cell Biology to the Clinic: Barriers and Opportunities. *Science* **287**, 1442–1446 (2000).

42. Ringe, J., Kaps, C., Burmester, G.-R. & Sittinger, M. Stem cells for regenerative medicine: advances in the engineering of tissues and organs. *Die Naturwissenschaften* **89**, 338–51 (2002).

43. ONeill, H. C. & Wilson, H. L. Limitations with in vitro production of dendritic cells using cytokines. *Journal of leukocyte biology* **75**, 600–3 (2004).

44. *The Laboratory Mouse*. (Elsecier Academic Press: San Diego, California, 2004).

45. Grubb, S.C., Churchill, G.A. & Bogue, M.A A collaborative database of inbred mouse strain characteristics. *Bioinformatics* **20**, 2857-9 (2004).

46. Haldane, J. B. . The Genetics of Cancer. *Nature* **132**, 265–267 (1933).

47. Bailey, D. . Recombinant-inbred strains. *Transplantation* **11**, 325–327 (1971).

48. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–62 (2002).

49. Petkov, P. M. *et al.* Development of a SNP genotyping panel for genetic monitoring of the laboratory mouse. *Genomics* **83**, 902–911 (2004).

50. Salinger, A. P. & Justice, M. J. Mouse Mutagenesis Using N-Ethyl-N-Nitrosourea (ENU). *CSH protocols* **2008**, pdb.prot4985 (2008).

51. Klein, J. A., Longo-Guess, C. M., Rossmann, M. P. & Seburn, K. L. The harlequin mouse mutation down-regulates apoptosis-inducing factor. *Nature* **419**, 367 (2002).

52. Bronson, R T, Lane, P W, Harris, B S, Davisson, M. T. Harlequin (Hq) Produces Progressive Cerebellar Cortical Atrophy. *Mouse genome* **87**, 110 (1990).

53. Miramar, M. D. *et al.* NADH oxidase activity of mitochondrial apoptosis-inducing factor. *The Journal of Biological Chemistry* **276**, 16391–16398 (2001).

54. Prtenjaca, A. & Hill, K. A Mutation frequency is not elevated in the cerebellum of harlequin/Big Blue(®) mice but Class II deletions occur preferentially in young harlequin cerebellum. *Mutation research* **707**, 53–60 (2011).

55. Susin, S. A. *et al.* Molecular characterization of mitochondrial apoptosis-inducing factor. *Nature* **397**, 441–446 (1999).

56. Laliberté, A. M., MacPherson, T. C., Micks, T., Yan, A. & Hill, K. A. Vision deficits precede structural losses in a mouse model of mitochondrial dysfunction and progressive retinal degeneration. *Experimental eye research* **93**, 833–41 (2011).

57. Srivastava, S. *et al.* Apoptosis-inducing factor regulates death in peripheral T cells. *Journal of immunology* **179**, 797–803 (2007).

58. Loeb, L. a, Bielas, J. H. & Beckman, R. a Cancers exhibit a mutator phenotype: clinical implications. *Cancer research* **68**, 3551–7; 3551-7 (2008).

59. Wright, J. H. *et al.* A random mutation capture assay to detect genomic point mutations in mouse tissue. *Nucleic acids research* **39**, e73 (2011).

60. Hill, K. A. *et al.* Spontaneous mutation in Big Blue mice from fetus to old age: tissue-specific time courses of mutation frequency but similar mutation types. *Environmental and molecular mutagenesis* **43**, 110–20 (2004).

61. Dollé, M. E. *et al.* Rapid accumulation of genome rearrangements in liver but not in brain of old mice. *Nature Genetics* **17**, 431–434 (1997).

62. Mcvicker, G. & Green, P. Genomic signatures of germline gene expression. *Genome Research* **20**, 1503–1511 (2010).

63. Quinlan, A. R. & Hall, I. M. Characterizing complex structural variation in germline and somatic genomes. *Trends in Genetics* **28**, 43–53 (2011).

64. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–94 (2011).

65. Churchill, G. a *et al.* The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nature genetics* **36**, 1133–7 (2004).

66. Chesler, E. J. *et al.* The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics. *Mammalian genome : official journal of the International Mammalian Genome Society* **19**, 382–9 (2008).

67. Aylor, D. L. *et al.* Genetic analysis of complex traits in the emerging collaborative cross. *Genome research* **21**, 1213–1222 (2011).

68. Churchill, G. A., Gatti, D. M., Munger, S. C. & Svenson, K. L. The Diversity Outbred mouse population. *Mammalian genome : official journal of the International Mammalian Genome Society* **23**, 713–8 (2012).

69. Valdar, W., Flint, J. & Mott, R. Simulating the collaborative cross: power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice. *Genetics* **172**, 1783–97 (2006).

70. Lindzey, G. & Thiessen, D. D. *Contributions to Behavior-genetic Analaysis: The Mouse as a Prototype*. 336 (Ardent Media: 1970).

71. Svenson, K. L. *et al.* High-resolution genetic mapping using the Mouse Diversity outbred population. *Genetics* **190**, 437–47 (2012).

72. Ohtsuka, M., Inoko, H., Kulski, J. K. & Yoshimura, S. Major histocompatibility complex (Mhc) class Ib gene duplications, organization and expression patterns in mouse strain C57BL/6. *BMC genomics* **9**, 178 (2008).

73. Ono, T. *et al.* Age-associated increase of spontaneous mutant frequency and molecular nature of mutation in newborn and old lacZ-transgenic mouse. *Mutation research* **447**, 165–177 (2000).

74. Dollé, M. E. T. & Vijg, J. Genome Dynamics in Aging Mice. *Genome Research* **12**, 1732–1738 (2002).

75. Dollé, M. E., Martus, H. J., Gossen, J. A., Boerrigter, M. E. & Vijg, J. Evaluation of a plasmid-based transgenic mouse model for detecting in vivo mutations. *Mutagenesis* **11**, 111–118 (1996).

76. Zhang, X. B., Urlando, C., Tao, K. S. & Heddle, J. A. Factors affecting somatic mutation frequencies in vivo. *Mutation Research* **338**, 189–201 (1995).

77. Nishino, H., Buettner, V. L., Haavik, J., Schaid, D. J. & Sommer, S. S. Spontaneous mutation in Big Blue transgenic mice: analysis of age, gender, and tissue type. *Environmental and Molecular Mutagenesis* **28**, 299–312 (1996).

78. Ono, T. *et al.* Spontaneous mutant frequency of lacZ gene in spleen of transgenic mouse increases with age. *Mutation Research* **338**, 183–188 (1995).

79. Cesta, M. F. Normal Structure , Function , and Histology of the Spleen. *Toxicologic Pathology* **34**, 455–65 (2006).

80. Cameron, I. L. Cell renewal in the organs and tissues of the nongrowing adult mouse. *Texas reports on biology and medicine* **28**, 203–248 (1970).

81. Cahan, P., Li, Y., Izumi, M. & Graubert, T. A. The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. *Nature genetics* **41**, 430–437 (2009).

82. Bross, L. *et al.* DNA double-strand breaks in immunoglobulin genes undergoing somatic hypermutation. *Immunity* **13**, 589–97 (2000).

83. Zagon, I. S., McLaughlin, P. J. & Smith, S. Neural populations in the human cerebellum: estimations from isolated cell nuclei. *Brain research* **127**, 279–82 (1977).

84. Sugino, K. *et al.* Molecular taxonomy of major neuronal classes in the adult mouse forebrain. *Nature Neuroscience* **9**, 99–107 (2006).

85. Yurov, Y. B. *et al.* Aneuploidy and confined chromosomal mosaicism in the developing human brain. *PloS one* **2**, e558 (2007).

86. Westra, J. W. *et al.* Aneuploid mosaicism in the developing and adult cerebellar cortex. *The Journal of comparative neurology* **507**, 1944–51 (2008).

87. Muotri, A. R. *et al.* Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**, 903–10 (2005).

88. Maeda, T. *et al.* Somatic DNA recombination yielding circular DNA and deletion of a genomic region in embryonic brain. *Biochemical and biophysical research communications* **319**, 1117–23 (2004).

89. Klaunig, J. E., Goldblatt, P. J., Hinton, D. E., Lipsky, M. M. & Trump, B. F. Mouse liver cell culture. II. Primary culture. *In Vitro* **17**, 926–934 (1981).

90. Baratta, J. L. *et al.* Cellular Organization of Normal Mouse Liver: A Histological, Quantitative Immunocytochemical, and Fine Structural Analysis. *October* **131**, 713–726 (2009).

91. Kelly, R., Bulfield, G., Collick, A., Gibbs, M. & Jeffreys, A. J. Characterization of a highly unstable mouse minisatellite locus: Evidence for somatic mutation during early development. *Genomics* **5**, 844–856 (1989).

92. Jennette, K. Chromate metabolism in liver microsomes. *Biological trace element research* **1**, 55-62 (1979)

93. Duncan, A. W. *et al.* Aneuploidy as a mechanism for stress-induced liver adaptation. *The Journal of clinical investigation* **122**, 3307–15 (2012).

94. Chaignat, E. *et al.* Copy number variation modifies expression time courses. *Genome research* **21**, 106–13 (2011).

95. Hill, K. A. *et al.* Tissue-specific time courses of spontaneous mutation frequency and deviations in mutation pattern are observed in middle to late adulthood in Big Blue mice. *Environmental and molecular mutagenesis* **45**, 442–54 (2005).

96. Maiti, S., Kumar, K. H. B. G., Castellani, C. A., O'Reilly, R. & Singh, S. M. Ontogenetic De Novo Copy Number Variations (CNVs) as a Source of Genetic Individuality: Studies on Two Families with MZD Twins for Schizophrenia. *PLoS ONE* **6**, 13 (2011).

97. Bielanska, M., Tan, S. L. & Ao, A. Chromosomal mosaicism throughout human preimplantation development in vitro: incidence, type, and relevance to embryo outcome. *Human reproduction Oxford England* **17**, 413–419 (2002).

98. Failla, G. The aging process and cancerogenesis. *Annals Of The New York Academy Of Sciences* **71**, 1124–1140 (1958).

99. Szilard, L. On the nature of the aging process. *Proceedings of the National Academy of Sciences of the United States of America* **45**, 30–45 (1959).

100. Curtis, H. J. Biological mechanisms underlying the aging process. *Science* **141**, 686–94 (1963).

101. Nohmi, T., Suzuki, T. & Masumura, K. Recent advances in the protocols of transgenic mouse mutation assays. *Mutation Research* **455**, 191–215 (2000).

102. Gossen, J. A. Efficient Rescue of Integrated Shuttle Vectors from Transgenic Mice: A Model for Studying Mutations in vivo. *Proceedings of the National Academy of Sciences* **86**, 7971–7975 (1989).

103. Kohler, S. W. *et al.* Spectra of spontaneous and mutagen-induced mutations in the lacI gene in transgenic mice. *ProcNatlAcadSciUSA* **88**, 7958–7962 (1991).

104. Blakey, D. H., Douglas, G. R., Huang, K. C. & Winter, H. J. Cytogenetic mapping of lambda gt10 lacZ sequences in the transgenic mouse strain 40.6 (Muta Mouse). *Mutagenesis* **10**, 145–8 (1995).

105. Swiger, R. R. Just how does the cII selection system work in Muta Mouse? *Environmental and molecular mutagenesis* **37**, 290–6 (2001).

106. Shifman, S. *et al.* A high-resolution single nucleotide polymorphism genetic map of the mouse genome. *PLoS biology* **4**, e395 (2006).

107. Yang, H. *et al.* A customized and versatile high-density genotyping array for the mouse. *Nature methods* **6**, 663–666 (2009).

108. BRLMM-P : a Genotype Calling Method for the SNP 5.0. *The Jackson Laboratory* 1–16 (2007).

109. Gresham, D. *et al.* Optimized detection of sequence variation in heterozygous genomes using DNA microarrays with isothermal-melting probes. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 1482–7 (2010).

110. Yalcin, B. *et al.* Commercially available outbred mice for genome-wide association studies. *PLoS genetics* **6**, (2010).

111. Zhang, W. *et al.* Genome-wide association mapping of quantitative traits in outbred mice. *G3 (Bethesda)* **2**, 167–74 (2012).

112. Yang, H. *et al.* Subspecific origin and haplotype diversity in the laboratory mouse. *Nature genetics* **43**, 648–655 (2011).

113. Fu, C.-P., Welsh, C. E., De Villena, F. P.-M. & McMillan, L. Inferring ancestry in admixed populations using microarray probe intensities. *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine - BCB '12* 105–112 (2012).

114. Bauchet, M. *et al.* Measuring European population stratification with microarray genotype data. *American journal of human genetics* **80**, 948–56 (2007).

115. Allen, M. J., Martinez-Martinez, J., Schroeder, D. C., Somerfield, P. J. & Wilson, W. H. Use of microarrays to assess viral diversity: from genotype to phenotype. *Environmental Microbiology* **9**, 971–982 (2007).

116. Berglund, J. *et al.* Novel origins of copy number variation in the dog genome. *Genome biology* **13**, R73 (2012).

117. Didion, J. P. *et al.* Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. *BMC genomics* **13**, 34 (2012).

118. McCue, M. E. *et al.* A high density SNP array for the domestic horse and extant Perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies. *PLoS genetics* **8**, e1002451 (2012).

119. Raman, H. *et al.* Diversity array technology markers: genetic diversity analyses and linkage map construction in rapeseed (Brassica napus L.). *DNA research : an international journal for rapid publication of reports on genes and genomes* **19**, 51–65 (2012).

120. Fadista, J. & Bendixen, C. Genomic position mapping discrepancies of commercial SNP chips. *PloS one* **7**, e31025 (2012).

121. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).

122. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Research* **19**, 1639–1645 (2009).

123. Maddatu, T. P., Grubb, S. C., Bult, C. J. & Bogue, M. A. Mouse Phenome Database (MPD). *Nucleic acids research* **40**, D887–94 (2012).

124. Hutchins, L. N. *et al.* CGDSNPdb: a database resource for error-checked and imputed mouse SNPs. *Database the journal of biological databases and curation* **2010**, 7 (2010).

125. Guide, U. Affymetrix ® Genome-Wide Human SNP Nsp / Sty. *Analysis* (2008).

126. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004).

127. Gascuel, O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution* **14**, 685–95 (1997).

128. Junier, T. & Zdobnov, E. M. The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell. *Bioinformatics* **26**, 1669–1670 (2010).

129. Tsang, S. *et al.* A comprehensive SNP-based genetic analysis of inbred mouse strains. *Mammalian genome official journal of the International Mammalian Genome Society* **16**, 476–480 (2005).

130. Lovmar, L., Ahlford, A., Jonsson, M. & Syvanen, A. C. Silhouette scores for assessment of SNP genotype clusters. *BMC genomics* **6**, 35 (2005).

131. Watkins-Chow, D. E. & Pavan, W. J. Genomic copy number and expression variation within the C57BL/6J inbred mouse strain. *Genome research* **18**, 60–6 (2008).

132. Machin, G. A. Some causes of genotypic and phenotypic discordance in monozygotic twin pairs. *American Journal of Medical Genetics* **61**, 216–228 (1996).

133. Bruder, C. E. G. *et al.* Phenotypically Concordant and Discordant Monozygotic Twins Display Different DNA Copy-Number-Variation Profiles. *Journal of Human Genetics* **82**, 763–771 (2008).

134. SNELL, G. D. & HIGGINS, G. F. Alleles at the histocompatibility-2 locus in the mouse as determined by tumor transplantation. *Genetics* **36**, 306–10 (1951).

135. Ohtsuka, M., Inoko, H., Kulski, J. K. & Yoshimura, S. Major histocompatibility complex (Mhc) class Ib gene duplications, organization and expression patterns in mouse strain C57BL/6. *BMC genomics* **9**, 178 (2008).

136. Bleiweiss, R. & Kirsch, J. A. W. Experimental analysis of variance for DNA hybridization: II. Precision. *Journal of Molecular Evolution* **37**, 514–524 (1993).

137. Hofstetter, J. R. *et al.* Genomic DNA from mice: a comparison of recovery methods and tissue sources. *Biochemical and Molecular Medicine* **62**, 197–202 (1997).

138. Bonner, J., Kung, G. & Bekhor, I. A method for the hybridization of nucleic acid molecules at low temperature. *Biochemistry* **6**, 3650–3 (1967).

139. Ling, L. L., Keohavong, P., Dias, C. & Thilly, W. G. Optimization of the polymerase chain reaction with regard to fidelity: modified T7, Taq, and vent DNA polymerases. *PCR methods and applications* **1**, 63–9 (1991).

140. Lindblad-Toh, K. *et al.* Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature genetics* **24**, 381–6 (2000).

141. Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A. & Ploner, A. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics* **21**, 3017–24 (2005).

142. Churchill, G. A. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics* **32**, 490–495 (2002).

143. Fang, Y. *et al.* A model-based analysis of microarray experimental error and normalisation. *Nucleic acids research* **31**, e96 (2003).

144. Finette, B. A., Kendall, H. & Vacek, P. M. Mutational spectral analysis at the HPRT locus in healthy children. *Mutation Research* **505**, 27–41 (2002).

145. Balin, S. J. & Cascalho, M. The rate of mutation of a single gene. *Nucleic acids research* **38**, 1575–1582 (2010).

146. Satta, Y., O’hUigin, C., Takahata, N. & Klein, J. Intensity of natural selection at the major histocompatibility complex loci. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 7184–7188 (1994).

147. Kumar, S. & Subramanian, S. Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 803–8 (2002).

148. Graubert, T. a *et al.* A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS genetics* **3**, e3 (2007).

149. Vicoso, B. & Charlesworth, B. Evolution on the X chromosome: unusual patterns and processes. *Nature reviews. Genetics* **7**, 645–53 (2006).

Appendix A - The University of Western Ontario Ethics Approval for Animal Use in Research

This appendix contains a copy of the ethics approval form for animal use in the Hill laboratory from the University Council on Animal Care and Animal Use subcommittee.

Kathleen Hill

From: eSiriusWebServer [esiriusadmin@uwo.ca]
Sent: September-01-11 11:34 AM
To:
Cc:
Subject: eSirius Notification - Annual Protocol Renewal APPROVED by the AUS 2009-033::2

AUP Number: 2009-033
PI Name: Hill, Kathleen
AUP Title: Mutational Mechanisms: Relevance and Role of Oxidative Stress

The YEARLY RENEWAL to Animal Use Protocol (AUP) 2009-033 has been approved.

1. This AUP number must be indicated when ordering animals for this project.
2. Animals for other projects may not be ordered under this AUP number.
3. Purchases of animals other than through this system must be cleared through the ACVS office. Health certificates will be required.

REQUIREMENTS/COMMENTS
Please ensure that individual(s) performing procedures on live animals, as described in this protocol, are familiar with the contents of this document.

The holder of this Animal Use Protocol is responsible to ensure that all associated safety components (biosafety, radiation safety, general laboratory safety) comply with institutional safety standards and have received all necessary approvals. Please consult directly with your institutional safety officers.

Submitted by: Thompson, Sharla H
on behalf of the Animal Use Subcommittee

The University *of* Western Ontario
Animal Use Subcommittee / University Council on Animal Care
Health Sciences Centre, ● London, Ontario ● CANADA – N6A 5C1
PH: 519-661-2111 ext. 86768 ● FL 519-661-2028
Email: auspam@uwo.ca ● http://www.uwo.ca/animal/website/

Appendix B - Annotations for the filtered SNP list

(http://www.uwo.ca/biology/Faculty/hill/)

This section contains an electronic file for the full annotations of the filtered SNP list.

Filtered SNP list annotation description

The annotated filtered SNP list contains information from my reannotations (headers highlighted in blue) and from the Didion *et al* 2012 reannotation (headers in orange). Probe sequences that are highlighted indicate the presence of a cut site for either the *Nsp*I (indicated in purple (5’-ACATGT-3’) or blue (5’-GCATGCC-3’)) or *Sty*I (indicated in green (5’-CCATGG-3’) or orange (5’-CCATGG-3’)) restriction enzymes. Note, sequences must be searched for in reverse (3’-5’) because the probe sequence in the annotation file is in the 3’-5’ orientation. If probes are cut by a single restriction enzyme, they were not removed from the analysis.

Annotations for the filtered SNP list

Column	Column Header	Description
A	PROBESET_ID	The Jackson laboratory probe ID for each SNP in the filtered SNP list
B	Chromosome	The chromosome that the SNP is located on
C	Chromosome position (bp)	The chromosome position in base pairs of the SNP
D	SNP	The nucleotide (A, G, C, or T) of allele A of the SNP
E	Alternate SNP	The alternate nucleotide (A, C, G, or T) of allele B of the SNP
F	rsNumber	The reference sequence number for the SNP
G	probe start position (+)	The start position of the probe sequence in base pairs on the positive strand
H	probe end position (+)	The end position of the probe sequence in base pairs on the positive strand
I	probe start position (-)	The start position of the probe sequence in base pairs on the negative strand
J	probe end position (-)	The end position of the probe sequence in base pairs on the negative strand
K	target sequence (5'-3') (+)	The target sequence that the probe on the positive strand will detect (5'-3')
L	target sequence (5'-3') (-)	The target sequence that the probe on the negative strand will detect (5'-3')
M	PROBE_SEQUENCE (3'-5') (+)	The probe sequence for the positive strand (3'-5')
N	probe sequence (3'-5') (-)	The probe sequence for the negative strand (3'-5')
O	RefGene	Genes identified from the UCSC known gene golden path database
P	known	Genes identified from the UCSC for the swissprot/uniprot database for the golden path
Q	in gene	Indicates whether a SNP is located within the region of the gene with a Yes or a No by combining information from "RefGene" and "known"
R	SNP source	Indicates from which strategy the SNP was identified from (Classical, NIEHS, B6, Wild, MSM, Y, or Mit)
S	C57BL/6J	Lists the nucleotide at the location of the SNP for a C57BL/6J mouse strain
T	CBA/CaJ	Lists the nucleotide at the location of the SNP for a CBA/CaJ mouse strain
U	MSM/Ms	Lists the nucleotide at the location of the SNP for a *Mus musculus molossinus*
V	Allele A strain	Indicates if a C57BL/6J (B6), CBA/CaJ (CBA) or MSM strain is detected by Allele A

Column	Column Header	Description
W	Allele B strain	Indicates if a C57BL/6J (B6), CBA/CaJ (CBA) or MSM strain is detected by Allele B
X	B6:CBA/CaJ Annotations	Indicates if the SNP is capable of detecting a B6 strain, a CBA/CaJ (CBA) strain, a B6 and a CBA/CaJ (CBA) strain, or neither a B6 nor a CBA/CaJ (CBA) strain.
Y	B6 same as CBA	Indicated if a B6 and a CBA/CaJ mouse contain the same allele
Z	Substitution type	The SNP's substitution type (transition or transversion).
AA	Orientation	The originally annotated probe orientation. There will always be two records for each probe set, one in the forward orientation (+) and one in the reverse orientation (-).
AB	Alignment score	Alignment score: -3 = Non-unique probe set: No probe in the probe set aligned uniquely, -2 = Unaligned: The probe failed to align with <= 1 mismatch, -1= Non-unique probe: This probe aligned equally well to more than one location in the genome, but other probes in the probe set aligned uniquely, 1 = Mismatch: The probe aligned uniquely with 1 off-target mismatch, 2 = Perfect: The probe aligned uniquely with no off-target mismatches
AC	Probes that match to strain	The IDs of the probes that map to this strain. There are eight probes on the array for every SNP probe set (two replicates each of four different sequences): * 2 for the reference allele in the forward orientation, * 2 for the variant allele in the forward orientation, * 2 for the reference allele in the reverse orientation, * 2 for the variant allele in the reverse orientation
AD	Probe base	Probe base: The target base of these probes.
AE	Probe sequence	Probe sequence: The genomic sequence these probes were designed to interrogate.
AF	Suboptimal alignments	Suboptimal alignments: Number of alternative alignments of lower quality than the best alignment.
AG	Chromosome	Chromosome: The chromosome to which the probe mapped. Probes from the same probe set may be mapped to different chromosomes
AH	Chromosome position (bp)	Target physical position (in bp): The physical position to which the target (polymorphic) base mapped

Column	Column Header	Description
AI	Target Genetic position (cM)	Target genetic position (in cM): The genetic position to which the target (polymorphic) base mapped. This is a sex-averaged position, and is not defined for chromosomes Y or M.
AJ	Target Base	Target base: The nucleotide at the target position in the genomic sequence.
AK	rsID	Target rsID: The rsID of the SNP at the target position.
AL	Target Deleted	Target deleted: 0 = The polymorphic base in the probe set (the position intended to interrogate the SNP in the genomic sequence) has been deleted in the genome sequence; 1 = no deletion.
AM	OTV position	OTV position: If the alignment score = 1, the position of the OTV (off-target variant).
AN	OTV type	OTV type: s = SNP, i = insertion, d = deletion
AO	OTV probe base	OTV probe base: The nucleotide at the OTV position in the probe. This will always be a single base unless the OTV type is 'i', in which case it will be empty
AP	OTV target position	OTV target position: The nucleotide at the OTV position in the genomic sequence. This will always be a single base unless the OTV type is 'd', in which case it will be empty.
AQ	OTV rsID	OTV rsID: rsID (if any) of the OTV.
AR	Orientation	Orientation: The orientation in which the probe mapped (+ or -).
AS	Start Position	Probe start position: Physical position to which the first base of the probe mapped.
AT	End Position	Probe end position: Physical position to which the last base of the probe mapped.
AU	Target Sequence	Target sequence: The genomic sequence to which the probe best aligns.
AV	Probe C/G fraction	Probe C/G fraction [0-1]: Fraction of bases in the probe that are C or G.
AW	SNP source	SNP source: See Yang 2009 for a description of the different data sources used to design the array
AX	Nsp I proximal	NspI proximal position: proximal (starting) position of the NspI fragment containing this probe.
AY	Nsp I distal	NspI distal position: distal (ending) position of the NspI fragment containing this probe.
AZ	Nsp I C/G fraction	NspI C/G fraction [0-1]: Fraction of bases in the NspI fragment that are C or G.

Column	Column Header	Description
BA	Sty I proximal	StyI proximal position: proximal (starting) position of the StyI fragment containing this probe.
BB	Sty I distal	StyI distal position: distal (ending) position of the StyI fragment containing this probe.
BC	Sty I C/G fraction	StyI C/G fraction [0-1]: Fraction of bases in the StyI fragment that are C or G.
BD	Ensembl Gene IDs	ENSEMBL gene IDs: ENSEMBL ID(s) (comma-delimited) of the gene within which this probe aligns (if any).
BE	Function	Function: Functional annotation of the gene within which this probe aligns (if any).
BF	Target coverage depth	Target coverage depth: 0 = the coverage depth was below the threshold, 1 = the coverage depth was above the threshold.
BG	Target sequence low coverage	Target sequence low coverage: Number of bases in the probe for which the genomic sequence had below-threshold coverage depth.

Appendix C - Supplementary Figures

This section contains supplementary figures used to aid in the understanding of genetic distance measures.

Table C-1: CEL file sample identifiers for each of the Hill Laboratory samples.

Mouse ID[a]	CEL file identifier[b]
904.9 Sp	SNP_mDIV_A4-SNP10_188_091610
904.11 Sp	SNP_mDIV_A5-SNP10_189_091610
911.49 Sp	SNP_mDIV_A6-SNP10_190_091610
911.50 Sp	SNP_mDIV_A7-SNP10_191_091610
904.9 Cl	SNP_mDIV_A8-SNP10_192_091610
904.11 Cl	SNP_mDIV_A9-SNP10_193_091610
911.49 Cl	SNP_mDIV_A10-SNP10_194_091610
911.50 Cl	SNP_mDIV_A11-SNP10_195_091610
911.143 Cl	SNP_mDIV_B1-SNP10_196_091610
911.148 Cl	SNP_mDIV_B2-SNP10_197_091610
300.6 Cl	DNA3158
300.6 Sp[c]	DNA3159
900.3 Li	DNA3160
900.3 Sp	DNA3161
911.50 Cl-2	DNA3162
911.50 Sp-2	DNA3163
904.9 Li[c]	DNA3253
904.11 Li[c]	DNA3254
904.12 Li[c]	DNA3255
911.49 Li[c]	DNA3256
911.47 Li[c]	DNA3257
911.50 Li	DNA3258
300.7 Cl-1	DNA3296
300.7 Cl-2	DNA3293
300.7 Cl-3	DNA3294
300.7 Sp-1	DNA3295
300.7 Sp-2	DNA3297
300.7 Sp-3	DNA3298
911.17 Sp	DNA3299
911.17 Cl	DNA3301
911.17 Li	DNA3302
911.49 Li	DNA3300
911.50 Li-2	DNA3303

[a] the identifier used for each of the samples including the mouse number, tissue type, and replicate number if there are replicates

[b] the specific identifier given to each specific CEL file
that correspond with each of the samples

[c] failed genotyping samples that were not included in analyses

Table C-2: Genotyping results for passing samples[a] after the 2nd round of genotyping using the filtered SNP list

MDGA samples identifier	Mouse Gender	Overall call rate[b]		Sample Type[c]	Mouse Strain
		336 samples	362 samples		
SNP_mDIV_C2-283_112108	male	99.88	99.88	BXD	BXD34
SNP_mDIV_A4-SNP08_002_103008	male	99.87	99.87	classical laboratory strain	C57BL/6J
SNP_mDIV_A9-382_012709	male	99.87	99.87	classical laboratory strain	C57BL/6NCr
SNP_mDIV_B1-385_012709	male	99.87	99.87	classical laboratory strain	C57BL/6NCr
SNP_mDIV_A9-SNP08_003_103008	female	99.84	99.85	classical laboratory strain	C57BL/6NJ
SNP_mDIV_B1-267_112108	male	99.84	99.84	BXD	BXD12
SNP_mDIV_A2-304_120908	male	99.83	99.83	BXD	BXD74
SNP_mDIV_B7-391_012709	male	99.83	99.83	classical laboratory strain	A/WySnJ
SNP_mDIV_A11-384_012709	male	99.82	99.82	classical laboratory strain	C57BL/6NTac
SNP_mDIV_C6-287_112108	male	99.82	99.82	BXD	BXD43
SNP_mDIV_B9-279_112108_2	male	99.82	99.81	BXD	BXD29
SNP_mDIV_C4-285_112108	male	99.81	99.81	BXD	BXD40
SNP_mDIV_C6-401_012709	male	99.80	99.81	classical laboratory strain	LT/SvEiJ
SNP_mDIV_C9-290_112108	male	99.82	99.81	BXD	BXD49
SNP_mDIV_A7-SNP10_191_091610	male		99.80		C57BL/6J/CBA/CaJ
SNP_mDIV_A3-421_022709	male	99.78	99.79	BXD	BXD100
SNP_mDIV_C8-SNP09_014_022709	male	99.79	99.79	consomic	(NOD.NON-Thy1 (N13F21)
SNP_mDIV_D3-409_012709	male	99.79	99.79	classical laboratory strain	SEC/1ReJ
SNP_mDIV_D6-298_112108_2	male	99.77	99.78	BXD	BXD67
SNP_mDIV_D8-414_012709	male	99.78	99.78	BXD	BXD9
SNP_mDIV_A11-265_112108	male	99.77	99.77	BXD	BXD8
SNP_mDIV_B1-SNP08_004_103008_4	male	99.76	99.77	classical laboratory strain	C57BL/6NJ
SNP_mDIV_B5-123_091708	male	99.75	99.76	classical laboratory strain	BPH/2J
SNP_mDIV_D2-408_012709	male	99.76	99.76	classical laboratory strain	SEC/1GnLeJ
SNP_mDIV_D9-261_111308	male	99.77	99.76	BXD	BXD1
SNP_mDIV_D11-263_111308	male	99.77	99.76	BXD	BXD5
SNP_mDIV_B1-SNP10_196_091610	male		99.76		C57BL/6J/CBA/CaJ

MDGA samples identifier	Mouse Gender	Overall call rate[b]		Sample Type[c]	Mouse Strain
		336 samples	362 samples		
SNP_mDIV_A6-424_022709	male	99.75	99.75	classical laboratory strain	IHOT1
SNP_mDIV_A8-381_012709	male	99.75	99.75	classical laboratory strain	C57BL/6NCrl
SNP_mDIV_B8-132_091708	male	99.75	99.75	classical laboratory strain	HPG/BmJ
SNP_mDIV_C11-406_012709	male	99.75	99.75	classical laboratory strain	NONcNZO5/LtJ
SNP_mDIV_A7-153_111308	male	99.74	99.74	classical laboratory strain	TKDU/DnJ
SNP_mDIV_C9-404_012709	male	99.74	99.74	classical laboratory strain	NOD/ShiLtJ
SNP_mDIV_B2-SNP10_197_091610	male		99.74		C57BL/6J/CBA/CaJ
SNP_mDIV_A1-SNP08_001_103008	female	99.73	99.73	classical laboratory strain	C57BL/6J
SNP_mDIV_A3-3_081308	male	99.72	99.73	classical laboratory strain	AKR/J
SNP_mDIV_A4-4_081308	male	99.73	99.73	classical laboratory strain	BALB/cByJ
SNP_mDIV_B6-274_112108	male	99.73	99.73	BXD	BXD20
SNP_mDIV_B10-21_081308	male	99.73	99.73	classical laboratory strain	FVB/NJ
SNP_mDIV_B11-281_112108	male	99.74	99.73	BXD	BXD32
SNP_mDIV_A5-SNP10_189_091610	male		99.73		C57BL/6J/CBA/CaJ
SNP_mDIV_A9-71_090908	male	99.72	99.72	consomic	C57BL/6J-Chr12/ForeJ
SNP_mDIV_A11-313_120908	male	99.71	99.72	BXD	BXD83
SNP_mDIV_A11-361_121608	male	99.72	99.72	BXD	BXD61
SNP_mDIV_B10-280_112108	male	99.73	99.72	BXD	BXD31
SNP_mDIV_A4-SNP10_188_091610	male		99.72		C57BL/6J/CBA/CaJ
SNP_mDIV_A6-SNP10_190_091610	male		99.72		C57BL/6J/CBA/CaJ
SNP_mDIV_A8-SNP10_192_091610	male		99.72		C57BL/6J/CBA/CaJ
SNP_mDIV_C3-398_012709	male	99.70	99.71	classical laboratory strain	DBA/1LacJ
SNP_mDIV_A10-SNP10_194_091610	male		99.71		C57BL/6J/CBA/CaJ
SNP_mDIV_A6-356_121608	male	99.70	99.70	BXD	BXD45
SNP_mDIV_C7-288_112108	male	99.70	99.70	BXD	BXD44
SNP_mDIV_A2-420_022709	male	99.69	99.69	BXD	BXD99
SNP_mDIV_A9-SNP10_193_091610	male		99.69		C57BL/6J/CBA/CaJ
SNP_mDIV_A11-SNP08_004_103008	male	99.68	99.68	classical laboratory strain	C57BL/6NJ

MDGA samples identifier	Mouse Gender	Overall call rate[b]		Sample Type[c]	Mouse Strain
		336 samples	362 samples		
SNP_mDIV_B10-394_012709	male	99.69	99.68	classical laboratory strain	BXSB/MpJ
SNP_mDIV_D4-410_012709	male	99.68	99.68	classical laboratory strain	SJL/Bm
DNA3294	male		99.68		C57BL/6J
SNP_mDIV_A3-SNP08_001_103008	female	99.68	99.67	classical laboratory strain	C57BL/6J
SNP_mDIV_B1-314_120908	male	99.68	99.67	BXD	BXD84
SNP_mDIV_D7-299_112108	male	99.67	99.67	BXD	BXD68
SNP_mDIV_D11-303_112108	male	99.66	99.66	BXD	BXD73
SNP_mDIV_A1-50_091708	male	99.64	99.65	classical laboratory strain	NZB/BINJ
SNP_mDIV_B1-432_022709	male	99.65	99.65	classical laboratory strain	ISS
SNP_mDIV_B9-373_121608	male	99.65	99.65	BXD	BXD103MK88
SNP_mDIV_C5-286_112108	male	99.66	99.65	BXD	BXD42
SNP_mDIV_D5-411_012709	male	99.65	99.65	classical laboratory strain	ST/bJ
SNP_mDIV_A9-429_022709	male	99.64	99.64	classical laboratory strain	IBWSR2
SNP_mDIV_B8-19_081308	male	99.64	99.64	classical laboratory strain	DBA/2J
SNP_mDIV_B9-138_091708	male	99.65	99.64	classical laboratory strain	P/J
SNP_mDIV_D2-SNP09_024_022709	male	99.64	99.64	classical laboratory strain	C57BLKS/J
SNP_mDIV_A8-427_022709	male	99.62	99.63	classical laboratory strain	ICOLD2
SNP_mDIV_A10-SNP08_004_103008	male	99.63	99.63	classical laboratory strain	C57BL/6NJ
SNP_mDIV_B11-88_090908	male	99.63	99.63	classical laboratory strain	C57L/J
SNP_mDIV_C3-284_112108	male	99.63	99.63	BXD	BXD39
SNP_mDIV_A8-154_111308	male	99.63	99.62	classical laboratory strain	TSJ/LeJ
SNP_mDIV_B1-362_121608	male	99.62	99.62	BXD	BXD86
SNP_mDIV_C7-333_120908	male	99.60	99.62	F1 hybrid	(C57BL/6JxBALB/cJ)F1
SNP_mDIV_A10-72_090908	male	99.61	99.61	consomic	C57BL/6J-Chr14/ForeJ
SNP_mDIV_A10-264_112108	male	99.62	99.61	BXD	BXD6
SNP_mDIV_D5-253_111308	male	99.61	99.61	classical laboratory strain	BALB/cJ
SNP_mDIV_B9-393_012709	male	99.60	99.60	classical laboratory strain	BDP/J
SNP_mDIV_C11-337_120908	male	99.58	99.60	F1 hybrid	(DBA/2JxC57BL/6J)F1
SNP_mDIV_D7-413_012709	male	99.60	99.60	classical laboratory strain	YBR/EiJ

MDGA samples identifier	Mouse Gender	Overall call rate[b]		Sample Type[c]	Mouse Strain
		336 samples	362 samples		
SNP_mDIV_A1-1_081308	male	99.60	99.59	classical laboratory strain	129S1SvlmJ
SNP_mDIV_A2-48_082108	male	99.59	99.59	classical laboratory strain	NON/ShiLtJ
SNP_mDIV_A3-49_082108	male	99.60	99.59	classical laboratory strain	NOR/LtJ
SNP_mDIV_B4-15_081308	male	99.58	99.59	classical laboratory strain	CBA/CaJ
SNP_mDIV_B8-392_012709	male	99.59	99.59	classical laboratory strain	AEJ/GnRk
SNP_mDIV_C4-93_090908	male	99.59	99.59	classical laboratory strain	LP/J
SNP_mDIV_A5-5_081308	male	99.59	99.58	classical laboratory strain	BTBRT+tf/J
SNP_mDIV_A7-SNP08_003_103008	female	99.59	99.58	classical laboratory strain	C57BL/6NJ
SNP_mDIV_D1-126_090908	male	99.58	99.58	classical laboratory strain	C3HeB/FeJ
SNP_mDIV_D8-256_111308	male	99.58	99.58	classical laboratory strain	CBA/J
SNP_mDIV_A5-423_022709	male	99.57	99.57	BXD	BXD102
SNP_mDIV_B3-387_022709	male	99.57	99.57	classical laboratory strain	129P1/ReJ
SNP_mDIV_C3-92_090908	male	99.58	99.57	classical laboratory strain	LG/J
SNP_mDIV_A1-282_120908	male	99.57	99.56	BXD	BXD33
SNP_mDIV_B1-68_091708	male	99.57	99.56	consomic	C57BL/6J-Chr10.3/ForeJ
SNP_mDIV_B5-271_112108	male	99.56	99.56	BXD	BXD16
SNP_mDIV_A11-SNP10_195_091610	male		99.56		C57BL/6J/CBA/CaJ
SNP_mDIV_B8-370_121608	male	99.54	99.55	BXD	BXD96
SNP_mDIV_A3-305_120908	male	99.54	99.54	BXD	BXD75
SNP_mDIV_B3-380_121608	male	99.51	99.54	F1 hybrid	(C57B6/JxDBA/2J)F1
SNP_mDIV_C8-32_081308	male	99.54	99.54	classical laboratory strain	NZW/LacJ
DNA3297	male		99.52		C57BL/6J
SNP_mDIV_B6-82_090908	male	99.52	99.51	consomic	C57BL/6J-ChrX.2/ForeJ
SNP_mDIV_C10-336_120908	male	99.50	99.51	F1 hybrid	(BALB/cJxC57BL/6J)F1
SNP_mDIV_B4-118_091708	male	99.51	99.50	classical laboratory strain	AEJ/GnLeJ
SNP_mDIV_D6-412_012709	male	99.52	99.50	classical laboratory strain	STX/Le
SNP_mDIV_D11-139_090908	male	99.50	99.49	classical laboratory strain	PN/nBSwUmabJ
DNA3298	male		99.49		C57BL/6J
SNP_mDIV_A4-150_111308_2	male	99.48	99.48	classical laboratory strain	SSL/LeJ

MDGA samples identifier	Mouse Gender	Overall call rate[b]		Sample Type[c]	Mouse Strain
		336 samples	362 samples		
SNP_mDIV_A7-7_081308	male	99.49	99.48	classical laboratory strain	C57BL/6J
SNP_mDIV_B5-389_012709	male	99.49	99.48	classical laboratory strain	129T2/SvEmsJ
SNP_mDIV_C6-30_081308	male	99.49	99.48	classical laboratory strain	NOD/ShiLtJ
SNP_mDIV_D2-339_120908	male	99.46	99.48	F1 hybrid	(NZW/LacJxC57BL/6J)F1
SNP_mDIV_D1-SNP09_023_022709	male	99.45	99.47	congenic	(NODxC57BLKS)F1
SNP_mDIV_A2-2_081308	male	99.48	99.46	classical laboratory strain	A/J
SNP_mDIV_A6-SNP08_002_103008	male	99.48	99.46	classical laboratory strain	C57BL/6J
SNP_mDIV_A7-425_022709	male	99.44	99.45	classical laboratory strain	IHOT2
SNP_mDIV_B2-75_103008_4	male	99.46	99.45	consomic	C57BL/6J-Chr18/ForeJ
SNP_mDIV_B11-22_081308	male	99.44	99.43	classical laboratory strain	KK/HIJ
SNP_mDIV_C2-91_090908	male	99.44	99.43	classical laboratory strain	JE/LeJ
SNP_mDIV_D1-407_012709	male	99.43	99.43	wild-derived laboratory strain	RBB/DnJ
SNP_mDIV_D6-254_111308	male	99.44	99.43	classical laboratory strain	129X1/SvJ
DNA3163	male		99.43		C57BL/6J/CBA/CaJ
SNP_mDIV_A5-SNP08_002_103008	male	99.43	99.42	classical laboratory strain	C57BL/6J
SNP_mDIV_A9-155_111308	male	99.43	99.42	classical laboratory strain	ZRDCT Rax+/ChUmdJ
SNP_mDIV_D8-345_120908	male	99.41	99.42	F1 hybrid	(C57BL/6JxAKR/J)F1
SNP_mDIV_A6-152_111308	male	99.43	99.41	classical laboratory strain	TALLYHO/JngJ
SNP_mDIV_A2-352_121608	male	99.40	99.40	BXD	BXD18
SNP_mDIV_A5-151_111308	male	99.42	99.40	classical laboratory strain	SWR/J
SNP_mDIV_B7-275_112108	male	99.42	99.40	BXD	BXD21
SNP_mDIV_C8-97_090908	male	99.42	99.40	classical laboratory strain	RIIIS/J
SNP_mDIV_B2-433_022709	male	99.39	99.39	classical laboratory strain	ILS
SNP_mDIV_B4-81_103008_4	male	99.39	99.39	consomic	C57BL/6J-Chr9/ForeJ
SNP_mDIV_B2-268_112108	male	99.39	99.38	BXD	BXD13
SNP_mDIV_A7-357_121608	male	99.38	99.37	BXD	BXD48
SNP_mDIV_A8-SNP08_003_103008	female	99.38	99.37	classical laboratory strain	C57BL/6NJ
SNP_mDIV_B9-86_090908	male	99.37	99.36	classical laboratory strain	C57BLKS/J

MDGA samples identifier	Mouse	Overall call rate[b]		Sample Type[c]	Mouse Strain
	Gender	336 samples	362 samples		
DNA3295	male		99.36		C57BL/6J
DNA3301	male		99.36		C57BL/6J/CBA/CaJ
SNP_mDIV_A6-6_081308	male	99.35	99.35	classical laboratory strain	C3H/HeJ
SNP_mDIV_D4-130_090908	male	99.35	99.35	classical laboratory strain	DLS/LeJ
SNP_mDIV_A2-148_111308	male	99.34	99.34	classical laboratory strain	SM/J
SNP_mDIV_C5-94_090908	male	99.36	99.34	classical laboratory strain	NZL/LtJ
SNP_mDIV_B10-143_103008_3	male	99.31	99.32	classical laboratory strain	RSV/LeJ
SNP_mDIV_D9-415_012709	male	99.34	99.32	BXD	BXD24
SNP_mDIV_B5-84_103008_4	male	99.32	99.30	consomic	C57BL/6J-ChrY/ForeJ
DNA3293	male		99.30		C57BL/6J
SNP_mDIV_D1-36_081308	male	99.29	99.27	classical laboratory strain	SJL/J
SNP_mDIV_A2-SNP08_001_103008	female	99.27	99.26	classical laboratory strain	C57BL/6J
SNP_mDIV_C7-31_081308	male	99.23	99.22	classical laboratory strain	NZO/HILtJ
SNP_mDIV_B10-140_091708	male	99.20	99.19	wild-derived laboratory strain	RBF/DnJ
SNP_mDIV_C1-89_090908	male	99.18	99.18	classical laboratory strain	C58/J
SNP_mDIV_A1-147_111308	male	99.19	99.17	classical laboratory strain	PN/nBSwUmabJ
SNP_mDIV_B9-20_081308	male	99.18	99.17	classical laboratory strain	DDY/JclSidSeyFrkJ
SNP_mDIV_D10-137_090908	male	99.16	99.14	classical laboratory strain	NZM2410/J
DNA3296	male		99.13		C57BL/6J
SNP_mDIV_A11-431_022709	female	99.13	99.11	classical laboratory strain	IBWSP2
SNP_mDIV_B7-18_081308	male	99.12	99.11	classical laboratory strain	DBA/1J
SNP_mDIV_D9-144_103008_3	male	99.14	99.11	classical laboratory strain	SB/LeJ
SNP_mDIV_A8-199_091708	male	99.13	99.10	classical laboratory strain	129S6
SNP_mDIV_D9-136_090908	male	99.12	99.10	classical laboratory strain	NU/J
DNA3162	male		99.10		C57BL/6J/CBA/CaJ
DNA3303	male		99.10		C57BL/6J/CBA/CaJ
SNP_mDIV_B11-141_091708	male	99.09	99.07	classical laboratory strain	RF/J
SNP_mDIV_A11-202_091708	male	99.06	99.06	CC-UNC G2:F1	OR1244m19
SNP_mDIV_A8-56_082108	female	99.06	99.03	classical laboratory strain	DDK/Pas

MDGA samples identifier	Mouse Gender	Overall call rate[b]		Sample Type[c]	Mouse Strain
		336 samples	362 samples		
SNP_mDIV_C9-120_090908	male	99.05	99.03	classical laboratory strain	ALS/LtJ
DNA3299	male		99.03		C57BL/6J/CBA/CaJ
SNP_mDIV_D8-300_112108	male	99.02	99.00	BXD	BXD69
SNP_mDIV_A4-481_012709	male	99.00	98.99	wild-derived laboratory strain	STRA
SNP_mDIV_C11-125_090908	male	99.02	98.99	classical laboratory strain	BUB/BnJ
SNP_mDIV_D10-262_111308	male	99.00	98.98	BXD	BXD2
SNP_mDIV_A5-482_012709	male	98.98	98.97	wild-derived laboratory strain	STRB
SNP_mDIV_D5-131_090908	male	99.00	98.97	classical laboratory strain	EL/SuzSeyFrkJ
SNP_mDIV_D11-417_012709	male	98.99	98.97	BXD	BXD65
SNP_mDIV_C9-335_120908	male	98.96	98.96	F1 hybrid	(C57BL/6JxNZW/LacJ)F1
SNP_mDIV_A10-201_091708	male	98.96	98.95	CC-UNC G2:F1	OR615m104
SNP_mDIV_A6-119_090908	male	98.95	98.94	classical laboratory strain	ALR/LtJ
SNP_mDIV_C6-95_090908	male	98.97	98.94	classical laboratory strain	PL/J
SNP_mDIV_C7-402_012709	male	98.94	98.93	wild-derived laboratory strain	MOR/RkJ
SNP_mDIV_D2-128_090908	male	98.95	98.92	classical laboratory strain	CE/J
SNP_mDIV_D3-129_090908	male	98.94	98.92	classical laboratory strain	CHMU/LeJ
SNP_mDIV_D8-185_082108	female	98.91	98.90	CC-UNC G2:F1	OR804f103
SNP_mDIV_C5-331_120908	male	98.87	98.89	F1 hybrid	(AKR/JxC57BL/6J)F1
SNP_mDIV_D10-145_103008_3	male	98.92	98.89	classical laboratory strain	SEA/GnJ
SNP_mDIV_A9-200_091708	male	98.88	98.88	CC-UNC G2:F1	OR496m20
SNP_mDIV_B8-85_090908	male	98.92	98.88	classical laboratory strain	C57BL/10J
SNP_mDIV_D2-178_082108	male	98.87	98.87	CC-UNC G2:F1	OR95m20
SNP_mDIV_D11-348_120908	male	98.87	98.87	F1 hybrid	(NZW/LacJxWSB/EiJ)F1
SNP_mDIV_D3-179_082108	female	98.85	98.85	CC-UNC G2:F1	OR95f16
SNP_mDIV_A6-483_012709	male	98.85	98.84	wild-derived laboratory strain	STLT
SNP_mDIV_A7-69_090908	male	98.84	98.82	consomic	C57BL/6J-Chr11.1/ForeJ
DNA3302	male		98.82		C57BL/6J/CBA/CaJ
SNP_mDIV_A1-350_121608	male	98.81	98.80	F1 hybrid	(WSB/EiJxNZW/LacJ)F1

MDGA samples identifier	Mouse Gender	Overall call rate[b]		Sample Type[c]	Mouse Strain
		336 samples	362 samples		
SNP_mDIV_B8-278_112108	male	98.83	98.80	BXD	BXD28
SNP_mDIV_C3-25_081308	female	98.81	98.80	CC-UNC G2:F1	OR294f18
SNP_mDIV_C4-26_081308	female	98.80	98.80	CC-UNC G2:F1	OR1109f19
SNP_mDIV_A2-64_090908	female	98.80	98.79	CC-UNC G2:F1	OR447f19
SNP_mDIV_B1-60_082108	male	98.80	98.79	CC-UNC G2:F1	OR1587m104
SNP_mDIV_A10-383_012709	male	98.78	98.76	classical laboratory strain	C57BL/6NTac
SNP_mDIV_D1-177_082108	female	98.76	98.74	CC-UNC G2:F1	OR88f16
SNP_mDIV_D9-44_081308	male	98.76	98.74	CC-UNC G2:F1	OR1400m105
SNP_mDIV_C7-462_012209	male	98.74	98.72	wild-derived laboratory strain	DJO
SNP_mDIV_C8-463_012209	male	98.71	98.70	wild-derived laboratory strain	DOT
SNP_mDIV_A2-51_091708	male	98.71	98.69	wild-derived laboratory strain	PERA/EiJ
SNP_mDIV_C1-23_081308	male	98.71	98.69	wild-derived laboratory strain	LEWES/EiJ
SNP_mDIV_C7-96_090908	male	98.71	98.69	wild-derived laboratory strain	RBA/DnJ
SNP_mDIV_C11-466_012209	male	98.68	98.67	wild-derived laboratory strain	DMZ
SNP_mDIV_A3-156_091708	male	98.69	98.66	wild caught	KCT222
SNP_mDIV_A10-58_082108	male	98.67	98.66	CC-UNC G2:F1	OR151m105
SNP_mDIV_A11-59_082108	female	98.67	98.66	CC-UNC G2:F1	OR1587f101
SNP_mDIV_D4-180_082108	male	98.67	98.66	CC-UNC G2:F1	OR656m18
SNP_mDIV_B2-90_091708	male	98.67	98.65	classical laboratory strain	I/LnJ
SNP_mDIV_B9-162_082108	male	98.67	98.65	wild caught	MWN1026
SNP_mDIV_C11-176_082108	male	98.66	98.65	CC-UNC G2:F1	OR88m19
SNP_mDIV_D11-146_103008_3	male	98.68	98.65	classical laboratory strain	SH1/LeJ
SNP_mDIV_B6-124_091708	male	98.67	98.64	classical laboratory strain	BPN/3J
SNP_mDIV_A11-11_081308	male	98.65	98.63	CC-UNC G2:F1	OR1005m105
SNP_mDIV_B3-98_091708	female	98.65	98.63	wild caught	RDS12763
SNP_mDIV_B4-388_012709	male	98.66	98.63	classical laboratory strain	129P3/J
SNP_mDIV_A6-183_091708	female	98.63	98.62	CC-UNC G2:F1	OR672f102
SNP_mDIV_B6-159_082108	female	98.60	98.61	wild caught	MWN1214
SNP_mDIV_D1-467_012209	male	98.63	98.61	wild-derived laboratory strain	BZO

MDGA samples identifier	Mouse Gender	Overall call rate[b]		Sample Type[c]	Mouse Strain
		336 samples	362 samples		
SNP_mDIV_A1-62_090908	male	98.61	98.60	CC-UNC G2:F1	OR294m21
SNP_mDIV_D5-181_082108	female	98.60	98.59	CC-UNC G2:F1	OR656f14
SNP_mDIV_C2-166_082108	male	98.57	98.58	wild caught	MWN1198
DNA3160	female		98.58		C57BL/6J x CBA/CaJ
DNA3158	male		98.57		C57BL/6J
SNP_mDIV_C5-460_012209	male	98.58	98.56	wild-derived laboratory strain	WMP
SNP_mDIV_D7-184_082108	male	98.57	98.55	CC-UNC G2:F1	OR804m105
SNP_mDIV_D2-294_112108	male	98.56	98.54	BXD	BXD60
SNP_mDIV_B9-142_103008_3	male	98.58	98.53	classical laboratory strain	RHJ/LeJ
SNP_mDIV_B3-14_081308	male	98.53	98.51	CC-UNC G2:F1	OR1048m20
SNP_mDIV_B7-160_082108	male	98.52	98.51	wild caught	MWN1194
SNP_mDIV_D3-38_081308	male	98.54	98.51	wild-derived laboratory strain	TIRANO/EiJ
SNP_mDIV_D9-186_082108	female	98.53	98.51	CC-UNC G2:F1	OR873f102
SNP_mDIV_A1-434_012209	male	98.51	98.50	wild-derived laboratory strain	BIK/g1
SNP_mDIV_B1-12_081308	female	98.52	98.50	CC-UNC G2:F1	OR1018f102
SNP_mDIV_C1-456_012209	male	98.52	98.50	wild-derived laboratory strain	WLA
SNP_mDIV_B2-76_090908	male	98.53	98.49	consomic	C57BL/6J-Chr19/ForeJ
SNP_mDIV_A11-444_012209	male	98.49	98.48	wild-derived laboratory strain	DCP
SNP_mDIV_A3-149_111308	male	98.49	98.47	wild-derived laboratory strain	SOD1/EiJ
SNP_mDIV_B11-164_082108	female	98.47	98.46	wild caught	MWN1030
SNP_mDIV_D6-343_120908	male	98.46	98.46	F1 hybrid	(PWD/PhJxNOD/ShiJ)F1
SNP_mDIV_A9-442_012209	male	98.47	98.45	wild-derived laboratory strain	DCA
SNP_mDIV_D5-219_103008_3	male	98.47	98.45	CC-UNC G2:F1	OR873m106
SNP_mDIV_D3-340_120908	male	98.43	98.43	F1 hybrid	(NZW/LacJxPWD/PhJ)F1
SNP_mDIV_D5-40_081308	female	98.43	98.42	CC-UNC G2:F1	OR1262f101
SNP_mDIV_A10-10_081308	female	98.41	98.40	CC-UNC G2:F1	OR1005f102
SNP_mDIV_A9-57_082108	female	98.41	98.38	CC-UNC G2:F1	OR151f102
DNA3160	female		98.37		C57BL/6J x CBA/CaJ
SNP_mDIV_C5-27_081308	male	98.37	98.36	CC-UNC G2:F1	OR1109m20

MDGA samples identifier	Mouse Gender	Overall call rate[b]		Sample Type[c]	Mouse Strain
		336 samples	362 samples		
SNP_mDIV_D6-182_082108	male	98.37	98.36	CC-UNC G2:F1	OR672m106
SNP_mDIV_D6-220_103008_3	male	98.37	98.36	CC-UNC G2:F1	OR978m21
SNP_mDIV_D7-221_103008_3	female	98.37	98.36	CC-UNC G2:F1	OR906f102
SNP_mDIV_C1-165_082108	female	98.34	98.35	wild caught	MWN1287
SNP_mDIV_C4-241_111308	male	98.36	98.35	F1 hybrid	(PWK/PhJxA/J)F1
SNP_mDIV_C5-207_103008_3	male	98.35	98.35	F1 hybrid	(129S1/SvImJIxPWK/PhJ)F1
SNP_mDIV_B1-227_111308	male	98.36	98.34	F1 hybrid	(C57BL/6JxPWK/PhJ)F1
SNP_mDIV_C3-205_103008_3	female	98.35	98.34	F1 hybrid	(129S1/SvImJIxPWK/PhJ)F1
SNP_mDIV_C5-400_012709	male	98.36	98.34	classical laboratory strain	DBA/2HaSmnJ
SNP_mDIV_C9-211_103008	male	98.35	98.34	F1 hybrid	(PWK/PhJx129S1/SvImJ)F1
SNP_mDIV_A5-378_121608	male	98.35	98.32	classical laboratory strain	C57BL/6J
SNP_mDIV_C4-206_103008_3	female	98.33	98.32	F1 hybrid	(129S1/SvImJIxPWK/PhJ)F1
SNP_mDIV_A4-52_082108	male	98.33	98.31	wild-derived laboratory strain	PERC/EiJ
SNP_mDIV_A4-437_012209	male	98.32	98.31	wild-derived laboratory strain	DEB
SNP_mDIV_D4-218_103008_3	female	98.33	98.31	CC-UNC G2:F1	OR496f18
SNP_mDIV_D7-344_120908	male	98.31	98.30	F1 hybrid	(PWD/PhJxNZW/LacJ)F1
SNP_mDIV_C10-212_103008_3	male	98.30	98.29	F1 hybrid	(PWK/PhJx129S1/SvImJ)F1
SNP_mDIV_C10-247_111308	male	98.30	98.29	F1 hybrid	(PWK/PhJxNOD/ShiJ)F1
SNP_mDIV_B2-228_111308	male	98.30	98.28	F1 hybrid	(CAST/EiJx129S1/SvImJ)F1
SNP_mDIV_D6-41_081308	female	98.29	98.28	CC-UNC G2:F1	OR1305f101
SNP_mDIV_C2-239_111308	male	98.25	98.24	F1 hybrid	(NOD/ShiJxPWK/PhJ)F1
SNP_mDIV_D3-217_103008_3	male	98.24	98.23	CC-UNC G2:F1	OR1611m121
SNP_mDIV_A11-224_111308	male	98.23	98.21	F1 hybrid	(A/JxPWK/PhJ)F1
SNP_mDIV_D10-347_120908	male	98.22	98.21	F1 hybrid	(CAST/EiJxNZW/LacJ)F1
SNP_mDIV_C8-210_103008_3	female	98.20	98.19	F1 hybrid	(PWK/PhJx129S1/SvImJ)F1
SNP_mDIV_B7-451_012209	male	98.20	98.18	wild-derived laboratory strain	DIK
SNP_mDIV_D2-216_103008_3	male	98.19	98.18	CC-UNC G2:F1	OR1305m105
SNP_mDIV_D4-252_111308_2	male	98.22	98.18	wild-derived laboratory strain	ZALENDE/EiJ

MDGA samples identifier	Mouse Gender	Overall call rate[b]		Sample Type[c]	Mouse Strain
		336 samples	362 samples		
SNP_mDIV_D5-342_120908	male	98.18	98.17	F1 hybrid	(PWD/PhJxC3H/HeJ)F1
DNA3300	male		98.17		C57BL/6J/CBA/CaJ
SNP_mDIV_D9-346_120908	male	98.15	98.16	F1 hybrid	(CAST/EiJxC3H/HeJ)F1
SNP_mDIV_B4-230_111308	male	98.15	98.15	F1 hybrid	(CAST/EiJxA/J)F1
SNP_mDIV_B10-87_090908	male	98.19	98.15	classical laboratory strain	C57BR/cdJ
SNP_mDIV_C9-246_111308	female	98.16	98.15	F1 hybrid	(PWK/PhJxNOD/ShiJ)F1
SNP_mDIV_B8-452_012209	male	98.16	98.14	wild-derived laboratory strain	DGA
SNP_mDIV_B3-229_111308	female	98.14	98.13	F1 hybrid	(CAST/EiJxA/J)F1
SNP_mDIV_B3-316_120908	male	98.16	98.12	classical laboratory strain	BPL/1J
SNP_mDIV_D1-249_111308	male	98.13	98.12	F1 hybrid	(WSB/EiJxCAST/EiJ)F1
SNP_mDIV_B8-234_111308	male	98.11	98.11	F1 hybrid	(CAST/EiJxNOD/ShiJ)F1
SNP_mDIV_C4-399_012709	male	98.14	98.11	classical laboratory strain	DBA/2DeJ
SNP_mDIV_D8-222_103008_3	male	98.12	98.09	CC-UNC G2:F1	OR906m104
SNP_mDIV_A10-223_111308	female	98.09	98.08	F1 hybrid	(A/JxPWK/PhJ)F1
SNP_mDIV_A6-SNP08_005_103108	male	98.08	98.07	F1 hybrid	(129X1/SvJxCAST/EiJ)F1
SNP_mDIV_D7-134_090908	male	98.13	98.07	classical laboratory strain	MRL/MpJ
SNP_mDIV_A8-225_112108	male	98.06	98.05	F1 hybrid	(C57BL/6JxCAST/EiJ)F1
SNP_mDIV_B8-161_082108	male	98.07	98.05	wild caught	MWN1279
SNP_mDIV_C3-240_111308	female	98.06	98.05	F1 hybrid	(PWK/PhJxA/J)F1
SNP_mDIV_A2-479_012709	male	98.06	98.03	wild-derived laboratory strain	BULS
SNP_mDIV_D2-250_111308	female	98.03	98.02	F1 hybrid	(WSB/EiJxPWK/PhJ)F1
SNP_mDIV_D3-251_111308	male	98.04	98.02	F1 hybrid	(WSB/EiJxPWK/PhJ)F1
SNP_mDIV_C7-209_103008_3	female	98.02	98.00	F1 hybrid	(PWK/PhJx129S1/SvImJ)F1
SNP_mDIV_B6-232_111308	male	98.01	97.98	F1 hybrid	(CAST/EiJxC57BL/6J)F1
SNP_mDIV_A3-480_012709	male	98.01	97.97	wild-derived laboratory strain	BUSNA
SNP_mDIV_C6-208_103008_3	male	97.97	97.96	F1 hybrid	(129S1/SvImJIxPWK/PhJ)F1
SNP_mDIV_C11-213_103008_3	female	97.97	97.96	F1 hybrid	(CAST/EiJx129S1/SvImJ)F1
SNP_mDIV_A5-53_082108	male	97.97	97.95	wild-derived laboratory strain	SF/CAMEiJ

MDGA samples identifier	Mouse Gender	Overall call rate[b]		Sample Type[c]	Mouse Strain
		336 samples	362 samples		
SNP_mDIV_C10-34_081308	male	97.99	97.95	wild-derived laboratory strain	PWD/PhJ
SNP_mDIV_B2-13_081308	female	97.97	97.94	CC-UNC G2:F1	OR1048f18
SNP_mDIV_A1-478_012709	male	97.96	97.92	wild-derived laboratory strain	STUS
SNP_mDIV_B10-236_111308	male	97.93	97.91	F1 hybrid	(CAST/EiJxPWK/PhJ)F1
SNP_mDIV_D11-189_082108	male	97.95	97.91	wild caught	BAG99
SNP_mDIV_A2-435_012209	male	97.91	97.88	wild-derived laboratory strain	MDH
SNP_mDIV_C6-243_111308	male	97.89	97.88	F1 hybrid	(PWK/PhJxC57BL/6J)F1
SNP_mDIV_A9-226_112108	female	97.85	97.85	F1 hybrid	(C57BL/6JxPWK/PhJ)F1
SNP_mDIV_C7-244_111308	female	97.85	97.84	F1 hybrid	(PWK/PhJxCAST/EiJ)F1
SNP_mDIV_A1-47_082108	male	97.86	97.83	wild-derived laboratory strain	MOLG/DnJ
SNP_mDIV_A5-158_091708	female	97.85	97.82	wild caught	RDS13554
SNP_mDIV_C11-248_111308	female	97.83	97.82	F1 hybrid	(WSB/EiJxCAST/EiJ)F1
SNP_mDIV_C5-242_111308	female	97.82	97.80	F1 hybrid	(PWK/PhJxC57BL/6J)F1
SNP_mDIV_B5-231_111308	female	97.77	97.76	F1 hybrid	(CAST/EiJxC57BL/6J)F1
SNP_mDIV_B11-237_111308	female	97.79	97.76	F1 hybrid	(CAST/EiJxWSB/EiJ)F1
SNP_mDIV_B1-445_012209	male	97.79	97.75	wild-derived laboratory strain	MBS
SNP_mDIV_B7-233_111308	female	97.70	97.68	F1 hybrid	(CAST/EiJxNOD/ShiJ)F1
SNP_mDIV_D3-469_012209	male	97.72	97.67	wild-derived laboratory strain	MCZ
SNP_mDIV_D11-46_081308	male	97.69	97.66	wild-derived laboratory strain	MOLD/RkJ
SNP_mDIV_C4-459_012209	male	97.68	97.65	wild-derived laboratory strain	22MO
SNP_mDIV_D7-255_111308_2	male	97.66	97.64	wild-derived laboratory strain	IS/CamRK
SNP_mDIV_D7-42_081308	female	97.65	97.62	CC-UNC G2:F1	OR1325f102
SNP_mDIV_B10-163_082108	female	97.60	97.59	wild caught	MWN1106
SNP_mDIV_B5-449_012209	male	97.59	97.57	wild-derived laboratory strain	MBT
SNP_mDIV_C5-169_082108	female	97.59	97.55	wild caught	BAG3
SNP_mDIV_B7-127_091708	male	97.59	97.54	wild-derived laboratory strain	CALB/RkJ
SNP_mDIV_D1-214_103008_3	female	97.54	97.53	F1 hybrid	(129S1/SvImJxCAST/EiJ)F1
SNP_mDIV_B5-16_081308	male	97.56	97.52	wild-derived laboratory strain	CZECHI/EiJ

MDGA samples identifier	Mouse	Overall call rate[b]		Sample Type[c]	Mouse Strain
	Gender	336 samples	362 samples		
DNA3258	male		97.50		C57BL/6J/CBA/CaJ
SNP_mDIV_C8-334_120908	male	97.51	97.49	F1 hybrid	(C57BL/6JxNOD/ShiJ)F1
SNP_mDIV_B11-395_012709	male	97.51	97.47	classical laboratory strain	C57BL/10ScNJ
SNP_mDIV_D4-39_081308	male	97.52	97.46	wild-derived laboratory strain	WSB/EiJ
SNP_mDIV_D5-471_012209	male	97.50	97.45	wild-derived laboratory strain	MH
SNP_mDIV_D11-477_012209	male	97.47	97.42	wild-derived laboratory strain	STUP
SNP_mDIV_B6-17_081308	male	97.46	97.41	wild-derived laboratory strain	CZECHII/EiJ
SNP_mDIV_B9-235_111308	female	97.44	97.41	F1 hybrid	(CAST/EiJxPWK/PhJ)F1
SNP_mDIV_C3-167_082108	male	97.44	97.40	wild caught	BAG74
SNP_mDIV_B4-448_012209	male	97.41	97.37	wild-derived laboratory strain	MBK
SNP_mDIV_D10-476_012209	male	97.40	97.35	wild-derived laboratory strain	STUF
SNP_mDIV_D8-43_081308	male	97.37	97.33	CC-UNC G2:F1	OR1325m106
SNP_mDIV_C4-168_082108	male	97.28	97.26	wild caught	BAG56
SNP_mDIV_C6-461_012209	male	97.29	97.25	wild-derived laboratory strain	CTP
SNP_mDIV_B3-190_082108	female	97.25	97.20	wild caught	BAG68
SNP_mDIV_B10-454_012209	male	97.18	97.15	wild-derived laboratory strain	MGA
SNP_mDIV_C2-24_081308	male	97.15	97.13	wild-derived laboratory strain	MOLF/EiJ
SNP_mDIV_C1-396_012709	male	97.16	97.10	classical laboratory strain	C57BL/10ScSnJ
SNP_mDIV_D10-45_081308	male	97.11	97.08	wild-derived laboratory strain	JF1/Ms
SNP_mDIV_B5-486_022709	male	97.03	96.98	wild caught	Yu2097m
SNP_mDIV_B8-489_022709	male	96.91	96.87	wild caught	Yu2115m
SNP_mDIV_B6-487_022709	female	96.90	96.85	wild caught	Yu2099f

[a] Samples are ordered in highest to lowest overall genotyping call rate based on the 364 sample set

[b] overall call rate refers to the total number of genotypes determined as AA, AB, or BB for each sample

[c] samples without a sample type listed are Hill laboratory samples

Appendix D – Genetic distance matrix samples used to create phylogenetic trees

This section contains examples of genetic distance matrices that were used to create the phylogenetic trees.

Table D-1: Genetic distance matrix representing each of the seven sample types used for generating phylogenetic trees.

	Sample								
Sample	**OR1611m121**	**MOR/RkJ**	**BXD32**	**(CAST/EiJx PWK/PhJ)F1_1**	**Yu2099f**	**MBT**	**129P3/J**	**C57BL/6J_1**	**KCT222**
OR1611m121	0	0.39086	0.41793	0.47254	0.51901	0.49304	0.41189	0.40568	0.42097
MOR/RkJ	0.39086	0	0.21885	0.56400	0.48733	0.48371	0.25829	0.11432	0.27100
BXD32	0.41793	0.21885	0	0.55446	0.44952	0.45648	0.18582	0.14327	0.23382
(CAST/EiJx PWK/PhJ)F1_1	0.47254	0.56400	0.55446	0	0.35138	0.30942	0.55158	0.55330	0.52558
Yu2099f	0.51901	0.48733	0.44952	0.35138	0	0.22966	0.44356	0.44492	0.43387
MBT	0.49304	0.48371	0.45648	0.30942	0.22966	0	0.45397	0.45325	0.43341
129P3/J	0.41189	0.25829	0.18582	0.55158	0.44356	0.45397	0	0.19394	0.23746
C57BL/6J_1	0.40568	0.11432	0.14327	0.55330	0.44492	0.45325	0.19394	0	0.23238
KCT222	0.42097	0.27100	0.23382	0.52558	0.43387	0.43341	0.23746	0.23238	0

Table D-2: Genetic distance matrix for C57BL/6J samples and the CBA/CaJ sample in The Jackson Laboratory dataset.

	Sample								
Sample	**CBA/CaJ**	**C57BL/6J_1**	**C57BL/6J_6**	**C57BL/6J_7**	**C57BL/6J_8**	**C57BL/6J_5**	**C57BL/6J_4**	**C57BL/6J_3**	**C57BL/6J_2**
CBA/CaJ	0	0.17828	0.17852	0.17875	0.18422	0.17781	0.17839	0.17953	0.17801
C57BL/6J_1	0.17828	0	0.00195	0.00227	0.00917	0.00095	0.00177	0.00326	0.00125
C57BL/6J_6	0.17852	0.00195	0	0.00236	0.00937	0.00121	0.00197	0.00343	0.00149
C57BL/6J_7	0.17875	0.00227	0.00236	0	0.00965	0.00152	0.00227	0.00366	0.00176
C57BL/6J_8	0.18422	0.00917	0.00937	0.00965	0	0.00845	0.00914	0.01045	0.00871
C57BL/6J_5	0.17781	0.00095	0.00121	0.00152	0.00845	0	0.00101	0.00252	0.00049
C57BL/6J_4	0.17839	0.00177	0.00197	0.00227	0.00914	0.00101	0	0.00320	0.00125
C57BL/6J_3	0.17953	0.00326	0.00343	0.00366	0.01045	0.00252	0.00320	0	0.00276
C57BL/6J_2	0.17801	0.00125	0.00149	0.00176	0.00871	0.00049	0.00125	0.00276	0

Table D-3: Genetic distance matrix example used to generate phylogenetic trees for all samples containing three tissues types from the Hill laboratory samples.

	Sample								
Sample	**911.50 Li-2**	**911.17 Li**	**911.17 Cl**	**911.49 Li**	**911.17 Sp**	**911.50 Cl**	**911.49 Cl**	**911.50 Sp**	**911.49 Sp**
911.50 Li-2	0	0.067051	0.064964	0.074507	0.066062	0.004296	0.067783	0.003377	0.067816
911.17 Li	0.067051	0	0.006448	0.084721	0.007757	0.066143	0.078524	0.065374	0.078497
911.17 Cl	0.064964	0.006448	0	0.08282	0.004626	0.06342	0.075835	0.062567	0.075799
911.49 Li	0.074507	0.084721	0.08282	0	0.083449	0.073643	0.010593	0.072823	0.010509
911.17 Sp	0.066062	0.007757	0.004626	0.083449	0	0.064734	0.077087	0.063883	0.077054
911.50 Cl	0.004296	0.066143	0.06342	0.073643	0.064734	0	0.065905	0.001284	0.065897
911.49 Cl	0.067783	0.078524	0.075835	0.010593	0.077087	0.065905	0	0.065123	0.000824
911.50 Sp	0.003377	0.065374	0.062567	0.072823	0.063883	0.001284	0.065123	0	0.065114
911.49 Sp	0.067816	0.078497	0.075799	0.010509	0.077054	0.065897	0.000824	0.065114	0

Table D-4: Genetic distance matrix example used to generate phylogenetic trees for tissue replicates of a B6 mouse from the Hill laboratory samples

	Sample					
Sample	**300.7 Sp-3**	**300.7 Sp-2**	**300.7 Cl-1**	**300.7 Sp-1**	**300.7 Cl-3**	**300.7 Cl-2**
300.7 Sp-3	0	0.002947	0.004195	0.003859	0.002407	0.004235
300.7 Sp-2	0.002947	0	0.004218	0.003861	0.002349	0.004245
300.7 Cl-1	0.004195	0.004218	0	0.004982	0.003683	0.005312
300.7 Sp-1	0.003859	0.003861	0.004982	0	0.003286	0.004917
300.7 Cl-3	0.002407	0.002349	0.003683	0.003286	0	0.00367
300.7 Cl-2	0.004235	0.004245	0.005312	0.004917	0.00367	0

Susan Eitutis

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

EDUCATION:

The University of Western Ontario, London, Ontario **September 2013**
M.Cl.Sc – Audiology

The University of Western Ontario, London, Ontario **September 2011 – present**
M.Sc. – Molecular Genetics

The University of Western Ontario, London, Ontario **September 2007 – April 2011**
B.Sc. Honours Specialization Genetics

AWARDS:

The University of Western Ontario **September 2011 – present**

- Western Graduate Research Scholarship (2011-2013)

Provincial Scholarships

- Received Ontario Graduate Scholarship (three terms beginning September 2013)

Teaching Awards

- Nomination for teaching assistant award (2013)

Awards for presentations

- Environmental Mutagen Society Student and New Investigator Travel Award ($500) (2012)
- 1st place poster presentation, Lawson research Day ($500 travel award) (2012)
- 2nd place poster presentation, Biology Graduate Research Forum award for excellence in scientific communication (2011)

The University of Western Ontario **September 2007 – April 2011**
B.Sc Honours Specialization Genetics

- Dean’s Honour list (2009-2010; 2010-2011)
- 2007 University of Western Ontario Entrance Scholarship (2007)

PRESENTATIONS:

Abstract submitted to international meeting

Eitutis ST, Locke MEO, Wishart AE, Daley M, Hill HA. Array-Based Genomic Diversity Measures Portray *Mus musculus* Phylogenetic and Genealogical Relationships, and Detect Genetic Variation Among C57BL/6J Mice and Between Tissues of the Same Mouse. Environmental Mutagenesis and Genomics Society, September 2013.

Marshall AE, Locke MEO, **Eitutis ST**, Daley M, Hill KA. An Array-Based Genomic Survey of Copy Number Variation Across Wild Caught and Inbred Mice Implicates Different Environmentally-Responsive Candidate Genes. Environmental Mutagenesis and Genomics Society, September 2013.

Wishart AE, Locke MEO, **Eitutis ST**, Daley M, Hill KA. The first application of three copy number variant detection pipelines for the Mouse Diversity Genotyping Array: metrics of concordance. Environmental Mutagenesis and Genomics Society, September 2013.

Peer-reviewed international meeting presentation

Eitutis ST, Wishart AE, Hill KA. Somatic mosaicism detected using the Mouse Diversity Genotyping Array reveals tissue-specific mutation patterns associated with the *harlequin* phenotype. Mouse Molecular Genetics, October 2012; *poster presentation.*

Wishart AE, Locke MEO, **Eitutis ST**, Hill KA. Patterns of recurrent and tissue-specific copy number variants in the mouse genome. Mouse Molecular Genetics, October 2012; *poster presentation.*

Hill KA, Locke MEO, Wishart AW, **Eitutis ST**, Butler J, Daley M. Hot or Not? Leveraging mouse genome diversity to identify hotspots of copy number variants. Mouse Molecular Genetics, October 2012; *platform presentation*.

Eitutis ST, Wishart AE, Hill KA. Genome-wide mutation detection using the Mouse Diversity Genotyping Array: Evidence for a mutation signature associated with a premature aging phenotype. American Society of Human Genetics October, 2011; *poster presentation.*

Wishart AE, **Eitutis ST**, Hill KA. Evidence for an altered profile of copy number changes in the cerebellum of the *harlequin* mouse mimic of human aging-associated neurodegeneration**.** American Society of Human Genetics October, 2011; *poster presentation.*

Eitutis ST, Wishart AE, Hill KA. Beyond Single Gene Mutation Targets: An Array Based, Genome-Wide Approach To The Study Of Somatic Mosaicism. Environmental Mutagen Society September, 2012; *poster presentation.*

Wishart AE, Eitutis ST, Hill KA. Patterns of Copy Number Changes Between Spleen and Cerebellum in the *harlequin* Mouse Model of Mitochondrial Dysfunction. Environmental Mutagen Society September, 2012; *poster presentation.*

Eitutis ST, Wishart AE, Hill KA. The Mouse Diversity Genotyping Array Profiles Tissue- and Genotype-Specific Mutations Across The Mouse Genome. Environmental Mutagen Society October, 2011; *oral and poster presentation.*

Wishart AE, **Eitutis ST**, Hill KA. Copy Number Changes Across the Mouse Genome Discovered Using the Mouse Diversity Genotyping Array Show Tissue and Genotype Specificity. Environmental Mutagen Society October, 2011; *poster presentation.*

Hill KA, **Eitutis ST**, Wishart AE. Genome-wide mutation analysis in a mouse mimic of human aging-associated neurodegeneration using the novel Mouse Diversity Genotyping Array. Mouse Genetics June, 2011; *poster presentation.*

Peer-reviewed regional/local meeting presentation

Butler JL, **Eitutis ST**, Locke MEO, Wishart AE, Daley M, Hill KH. HD-CNV: Hotspot detection of copy number variants. Lawson Research Day, 2012; *poster presentation.*

Eitutis ST, Wishart AE, Hill KA. A genome-wide genotyping array detects clusters of mutations across the genome of the *harlequin* mouse model of neurodegeneration. University of Western Ontario Biology Day March, 2011; *oral presentation.*

Wishart AE, **Eitutis ST**, Hill KA. A novel high-resolution genotyping array for the mouse detects putative copy number changes associated with a neurodegenerative phenotype. Lawson Research Day March, 2011; *poster presentation.*

Eitutis ST, Wishart AE, Hill KA. Genome-wide mutation analysis using the novel Mouse Diversity Genotyping Array. Lawson Research Day March, 2011; *poster presentation*.

Eitutis ST, Wishart AE, Hill KA. A genome-wide genotyping array detects clusters of mutations in the genome of the *harlequin* mouse model of neurodegeneration. Ontario Biology Day March, 2011; *oral presentation*.