8-10-2018 3:00 PM

A bioinformatics analysis of contributors to false discovery for a mouse genotyping array

Nisha Patel
The University of Western Ontario

Supervisor
Hill, Kathleen A.
The University of Western Ontario

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

Abstract

Microarray experiments employing massively-parallel hybridization are valuable for the study of genetic variation, however, errors during hybridization and limitations of single-species design must be considered for use within and across species. The Mouse Diversity Genotyping Array (MDGA) is a low cost, high-resolution microarray with probes that bind to target DNA for variant detection. Errors associated with probe design and incomplete protein removal from target DNA lead to false discovery and thus necessitate examination of probe suitability and target DNA availability. Bioinformatics methods were used to carry out confirmation of probe annotations, assessment of DNA accessibility for hybridization to probes, and prediction of the theoretical ability of MDGA probes to hybridize cross-species to naked mole-rat genomic DNA. The results are a filtered probe list demonstrated to reduce false discovery, a suggested approach to assess biases arising from protein-bound DNA, and predictions for cross-species application of the MDGA to naked mole-rat samples.

Keywords

Mouse Diversity Genotyping Array, *Mus musculus*, hybridization, microarray probes, copy number variants, single nucleotide polymorphism, SNP genotyping, cross-species hybridization, naked mole-rat

Co-authorship statement

Nisha Patel completed the work presented in this thesis under the supervision and financial support from Dr. Kathleen Allen Hill. This thesis is presented in monographic format. Nisha Patel performed the bioinformatics analyses presented in this thesis and contributed to a publication by *Locke et al.*[1]. Dr. Kathleen Hill is a senior author on all publications arising from the research presented due to her role in project design, supervision, literature research, data analysis, and assistance with writing. The following people contributed to the publication: M Elizabeth O Locke, Maja Milojevic, Susan T. Eitutis, Andrea E. Wishart, and Mark Daley.

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Kathleen Hill, for the patient guidance, encouragement, and support she has provided throughout the duration of my time as her student. I am extremely lucky to have a supervisor who cared so much about my progress and stood by me to the end. Her knowledge and patience were essential to the progression and completion of this thesis. I would also like to express gratitude to my committee advisors Dr. Susanne Kohalmi and Dr. Shiva Singh who have continued to support me throughout the duration of my project.

Additionally, I must thank all members of the Hill Laboratory, past and present, for their assistance and friendship. I especially appreciate Beth Locke, Maja Milojevic, Rachel Dawn Kelly, and Freda Qi for their willingness to tackle problems together. Lastly, I wouldn't have been able to make it through the ups and downs during my years as a student if it weren't for Nilesh Patel. I can't thank him enough for all his support.

Table of contents

List of figures

List of tables

List of abbreviations

16S rRNA	16S ribosomal RNA
aCGH	Array Comparative Genomic Hybridization
ANOVA	Analysis of Variance
B6	C57BL/6J
BED	Browser extensible data [*file extension*]
BLAST	Basic Local Alignment Search Tool
bp	Base pairs
CEL	Cell intensity file [*file extension*]
CGD	Center for Genome Dynamics
CGH	Comparative genomic hybridization
CNV	Copy number variant
DAVID	Database for Annotation, Visualization, and Integrated Discovery tools
DHS	Deoxyribonuclease I hypersensitive site
DNase I	Deoxyribonuclease I
DNA	Deoxyribonucleic acid
dsDNA	Double-stranded DNA
GO	Gene Ontology
ENCODE	Encyclopedia of DNA Elements
FASTA	FAST-All [*file type]*
FTP	File Transfer Protocol
IGP	Invariant genomic probe
kb	Kilobase

Mb	Megabase
MDGA	Mouse Diversity Genotyping Array
MGI	Mouse Genome Informatics
MT	Mitochondria
NCBI	National Center for Biotechnology Information
PCR	Polymerase chain reaction
RE	Restriction enzyme
SD	Segmental duplication
SNP	Single nucleotide polymorphism
SV	Structural variation
T_m	Melting temperature
UCSC	University of California Santa Cruz

Chapter 1 Introduction

1.1 Research motivation

The study of genomic variation enables us to understand the genetic makeup of an individual, a population's genetic structure, and the functional consequences of genetic variants that underlie health, disease, and evolution. The laboratory mouse is a valuable organism for understanding genetic variation across the genome for many reasons including convenience, ease of study (i.e., an abundance of existing mouse resources and tools), and cost-effectiveness. While historically useful, the laboratory mouse does not represent all of the variation found in nature. For that reason, it is also important to study non-model organisms that collectively capture a greater breadth and depth of variation present across species in the natural world.

There are two major challenges in the study of genetic variation in model and non-model organisms: the lack of low cost, high-throughput genotyping technologies and errors associated with massively-parallel hybridization that lead to false discovery of genetic variants. Microarray technologies employ massively-parallel hybridization to detect genetic variants however experiments that rely on the simultaneous hybridization of sample DNA to millions of probes to assess hundreds of thousands of loci are error-prone. The Mouse Diversity Genotyping Array (MDGA), a microarray for the study of genetic variation in Mus species, is no exception. Inaccuracies in microarray data can lead to an inaccurate and poor understanding of true biological variation. A valuable approach in assessing potential sources of error in microarray data is the use of bioinformatics methods.

Errors associated with probe design and annotation are not uncommon and validation of probe suitability is necessary[1–3]. A microarray experiment requires successful hybridization of probe to target DNA, however any errors in sample target DNA isolation can leave DNA inaccessible for hybridization to probes[4]. Assessment of array data is therefore required. Furthermore, the lack of low cost, high-resolution genotyping technologies for non-model organisms requires alternate approaches to carry out

population genetics studies. Theoretically predicting probe suitability for cross-species study in naked mole-rat genomic DNA can demonstrate the potential applicability of the MDGA for cross-species use and naked mole-rat studies.

1.2 Variation exists in genomes both within and across species

Genomic variation exists within an organism, between individuals of the same population, between populations, and between species. This diversity in DNA sequences of genomes is created by mutations and the specific DNA sequences that differ across genomes are referred to as genetic variants. Genetic variants can encompass stretches of DNA from as small as one base pair to millions of base pairs. When a single nucleotide at a specific genomic position differs in a number of individuals relative to the population (at least one percent of the population), it is referred to as a single nucleotide polymorphism (SNP). Variants can also exist as large-scale structural variants whereby large segments of the genome are affected. Structural variants that involve the loss or gain of segments of the genome are referred to as copy number variants (CNVs). Characterizing CNVs and SNPs is important in order to determine the extent of variation between individuals, populations, and species and also to study the role of variation in disease and evolution.

The most frequent type of genetic variant is a SNP[5]. Polymorphism for a single nucleotide refers to the existence of one of two possible alleles at a particular genomic position; determining which alleles exist at specific SNP sites is called SNP genotyping. More than 1.4 million SNPs were identified in the first initial sequencing of the human genome[6]. Since then, SNPs have been identified in many populations of many organisms and there are databases of known SNPs[7,8]. SNPs are found to occur on average every 300 nucleotides with about 10 million SNPs in the human genome[9]. Polymorphic sites in the genome are common, may affect fitness, and are important in evolution[10]. Determining the frequency of these alleles at polymorphic sites across populations and species directly informs researchers of the genetic diversity present in organisms and the population structure across organisms.

Copy number variation involves deletions or duplications of segments of DNA greater than 500 bp in size[1]. The deletions and duplications alter the diploid state of DNA where diploid refers to two sets of inherited chromosomes (maternal and paternal) and therefore two copies of each genomic locus along a chromosome. Deletions and duplications can create differences in copy number state of certain segments of DNA relative to the reference genome (copy number state of 2 for a diploid genome). CNVs, by virtue of encompassing large segments of the genome, play a major role in genomic variability[11,12].

Rates of evolution and evolutionary relationships between populations and species

The variable copy numbers that persist in populations can give insight into rates of evolution as well as evolutionary relationships between populations and species. The rate of adaptive change is directly influenced by the gene duplication rate[13]. Gene duplications in particular persist in populations due to their role in generating novel genes that have an evolutionary advantage[14]. For example, homologs of the AMY1 (Amylase, alpha 1A) locus in the human genome are found across a variety of primates suggesting the *AMY1* gene must have arisen in an ancestor of present day humans and primates[15]. CNVs that persist in populations as duplications are considered to be segmental duplications[16]. Segmental duplications (SDs) are segments of duplicated DNA (> 1 kb) that are highly homologous with high sequence identity (or 90% or greater sequence identity among duplicates). SDs are associated with a significant proportion of novel CNVs and permit direct assessment of rates of evolution where fewer SDs suggest a slower pace of evolution[15,17,18]. CNVs lead to adaptive variability and can provide much insight into the evolution of species.

1.3 The laboratory mouse is a valuable organism for studies of genetic variation

The mouse serves as an invaluable model organism for the study of genetic variation due to the existence of a fully sequenced and annotated genome, comprehensive genetic databases and tools, and the wide depth and breadth of pre-existing literature. Characteristics such as its relatively small size and short generation time make the mouse a very economically viable and practical mammalian organism for genomic studies in a

laboratory setting. The use of mice (*Mus musculus*) in research can be traced back to at least the 1800s and as a result, much is known about the laboratory mouse[19]. Laboratory mouse strains that are in use today were started in 1921 by the mating of two mice code-named C52 and C57 for over 200 generations[20]. One of the oldest, most widely used, and best characterized inbred mouse strains arising from this inbreeding is the C57BL/6J (B6). The B6 mouse is the second mammalian species after the human to have its DNA sequenced[21].

The laboratory mouse serves as a model organism for the study of human CNV due to genomes of similar size, content, and organization[22]. Shared evolutionary ancestry is evident from homologous genes being located in large blocks of syntenic regions as well as analyses showing 40% sequence alignment between mouse and human genomes[23]. About 99% of mouse genes have a homologue in the human genome and 80% of mouse genes have an orthologous counterpart in the human genome[21]. Not only have CNVs been detected in both the mouse and human genomes, but the variants also show a high degree of sequence identity, making mouse CNV research relevant to CNV profiles in humans[11,12,24,25].

1.4 Measuring single nucleotide polymorphism diversity and detecting copy number variation: Current technologies

Technological advances over the last decade have allowed researchers to conduct genome-wide studies of variation in humans and across species. Earlier studies using cytogenetic techniques such as fluorescent *in situ* hybridization (FISH) allowed for the observation of the physical structure of chromosomes. FISH involves the binding – or hybridization – of highly complementary probe sequences of different lengths and DNA sequences on the chromosome; fluorescent labeling of the probes allows for the detection of specific sequences on a chromosome. While cytogenetic techniques were initially useful for the detection of larger variants, major disadvantages are low genomic resolution and poor detection of smaller variants[26]. To date, the detection of SNPs and CNVs across the genome is primarily carried out through one of two approaches: next-generation sequencing (NGS) and microarray technology. Sequencing techniques today

are far superior to the original Sanger sequencing technique used to first sequence the human genome in the early 2000s for the Human Genome Project[6,27]. Present day sequencing is carried out using NGS methodologies for both SNP and CNV detection[28,29]. While NGS technologies are powerful, they are also very expensive and not feasible for many population surveys of genome-wide genetic variation. CNV detection also required development of appropriate bioinformatics approaches which have only recently been refined[30–32]. Microarray technology is currently an important tool, particularly in studies of model (and non-model) organisms. Microarrays allow for the detection of variants by employing massively-parallel hybridization – the simultaneous binding of millions of probes to complementary target DNA from samples of interest. Microarray technology has been useful for SNP genotyping and more recently, whole-genome CNV discovery across a variety of species[33–35].

A microarray, commonly known as an array, is a small chip that has millions of microscopic probes attached to its solid surface (typically a microscopic slide). The physical space on a chip along with the attached probes is referred to as a DNA feature. Probes are synthesized directly onto the surface of the chip using a special printing process called photolithography that involves relies on UV light and chemical synthesis technologies. The probes are single-stranded DNA molecules that are designed to be complementary at one location to a fully sequenced reference genome of the species. Each probe has a particular sequence that under optimal conditions can recognize and specifically bind to sample DNA according to the principle of complementary base pairing – a thermodynamic process called hybridization. Ultimately, it is the hybridization of target sample DNA to probe sequences on the array that allows for the detection of SNPs and CNVs. The probes themselves can range in size from 10 bases to many kilobases. Each probe consists of a unique nucleotide probe sequence associated with a genomic position in the reference genome.

There are two types of microarrays that employ hybridization for CNV detection: SNP arrays and comparative genomic hybridization (CGH) arrays. Array-based CGH requires co-hybridization of both test and reference sample DNA for the detection of copy number

variants (through comparison). Array-based CGH is limited in use because it is not designed for SNP genotyping. For this reason, the SNP array is the most commonly used microarray for SNP and CNV genotyping. The SNP array was initially developed for the simultaneous genotyping of SNP alleles at many known loci of a genome with probes designed to detect one of two alleles at a single locus. Additionally, the SNP array detects CNVs through probes designed specifically for regions of the genome that may exist in copies of a variable number. Copy number gains and losses are determined by comparing the number of copies of specific DNA sequences to the reference genome. The SNP array is a powerful tool for genome-wide SNP genotyping and CNV discovery due to the hybridization of millions of DNA sequences to probes, referred to as massively-parallel hybridization.

The underlying principle of microarray technologies is the successful and efficient hybridization of millions of probes to complementary strands of target sample DNA. The structure of double-stranded DNA (dsDNA) and more specifically, the binding affinity between two strands of DNA, is affected by thermodynamic factors – heat, work, and temperature – in predictable ways[36]. Hybridization between two complementary strands of DNA involves interactions between the nucleotide bases to result in an energetically preferred single complex, referred to as a duplex. An increase in the number of complementary base pairs of a given double-stranded DNA sequence results in a more stable, hybridized duplex due to more hydrogen bonds and stronger hydrophobic interactions between the base pairs. Thus, the binding affinity between strands of complementary sequences (probe to target) is affected by the degree of mismatching between the two sequences.

1.5 The Mouse Diversity Genotyping Array is a high-throughput tool for the study of genetic variation in the mouse

The Mouse Diversity Genotyping Array (MDGA; Affymetrix®, Santa Clara, CA) is a SNP microarray designed for the detection of both SNPs and CNVs in the mouse genome. It is the first mouse array that has the ability to capture genetic variants across

the mouse phylogeny[37] and captures genetic variation in laboratory mice including the highly inbred classical and genetically diverse wild-derived mouse strains. Two distinct probe types, SNP probes and invariant genomic probes (IGP), were designed for SNP and CNV detection across the genus Mus. There are about 4.9 million SNP probes that target known SNPs at 623,124 loci across 8 mouse strains and about 1.8 million IGPs that target about 200,000 exons of genes[37]. A unique genomic position, or locus, is targeted by more than one probe to account for both the sense and antisense DNA strands and potential alleles (for SNPs). In aggregate, probes were designed such that genic sequences of exons for CNV detection and sequences containing SNPs for SNP detection can be targeted. Probes cover the genome across all chromosomes with even distribution.

All probes on the MDGA have 25mer sequences that target sample DNA sequences based on complementary base pairing[36,37]. The probes are designed such that the probe and target DNA sequences hybridize with high specificity. High specificity refers to more specific binding i.e., probe sequences hybridize more readily to complementary target sequences when there is no mismatch in base pairs. The ability to call SNP genotypes and CNVs depends on successful hybridization between probe-to-target DNA from samples of interest. DNA with base pair mismatches to the probe sequence are less likely to bind to the probes and remain bound as stable duplexes[36]. Only duplexes formed between probe-to-target DNA are observable for SNP and CNV analysis.

Hybridization of probe to fluorescently labelled target DNA results in the emission (upon excitation by a scanner) of an observable fluorescent intensity that can be visualized and analyzed. Relative measures of fluorescent intensities indicate hybridization success and higher relative measures are associated with higher amounts of target DNA of interest. Higher concentrations of target DNA bound to a certain probe spot results in a relatively higher fluorescent intensity measure at the probe spot[38,39]. Detection of a SNP or copy number event is usually inferred from the probe signal intensity when compared against the reference genome. A probe with high specificity should theoretically provide a signal only in the presence of the target molecule. Each probe is associated with a region in the mouse genome and can thus be computationally analyzed for CNV or SNP detection.

SNP probes exist in sets of eight that target a single SNP locus (Figure 1)[37]. At each SNP locus, the probes are capable of detecting one of two possible alleles, designated as allele A or allele B. A SNP probe set is comprised of four probes targeting the SNP locus on the sense strand and four probes targeting the SNP locus on the antisense strand. Of the four probes designed for each strand, two probes target the first possible allele (allele A) and two probes target the second possible allele (allele B) at a known position. The four SNP probes on each strand differ in the one base at the SNP location that can be either allele A or B. Despite targeting the same SNP alleles, two of the four SNP probes targeting each SNP allele can be offset by up 10 bp apart. SNP probes are designed to be redundant for more accurate SNP genotyping. Redundancy is achieved by having two identical probe sequences for each strand and for each allele and by having the probes slightly offset. SNP probes are useful not just for identifying SNPs, but also for CNV analysis. The redundancy in the SNP probes gives greater confidence in calls and is used in concert with IGP probes in CNV calling.

The MDGA also contains about 1.8 million IGPs devoid of SNPs that target 916,269 unique exonic regions in the mouse genome with two IGPs per exonic locus (Figure 1)[37]. IGPs were designed to target 93.4% of over 200,000 exons (Ensembl version 49). Since the mouse genome contains about 200,000 exons, the MDGA provides good coverage of the exons in the mouse[40,41]. Each exon is covered by three unique IGPs, one at each of the proximal, medial, and distal locations on both the sense and antisense strand of the exon. A total of 6 IGPs comprise an IGP set to target one exon and can be referred to as an IGP set. IGPs. All IGPs on the MDGA have the ability to detect CNVs as either deletions or duplications relative to the reference genome.

1.5.1 Elements of the experimental pipeline can affect false discovery

A microarray experiment involves preparation of sample DNA for hybridization to the chip. Detection of genetic variants is possible only through the hybridization of isolated, pure target DNA to complementary probe sequences on the chip; incomplete DNA isolation can affect discovery of variants[4]. Fluorescence of varying intensities from probe-to-target DNA hybridization are converted into raw fluorescent intensity values

that can be computationally analyzed. Various algorithms use probe annotations, along with fluorescent intensity data for each probe, to assign genotypes at known SNP loci and call for putative CNVs. Issues with target DNA or probe adversely affect hybridization success

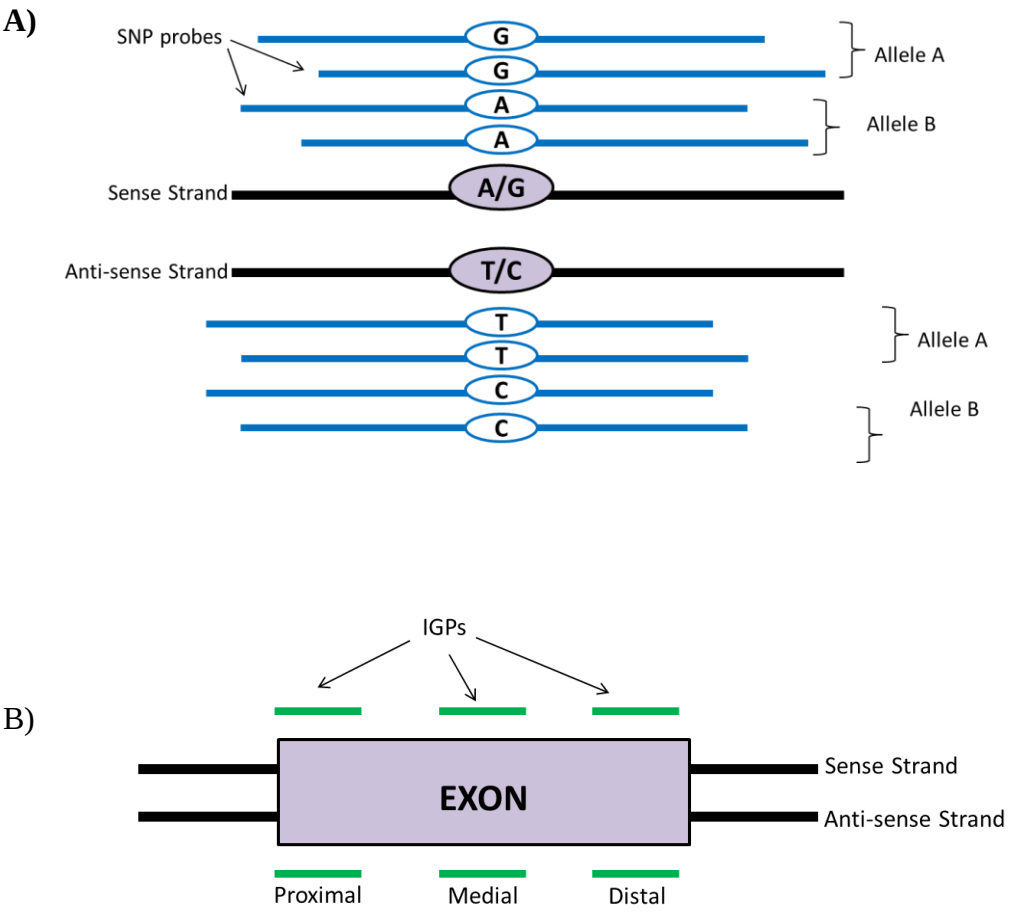

Figure 1 The Mouse Diversity Genotyping Array (MDGA) has two types of probes, single nucleotide polymorphism (SNP) probes for SNP genotyping and invariant genomic probes (IGPs) for copy number variation (CNV) calling. (A) SNP probes exist in sets of eight that target a single SNP locus to detect one of two possible alleles, allele A and allele B. At each SNP locus, four SNP probe sequences target SNP allele A and four SNP probe sequences target SNP allele B. Two of the four SNP probes targeting

each SNP allele can be offset by up to ten base pairs apart. **(B)** Invariant genomic probes (IGPs) target proximal, medial, and distal regions of the exons of genes along both sense and antisense strands. An IGP set consists of six IGPs that target a Mus exon.

and can lead to false discovery[42].

Isolated DNA is hybridized to the MDGA probes

Four general steps are involved prior to hybridization: DNA extraction, restriction enzyme digestion, polymerase chain reaction (PCR), and fluorescent tagging of sample processed DNA. DNA extraction kits with standardized protocols for optimal yields of pure, uncontaminated DNA are commonly used. However, certain tissues may require more DNA purification steps to remove higher levels of tissue-specific contaminants, such as proteins, lipids, and RNA. If protein-removal is ineffective, sample DNA is unable to undergo restriction enzyme digestion – important for complexity reduction of the genome prior to hybridization. Optimally-sized DNA fragments are amplified and fluorescently tagged such that hybridization of the DNA to a probe results in emission of fluorescence. The fluorescence associated with each probe depends on the amount of target DNA in the sample as well as the affinity between target and probe sequences. A scanner is then used to detect the relative fluorescent intensities at each array spot and the information is outputted as raw fluorescent intensity data in a file format called CEL.

SNP genotyping and CNV calling are carried out computationally using algorithms

Since the raw fluorescent intensities do not provide sufficient accuracy in identifying a SNP, a step called SNP genotype calling is performed. SNP genotyping involves applying the BRLMM-P algorithm to read each SNP location in the genome and estimating the SNP allele present (Affymetrix® Power Tools from the Affymetrix Genotyping Console™)[43]. In a diploid genome, each SNP is composed of two allele calls from two homologous chromosomes, with A and B representing the two possible alleles. Estimation of the SNP allele at each SNP locus is based on the probability of probe signals being grouped into clusters representing one of three SNP genotypes: heterozygous AB, homozygous AA, or homozygous BB. All signals for each genotype

group should look similar. If a SNP signal does not meet a high enough probability (normally a percentage of calls of over 90% to 97%) to cluster into one of the three genotype groups, it is discarded as a no call and in doing so, removes any ambiguous results that may arise from low quality DNA, poor hybridization, or poor chip quality[32,43,44]. A percent call is simply the number of SNP loci that are given a genotype call of AA, AB, or BB divided by the total number of SNP loci on the chip. Any errors associated with SNP probes or SNP probe to target DNA hybridization can affect percent calls and therefore false discovery rates.

Putative CNVs are called using PennCNV, an open source software tool that uses an algorithm called the Hidden Markov Model (HMM)[45]. HMM models designed for CNV detection are based on the assumption that copy number states of probes are associated with specific probabilities. Specifically, the HMM is used to determine the state of a given probe by using its fluorescent intensity value and the state of the closest prior probe on DNA landscape. CNV calls are made as gains or losses relative to the reference diploid fluorescent intensity of two (diploid cells contain two homologous copies of each chromosome). A higher fluorescent intensity at a given microarray spot indicates that higher levels of hybridization has occurred i.e., that there are relatively more copies of a sample that have hybridized to the probes[38,39]. A deletion refers to a loss of one or two copies and a duplication refers to a gain of one or more copies of the particular genomic segment. The distance between probes plays a factor in the likelihood of the copy numbers for the probes being the same. Furthermore, a minimum of three consecutive probes of the same copy number state covering a region is required for a putative CNV to be called[1]. These consecutive probes must be in the correct order for the CNV call to be made. All CNV calls must also be between 500 bp to 1 Mb in size[1]. It is important to note that any errors associated with target DNA and IGP hybridization can adversely affect calls and lead to the false discovery of CNVs.

1.6 Massively-parallel hybridization is successful under ideal conditions

Conditions for hybridization directly influence the binding specificity between the target and probe sequences. Ideal conditions for hybridization are determined from the melting temperature (T_m) of a particular sequence where T_m is the melting temperature at which half of the total number of DNA strands are free (single-stranded) and half are double-stranded (occupied)[36]. T_m depends on the length of the DNA sequence and its specific nucleotide composition. The T_m is calculated based on the energy required to separate hybridized strands of a sequence. High temperatures cause the dsDNA to dissociate and exist as two independent strands (the energy of all molecules is weaker)[36]. Ideal conditions for a microarray hybridization are estimated from the T_m for the specific probe and DNA sequences where hybridization between perfectly complementary sequences occurs favourably to create a significantly more thermodynamically stable duplex (compared to mismatched sequences)[46].

Hybridization stringency refers to the extent to which hybridization between mismatched sequences can occur and is directly influenced by how strict hybridization conditions are set. Conditions are set to manipulate binding specificity between probe and target DNA[42]. High stringency conditions allow for more specific binding (i.e., binding between mismatched sequences is less favourable). However, if stringency conditions are too high, the probe cannot bind to its target readily because conditions are too demanding and if stringency conditions are too low, the probe is more likely to bind to target DNA that isn't complementary (low specificity). Hybridization stringency is related to the purpose of hybridization. High stringency conditions are typical when highly specific binding between probe and target sequence bases is required such as when targeting particular DNA in the genome. However, low stringency conditions can be used when some sequence mismatch is expected, such as in the case of inter-organismal comparisons[47,48].

Hybridization stringency is affected by four major factors: temperature, salt, sequence length, and sequence composition. High stringency conditions involve increasing the temperature closer to the T_m of the DNA molecules or decreasing the salt concentration

which results in increased specificity i.e., binding to be more specific[42,49]. Conversely, low stringency conditions are achieved by decreasing temperature well below the T_m of the DNA molecules or increasing salt concentrations (with salt referring to a saline solution that contains sodium chloride and sodium citrate).

Probe length is another factor that greatly influences hybridization. Longer probes hybridize to complementary target sequences under more stringent conditions while shorter probes hybridize optimally under less stringent conditions. It is therefore important that probes are designed to be similar in length and have a uniform T_m in order to have a shared thermodynamic profile under which hybridization can occur[42]. Differently sized probes that do not share the uniform T_m bind less specifically under the given set of conditions. Lastly, the probe sequence itself affects specificity during hybridization. More G and C bases in a duplex result in a higher T_m because of the increased number of hydrogen bonds between G and C bases relative to A and T bases. It is for this reason that probe design should aim to be homogenous with regards to the ratio of G and C to A and T bases across probes. Sequence features such as GC runs or mononucleotide repeats in some probes in a massively-parallel hybridization experiment will affect hybridization success of probe-to-target sequences.

The challenges of massively-parallel hybridization

Simultaneous hybridizations of thousands of probes to target DNA means that specificity isn't guaranteed even if there is a perfect match between probe and target DNA sequences. According to the Watson-Crick model of base pairing when there are many correctly paired – despite some mismatched – bases, the thermodynamic penalty of a few mismatches can be overridden to result in imperfectly matched probe to target binding[36]. Having some probes with lowered probe specificity (mismatched binding) is inevitable to some small degree during a high stringency, massively-parallel hybridization experiment. However, probe specificity that is too low leads to hybridization of non-target sample DNA. Errors can be minimized by regulating probe specificity through an accurate temperature calculation, improving how easily a particular sequence can be recognized

by having more probes (probe density and redundancy), and PCR amplifying target sequences[36,50].

Two sources of error can contribute to higher false positive and false negative errors in variant calls from microarray studies: 1) inaccurate probe design and probe annotations, and 2) incomplete protein removal during DNA isolation leaving target DNA inaccessible for hybridization to probes. Bioinformatics methods and software tools make it possible to computationally examine and better understand large sets of biological data to decrease false positive and false negative errors. First, probes not meeting design criteria along with incorrect probe annotations directly influence binding affinity between complementary target and probe sequences and lead to inaccurate variant calls. Second, high-quality, intact target DNA is necessary for successful hybridization. Isolation procedures that do not effectively remove proteins from protein-bound DNA may increase the likelihood for deletions because the DNA was unavailable for hybridization to the array. Inadequate DNA isolation can adversely affect hybridization success and reliability in array data. Computationally examining the two sources of error can reduce false positive and false negative rates known to plague variant calls.

1.6.1 First source of error: Flawed probe design and inaccurate annotations reduce reliability and accuracy of variant calls

Previous studies have determined that SNP probes on the MDGA did not meet design specifications and that inconsistencies in probe design or probe annotations adversely affected genotyping accuracy[2,3]. The physical probes on the array are computationally annotated, that is, a set of information exists for each probe on the array. The information is known as metadata (available for download from the Center for Genome Dynamics at http://cgd.jax.org/datasets/diversityarray/CELfiles.shtml) and includes information such as probe sequence, target sequence, genomic location, and unique identifiers that corresponds to each probe on the array. The metadata are based on the mouse reference genome for which the MDGA probes were designed – build 37 from the National Center for Biotechnology Information (NCBI)[37]. Computational removal and correction of SNP probe annotations prior to SNP genotyping steps were found to increase genotyping

accuracy[2,3]. The results from these studies demonstrate the importance of assessing probe design specifications and associated annotations to increase reliability in variant calls. While MDGA SNP probes have been previously filtered to exclude poorly performing probes from SNP genotyping steps, IGPs on the MDGA have not[2,3].

IGPs on the MDGA were designed to meet a specific set of criteria. Meeting probe design specifications and having accurate probe annotations are important for consistent hybridization success and more reliable CNV calls. Successful hybridization between probe sequences and target DNA are calculated based on multiple factors related to probe design. For example, calculation of the optimal temperature for hybridization to the MDGA is based on probe size of 25 nucleotides; inconsistent probe lengths decrease probe specificity and can result in altered hybridization success[37]. Similarly, reliable CNV calls are only possible if probe annotations are correct i.e., probes annotated to incorrect genomic locations results in erroneous CNV calls.

Bioinformatics tools allow for probe annotation validation. Automating the process of examining large sets of data is possible through the use of programming languages such as Python to carry out specific tasks. In this case, MDGA probes can be computationally assessed based on specific inclusion and exclusion criteria to ensure suitability for hybridization to experimental target DNA, Probes meeting each inclusion and exclusion criterion are referred to as stringent probes and indicate optimal probe design for hybridization to target DNA.

1.6.2 Second source of error: Ineffective removal of proteins from target DNA may result in a bias for deletions in microarray studies

DNA extraction protocols designed to purify DNA of proteins may not be effective across all tissue types[4]. The ever-changing landscape of protein binding across the genome differs depending on the type of tissue, age of the tissue, and the replicative status (i.e., mitosis) of the tissue[5,51–59]. Not taking into consideration levels of protein-binding in different sample types can lead to some DNA samples being protein-bound and blocked from undergoing the DNA preparation steps (restriction enzyme digestion,

adaptor ligation, PCR amplification, and fluorescent tagging) for hybridization to the array. Ultimately, the ability to accurately genotype SNPs and call CNVs is diminished[4].

Chromatin state varies differs between cell and tissue types

Protein-bound DNA refers to DNA that exists in the nucleus in a highly compact structure with the help of proteins. Nuclear DNA is packaged into chromosomes that consist of DNA tightly wound around a small group of proteins called histones that support its structure. This compact DNA-protein complex is called chromatin and exhibits high stability due to the attraction between the negatively-charged DNA and positively-charged histones[60]. The extent to which DNA is associated with histones is directly related to the state in which genomic DNA is packaged within a cell. There are two possible chromatin states within a cell: closed state (silent heterochromatin) and open state (active euchromatin). Heterochromatin is a very tightly packed form of DNA that is not being actively transcribed (inactive) and serves to protect the integrity of DNA. Heterochromatin is involved in gene regulation by silencing genes. Euchromatin is a lightly packed form of chromatin that is typically enriched in genes and is being actively transcribed (active). In this state, chromatin is unwound and temporarily displaced of histones for the DNA to be accessible to polymerases and other enzymes for cellular processes (i.e., transcription and replication).

Chromatin landscapes of open and closed DNA exists in tissue- and cell-type-specific ways to drive cell-specific genic expression over time[55,56,61,62]. DNA methylation in particular plays a major role in tightly regulating gene expression across tissues with some of the highest levels of DNA methylation being found in the mammalian brain[56–58,61]. Global changes in chromatin accessibility to key players in DNA replication also underlie the turning on and off of genes necessary during different stages of development and life. For example, chromatin modifications lead to the cell-type-specific gene expression changes required for initiation of cellular differentiation during embryogenesis. Chromatin accessibility varies drastically during early development and late life within and between tissues and is linked to tissue-specific proliferative potential and patterns of variation[53].

Tissue-specific characteristics can pose a challenge for DNA isolation and hybridization

Chromatin status, homogenous cell types (i.e., liver), and replicative status are known to affect hybridization to create tissue-specific biases in array data[4]. Inaccessible DNA that is highly-protein bound by histones can make it difficult for the complete removal of proteins during DNA purification and can effectively prevent target DNA from hybridizing to the array. A notable example is sperm whereby the majority of sperm chromatin exists in a closed state that can result in higher levels of protein-bound DNA and therefore inaccessible for hybridization[63]. Similarly, cerebellar tissue DNA is also highly protein-bound relative to the spleen and liver and may require more fine-tuning of the extraction protocol to ensure hybridization is possible[4,54,61]. Certain cell types or tissues may require more purification steps prior to hybridization. Brain tissue such as the cerebellum is protein- and lipid-rich – tissue-specific characteristics that can contaminate DNA[64]. The presence of high levels of nuclease that are typical of the spleen also need to be inactivated to ensure that the effects of nucleases in the DNA are reduced. While extraction protocols are meant to take into account the different levels of contamination, chromatin structure and other tissue-specific factors affecting DNA availability is unknown.

Complete digestion of proteins is typically achieved through digestion with the enzyme proteinase K, a broad-range protease that degrades proteins and inactivates nucleases (i.e., DNases) that may degrade DNA during purification[65]. Undamaged, uncontaminated protein-free DNA in mammals is typically isolated through proteinase K digestion and requires two conditions: high enough concentration of the enzyme and long enough incubation time to allow for complete digestion. Due to the differing nature of protein-bound landscapes across the genome from tissue to tissue, a one-for-all incubation time and concentration may not be sufficient for one tissue. For DNA samples from different mouse tissue and cell types, there is no such standard tissue-specific systematic approach. Tissue and cell-type-specific chromatin characteristics and other factors need to be considered when isolating DNA and analyzing CNV.

DNase I sensitivity assays, commonly used to map out chromatin accessibility across the genome, preferentially cleaves DNA at open, accessible sites called DNase I hypersensitive sites (DHSs). Chromatin structural changes occur at active gene loci and at regulatory elements to result in a more open chromatin structure, one that is involved in active transcription and also sensitive to DNase I[59]. DHSs in chromatin are used extensively to map out open and closed DNA regions across many organisms including the mouse[52]. DNase I sensitivity measures can be found as tracks, or downloadable annotation datasets, for various tissues and cell lines in the mouse (curated by the Encyclopedia of DNA Elements (ENCODE) Consortium)[66].

Since it is not possible to assess the effectiveness of DNA preparation protocols in externally sourced data arising from past experiments, bioinformatics tools can be used to assess experimental putative CNV calls. CNV calls can be compared to known tissue-specific landscapes of open and closed DNA. While this would be a tedious task if done manually, software programs and writing in-house code can allow a biologist to work with and examine large sets of genome-wide data. A post-hoc assessment of data arising from a microarray experiment can allow for the identification of any potential inconsistencies in data arising from inaccessible DNA.

1.7 Cross-species application of microarrays allows for the study of genetic variation of non-model organisms

Unsequenced genomes and understudied non-model species are difficult to examine because of a lack of genomic resources (i.e., gene or variant databases) and technologies, and high costs for genome sequencing and annotation. Alternative methods to study non-model genomes are necessary, such as the application of SNP arrays across species for both CNV and SNP study and referred to as cross-species hybridization.

1.7.1 SNP arrays can be useful for cross-species applications

Successful cross-species hybridization for the identification of SNPs is reported in literature, especially for closely-related species (less than 3 million years of divergence)[67,68]. For example, cross-species application of bovine and ovine SNP arrays generated about 2200 polymorphic SNPs in European and American bison and 850 SNPs

in bighorn and thinhorn sheep, respectively[67,69]. Cross-species amplification has been demonstrated to be successful when using medium density or higher SNP arrays (50,000+ markers) in closely-related non-model species. Cross-species hybridization for SNP genotyping can also generate a large number of markers with relatively low cost per locus. However, a major limitation of cross-species amplification is the phylogenetic distance between species. The ability to identify SNPs cross-species diminishes greatly with increased phylogenetic distance, where the proportion of polymorphic SNPs exponentially decreases with phylogenetic distance, dropping to around 5% for species that have diverged 3 million years ago (mya)[70]. This poses a problem for researchers studying distantly-related species for which technologies are lacking.

Cross-species hybridization to phylogenetically divergent taxa has led to varying levels of success. One study generated different percentages of calls and numbers of polymorphic loci across taxa by applying an equine SNP array to wild relatives of the horse such as zebras, asses, tapirs, and rhinoceros[71]. High-density bovine arrays have been successfully applied in more distantly related species such as the antelope (divergence of 24 mya) and deer (divergence of up to 30 million years) to identify about 150 and 1050 polymorphic SNPs of ~54,000 SNP loci present in the bovine genome[72,73]. Estimation of evolutionary relationships among higher ruminants such as deer and giraffes (up to 29 million years of divergence) was also possible by using a bovine SNP array[72,73]. One particular study genotyped 678 higher ruminants representing 61 species and demonstrated success in resolving phylogeny for the diverse group of species based on almost 41,000 genome-wide cross-amplifying SNPs[74]. Furthermore, cross-species hybridization using ancient DNA from an extinct species was also demonstrated[74]. Lastly, cross-species hybridization of the canine array to 24 Antarctic fur seal individuals (*Arctocephalus gazella*; 44 million years of divergence) generated 173 SNPs (0.5% of canine SNP loci) that harbor highly conserved genomic regions[75].

The utility of cross-species hybridization for the study of CNVs has been demonstrated across various closely related species: bovine-goat, bovine-sheep, chicken-turkey, chicken-duck, and human-primates[18,76–78]. Cross-species hybridization allowed for the

first studies of CNV in primates and was routinely carried out over the last decade. Specifically, microarrays designed for the human genome were used to identify the first putative sites of CNV in great apes such as the chimpanzee, gorilla, bonobo, orangutan, and rhesus macaque[79–83].

1.7.2 The naked mole-rat as a candidate for cross-species hybridization of the Mouse Diversity Genotyping Array

The naked mole-rat is a valuable non-model organism to study for several unique characteristics in biology, genetics, and population structure. Though the naked mole-rat is similar in size to the mouse (27-30 grams), it is the longest lived rodent with a lifespan of 32 years compared to just three to five years for the mouse[84]. The naked mole-rat lives five times longer than predicted based on body size, similar to that of a human[85]. The naked mole-rat is the first reported mammal that does not undergo typical mammalian aging and susceptibility to disease because good health is maintained for at least 66% of its life (the equivalent of an 80 year old human showing a biological age of 30)[86,87]. The naked mole-rat also exhibits resistance to tumors, in stark contrast to the majority of laboratory mice (C57BL/6J) that die of cancer and show lesions and small non-lethal tumors[87–89]. The subterranean burrowing rodent is native to the eastern horn of Africa and exhibits unique adaptations to its dark, hypoxic environment rich in carbon monoxide and ammonia. The naked mole-rat shows extremely high levels of inbreeding and is also one of only two known eusocial mammals. Eusocial mammals, like bees and other social insects, live in colonies of 75-300 individuals. Each colony consists of a single breeding female and one to three breeding males with the remaining members being sterile workers that carry out communal tasks such as food collection and tunnel excavation. Once the queen dies, a sterile female becomes the queen by losing her sterility[90].

Much of the value in studying the naked mole-rat genome comes from its comparison to genomes of other mammals, particularly the mouse and human genomes. The naked mole-rat lineage diverged from the mouse and humans lineages about 70 mya and 90 mya, respectively[84]. Although limited genomic annotations, resources, and technologies exist for the naked mole-rat, there are currently two fully sequenced reference genomes –

one male and one female naked mole-rat. Initial genomic analyses suggest a low mutation rate due to a reduced level of polymorphisms found in the naked mole-rat[84,91]. Furthermore, 93% of the naked mole-rat genome shows synteny to human, mouse, or rat genomes and thus, allows for this species of interest to be studied alongside classical laboratory mouse and human in a comparative manner[84]. A recent study suggests that 88% of human genes (about 17,000 genes) have a naked mole-rat ortholog[92]. The naked mole-rat is predicted to have 22,561 genes – comparable to other mammals such as the human (22,389), mouse (23,317), and rat (22,841) genomes[5,21,84].

1.7.3 Probes on the array may be suitable for cross-species hybridization

The cross-species application of arrays necessitates the consideration of whether the array probes themselves are suitable for hybridization to DNA from a species of interest. While cross-species hybridization can be carried out for unsequenced species (i.e., canine array for seals) or for species that are already sequenced (human array for chimpanzees), it is unknown whether the MDGA has practical utility for cross-species hybridization in a previously sequenced distant species – the naked mole-rat. Predicting the potential for probes on the MDGA to interrogate naked mole-rat samples can be carried out by aligning the probe sequences to similar sequences in the naked mole-rat reference genomes. There are currently two fully sequenced reference genomes – one male (build 1) and one female (build 2) – of the naked mole-rat (via whole-genome shotgun assembly). Since sequencing methods cannot read the entire genome at one time, small sequence fragments are read one at a time and then assembled and linked together into a scaffold (with occasional gaps). Not all genomes are created equally; the naked mole-rat assemblies are much lower in quality than the mouse genome which is expected since the scaffolds are yet to be mapped to physical locations in the two genomes. Scaffolds are associated with chromosomes but exact positional information by chromosome has not been determined.

Cross-species application of the SNP array across divergent sets of taxa is challenging for two reasons: firstly, polymorphic SNPs decrease exponentially with an increase in phylogenetic distance before leveling off after about 5 million years of divergence and

secondly, a lack of appropriate CNV probes make cross-species hybridization difficult[70]. SNP probes on the MDGA are variable in nature (sites of polymorphism) and are less likely to be conserved cross-species. IGPs on the other hand are expected to be conserved due to the targeting of almost all known invariant exons in the mouse though not for CNV calling.

CNV detection in the naked mole-rat genome is unlikely due to the lack of appropriate IGPs. However, the MDGA by virtue of its IGP design and high number of targeted loci, may be useful cross-species hybridization. The IGPs are designed such that highly conserved exons of genes (invariant genomic regions) can be interrogated and a subset of the IGPs are known to cover ultraconserved regions. The probes are also designed to capture maximum diversity present in mice from the C57BL/6J to wild-caught mice and therefore diverse genomes of the mouse are represented (reduces bias in probe sequences). Furthermore, because the IGPs cover almost all of the exons in the mouse genome (about 20,000 genes) and represent varying levels of evolutionary conservation, the IGPs have unbiased coverage of genes, some of which may exist in phylogenetically distant organisms such as the naked mole-rat.

While polymorphic SNPs decrease with divergence, successful cross-species application of a SNP array to identify polymorphic SNPs may be possible given a large number of loci on the array and a large enough sample size of the species under study. The MDGA is a high-density SNP array targeting over 600,000 SNP loci across the entirety of the mouse genome, potentially allowing for the generation of even a small number of SNP markers. Though cross-species hybridization may not be as successful across divergent taxa based on theoretical expectations, discovery of even a limited number of SNP markers is valuable for the generation of useful markers in unstudied genomes of naked mole-rat colonies. Cross-species application of the MDGA may be valuable for comparative genomic studies by allowing for the discovery of polymorphic loci, homologues, evolutionarily and biologically significant regions of the genomes, as well as the study of the genetic diversity of naked mole-rat colonies.

1.8 Central goal and specific objectives

Central goal: The massively-parallel hybridization of Mouse Diversity Genotyping Array probes to target DNA is error-prone. Bioinformatics methods will be used to validate whether probes on the array are well designed such that high hybridization success to sample DNA is expected, that sample DNA is accessible for DNA to probe hybridization, and that the Mouse Diversity Genotyping Array is conducive to cross-species hybridization.

Objective 1) To computationally assess Mouse Diversity Genotyping Array probe annotations based on probe design criteria such that only appropriately designed probes with accurate probe annotations are used for reliable SNP genotyping and CNV calling.

Objective 2) To conduct a post-hoc examination of the association between detected CNVs and regions of known closed DNA such that deletions preferentially located within closed regions are predicted to be protein bound and unavailable for hybridization.

Objective 3) To predict the theoretical ability of the Mouse Diversity Genotyping Array probes to hybridize cross-species to naked mole-rat target DNA through single-locus complementarity of probe sequences to two naked mole-rat reference genomes.

Chapter 2 Materials and Methods

2.1 Assessment of probe suitability and annotation accuracy for hybridization to target mouse DNA

Original IGP annotation files were downloaded from the Center for Genome Dynamics website (http://cgd.jax.org/datasets/diversityarray/CELfiles.shtml). A previously filtered list of SNP probe annotations was obtained for further validation[2,3]. A total of 1,195,516 IGP annotations targeting 597,758 unique exonic loci and 523,322 SNP probe annotations for each unique SNP locus were compiled to Microsoft Excel. Computational assessment of probe annotations based on probe design criteria was carried out using in-house script written in Python (programming language) coupled with a local blasting program called BLAST+ and build 37 of the mouse genome database (UCSC:mm9). BLAST+ allows for the comparison of biological sequences and provides data on regions of similarity between the sequences. Specifically, an algorithm called Basic Local Alignment Search Tool (BLAST) was used to compare a query sequence to the sequences found within a database of sequences (i.e., mouse reference genome). BLAST results identify sequences within the database that share sequence identity with the query sequence above a certain statistical threshold. The BLAST+ executable program was downloaded from the National Centre for Biotechnology Information (NCBI) website (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download). The mouse genome database (build 37) was downloaded in FASTA format, a commonly used text-based format containing sequence data (nucleotides) available at https://hgdownload.cse.ucsc.edu/goldenPath/mm9/chromosomes/.

IGPs were determined to be suitable for subsequent analyses if design specifications detailed by Yang *et al.* were met[37]. IGP annotations were assessed based on four of the design specifications: 1) IGP sequences are murine 2) IGP sequences are unique in the genome (non-repetitive) 3) IGPs have correctly annotated base pair start and end positions and chromosomal location, and 4) IGPs are 25 nucleotides in length. Confirmation for each SNP probe sequence annotation containing the correct SNP allele at the correct SNP locus for each SNP probe had not been previously carried out. Two

criteria were computationally confirmed: 1) the SNP allele being interrogated by each probe must be consistent with the indicated SNP allele, and 2) the SNP allele must also be positioned at the correct base pair location in the probe sequence as indicated in the annotation file by Yang *et al.*[37].

The in-house script was written such that each line of the probe list was sorted if specified criteria were met. Each probe annotation that did not meet the criteria was separated into a separate file. Probe annotations that did meet criteria were then compiled into a secondary list for local BLAST. A second script was written to call for the function "blastn" from the BLAST+ executable program that utilizes the NCBI C++ Toolkit. Local BLAST was performed against downloaded mouse genome database (build 37). A Windows command line function called "makeblastdb" was called to generate a local database of the mouse genome that the "blastn" function could access. The script read the annotated probe list in as input, the "blastn" function carried out alignment of each probe sequence against the mouse genome, and the probe annotations meeting specified criteria were recorded to an excel file as output.

All IGP sequence annotations that met the four specified design criteria were deemed as stringent probes, verified to be suitable for hybridization success between probe and target DNA [37]. Probes, for which annotations could be corrected, were manually corrected on Excel and added to the list of verified probes. Specifically, the mismatched start and end positions of 242 IGPs on the array were replaced by positions in NCBI's build 37 of the mouse genome (UCSC:mm9). Probe annotations that could not be corrected while also not meeting design criteria were computationally removed from the probe list used for SNP genotyping and CNV calling. The SNP probe annotations that were found to be incorrect were computationally removed from the previously filtered SNP probe list[3]. SNP annotations having the correct SNP allele at the correct SNP locus were deemed suitable for SNP detection as outlined by Yang *et al.*[37].

2.2 Mapping copy number variant regions alongside DNase I sensitive regions across the genome

CNV calls derived from four different tissues across two sets of samples were obtained for analysis: a reference set (Jackson Laboratory) and an in-house experimental set (Hill Laboratory)[1,93,94]. CNV calls across both the reference and in-house experimental sets were determined based on well performing, stringent probes that meet design specifications as outlined by Yang *et al.* and target 496,900 SNP loci and 435,167 unique exonic regions[1,37,94]. CNV data from the mitochondria and chromosome Y were excluded because of a relatively low number of probes that exist to target sequences on Mus chromosomes Y and MT (mitochondrial DNA). CNV data from chromosome X and Y were not available at the time of study.

The reference sample set published by Locke *et al.* consists of putative CNVs found in 351 adult Mus tail samples (Table 2.1)[1]. The CNVs in this study were determined from publically available Mouse Diversity Genotyping Array CEL files from the Center for Genome Dynamics at the Jackson Laboratory (http://cgd.jax.org/datasets/diversityarray/CELfiles.shtml). The 351 CEL files contain raw array intensity data for samples from 120 classical laboratory strains, 58 wild-derived strains, 10 consomic strains, 1 congenic strain, 44 BXD recombinant inbred strains, 40 CC-UNC, G2:F1 strains, 55 F1 hybrids and 23 wild-caught mice. The CNV calls generated from the Jackson sample set represent the most extensive CNV (specifically, germline CNV) analysis to date of mouse tail samples with 331 of 351 samples meeting quality control standards[1]. A total of 9,634 putative autosomal CNVs were called, of which 5656 were deletions and 3978 were duplications (a ratio of 1.42:1). The autosomal CNVs represent 6.87% of the mouse reference genome.

Table 2.1 Copy number variant (CNV) data from a publically available reference sample set of 351 mouse tail samples[1]

Mouse strain[2]	Number of samples	Total number of CNVs	Total number of CNVs by state[3]				CNV loss/gain ratio[4]
			0	1	3	4	
Classical laboratory strain	120	2824	424	867	887	646	0.84
Congenic	1	12	0	6	3	3	1
Consomic	10	296	8	192	53	43	2.08
BXD	44	680	67	364	149	100	1.73
Wild-derived laboratory strains	58	2611	1214	594	361	442	2.25
F1 hybrid	55	1370	35	707	422	206	1.18
CC-UNC G2:F1	40	872	16	440	280	136	1.1
Wild caught	23	969	231	491	109	138	2.92
Total	**351**	**9634**	**1995**	**3661**	**2264**	**1714**	**1.42**

[1]Copy number variant calls were published by Locke *et al*. and discovered using publically available Mouse Diversity Genotyping Array data[1]. Mouse Diversity Genotyping Array CEL files were downloaded from the Center for Genome Dynamics at the Jackson Laboratory (http://cgd.jax.org/datasets/diversityarray/CELfiles.shtml).

[2]Mouse strains describe the genetic background of 351 mouse tail samples previously hybridized by the Jackson Laboratory to the Mouse Diversity Genotyping Array.

[3]CNVs are classified by state as either deletions (copy number state of 0 or 1) or duplications (copy number state of 3 or 4).

[4]CNV deletions and duplications are referred to as losses and gains, respectively.

The Hill sample set consists of CNV calls from 12 cerebellum, 5 liver, and 10 spleen tissue samples (Table 2.2)[3,94]. The samples include replicates of the same tissue from the same mouse, as well as multiple tissue types from the same mouse and were obtained for the study of somatic CNVs. All mouse tissue samples were derived from adult mice ranging in age from 4.4 months to 11.4 months with either CBA/CaJ (8 mice) or C57BL/6J (2 mice) mouse strain background. A total of 363 putative autosomal CNVs were called with a deletion to duplication ratio of 1.19:1. Total number of CNVs by tissue type are as follows: 184 (cerebellum), 31 (liver), 148 (spleen) with loss to gain ratios of 1.11, 1.82, and 1.18 for each tissue respectively.

To determine probable open and closed areas of DNA across the genome, annotations for DNase I hypersensitive regions across four tissues were obtained (Table 2.3). The genomic intervals for DNase I sensitivity signals across the mesoderm, cerebellum, liver, and spleen were found as annotation tracks on UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=mm9&g=wgEncodeUwDNase). DNase I sensitivity measures for the cerebellum, liver, and spleen were derived from adult C57BL/6J mice at 8 weeks of age while measures for the mesoderm tissue were derived from a mice of CD-1 background at embryonic day 11.5[51,66]. The cerebellum, liver, mesoderm, and spleen tissues were chosen for their comparability to the cerebellum, liver, tail, and spleen tissues used for CNV detection in previous studies.

Original signal annotation tracks were available in BigWig format, an indexed binary file format that allows for the display of dense, continuous data that can be displayed from the Genome Browser as a graph. The tracks are annotation files containing measures (signals) of DNase I sensitivity. Signal tracks in BigWig format were simplified to wiggle (wig) format using a conversion program called bigWigToWig from UCSC available from the directory of binary utilities at http://hgdownload.cse.ucsc.edu/admin/exe/. The wig files were then converted to annotation tracks in BED format using an open sourced Linux-based genomic toolset called BEDOPS (v2.4.14)[95]. An annotation track in BED format must consist of a minimum of three required fields: chromosome (chrom), start

position (chromStart), and end position (chromEnd). The resulting BED files were then used as data

Table 2.2 Copy number variant (CNV) data for an in-house sample set derived from three mouse tissues[1]

Mouse tissue[2]	Number of samples	Total number of CNVs	Total number of CNVs by state[3]				CNV loss/gain ratio[4]
			0	1	3	4	
Cerebellum	12	184	28	69	53	34	1.11
Spleen	10	148	18	62	50	18	1.18
Liver	5	31	10	10	8	3	1.82
Total	**27**	**363**	**56**	**141**	**111**	**55**	**1.19**

[1]CNV calls were obtained from the Hill Laboratory (Milojevic, unpublished).

[2]Mouse tissue samples include replicates of the same tissue from the same mouse as well as multiple tissue types from the same mouse. Tissue samples were taken from adult mice ranging in age from 4.4 months to 11.4 months with a genetic background of either CBA/CaJ (8 mice) or C57BL/6J (2 mice).

[3]CNVs are classified by state as either deletions (copy number state of 0 or 1) or duplications (copy number state of 3 or 4).

[4]CNV deletions and duplications are referred to as losses and gains, respectively.

Table 2.3 Previously published DNase I accessibility data for mouse tissues exist as tracks

Tissue type	Mouse strain	Age	Track type[1]	Track Name[2]
Cerebellum	C57BL/6J	Adult 8 Weeks	Signal[3]	Cerebellum DNaseI HS Signal Rep 1
Liver	C57BL/6J	Adult 8 Weeks	Signal	Liver C57BL/6 Adult 8 Weeks DNaseI HS Signal Rep 1
Mesoderm	CD-1	Embryonic day 11.5	Signal	Mesoderm DNaseI HS Signal Rep 1
Spleen	C57BL/6J	Adult 8 Weeks	Signal	Spleen DNaseI HS Signal Rep 1

[1]DNase I sensitivity signal annotation tracks curated by the ENCODE Consortium were downloaded from the University of California, Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=mm9&g=wgEncodeUwDNase)[66].

[2]Tracks exist as three types of sequence data: Hotspots, Peaks, and Signals.

[3]Signals are defined as the density of tags mapping within a 150 base pair sliding window[66].

containing DNase I accessible, or open regions across the genomes of four tissues. Custom annotation tracks representing genomic intervals of CNV regions were created for each tissue type. To determine the degree of overlap between DNase I sensitive and CNV regions of the genome by tissue type, an open-source Linux-based genomic arithmetic software called BEDTools (v2.18) was used[96].

To generate BED files of genomic intervals of DNase I inaccessible, or closed DNA, the inverse of the DNase I accessible intervals was found using the "complement" function on BEDTools. All complements were based on chromosomal start and end positions from NCBI's build 37 of *Mus musculus* (UCSC:mm9). The complemented BED files representing genomic intervals of closed DNA for each tissue type were compared with the genomic intervals of the CNV regions from each tissue. The BEDTools "intersect" function was used to enumerate each and every intersection between the two sets of genomic intervals to determine degree of overlap between the CNVs and closed DNA. Only the autosomal genome (chromosomes 1-19) was considered for the analyses. Output files containing intersection data between genomic intervals of closed DNA and CNV regions by tissue type were obtained and analyzed for degree of overlap.

DNase I signal annotation tracks were further examined to discover any tissue-specific patterns of open and closed regions across the four tissues. Tissue-specific open or closed regions of DNA are segments of DNA along the chromosomes that are not found as open or closed in the other tissues, that is, the intervals of open and closed DNA are unique to the tissue. This was accomplished by computationally comparing each tissue track to the conglomerate of the other three tissue tracks. This was achieved by compiling all genomic intervals, sorting the intervals based on start and end position by chromosome, and merging intervals that overlap. The conglomerate file representing shared regions between the three tissues was intersected with the tissue track of interest to determine shared regions of DNA across all four tissues. This track represents the shared regions of all four tissues, that is the union all four tracks. To determine unique patterns of open and closed DNA for the tissue of interest, the intervals from the tissue of interest were

subtracted from the union of all tissues. The process was repeated for each tissue of interest to determine tissue-specific closed and open DNA.

2.4 Compiling a list of stringent probes suitable for cross-species hybridization

Probes predicted to be suitable for cross-species hybridization were determined by BLAST, or aligning the probe sequences to similar sequences in two reference naked mole-rat genomes (Table 2.4). All probe sequences were BLAST searched against each existing naked mole-rat genome build (1 and 2) that was downloaded from the NCBI FTP server (ftp://ftp.ncbi.nlm.nih.gov/genomes/Heterocephalus_glaber/). The script was written such that BLAST results of probe sequences complementing without mismatch to one unique locus in the naked mole-rat were recorded to one output file and probe sequences complementary to more than one locus were recorded to a second output file. All other BLAST results were excluded.

A probe was considered to be complementary to the naked mole-rat genome if all 25 nucleotides of the probe sequence aligned perfectly with complementarity (with no mismatch in nucleotides) to a sequence in the naked mole-rat genome. Any probe sequences found to align to more than one locus in one genome build would result in competition for probe-DNA hybridization and were thus excluded from a stringent list of probes targeting unique loci only. Probes found to be complementary without mismatch to sequences in both genome builds were compiled as a list of probes that theoretically should be able to reliably hybridize to naked mole-rat genomes (in the absence of genetic variation or *de novo* mutations) with high hybridization success.

Table 2.4 Two naked mole-rat reference genome builds were used to predict hybridization between Mouse Diversity Genotyping Array (MDGA) probes and naked mole-rat samples

Genome Assembly[1]	Build 1[2]	Build 2[3]
Name of genome build	HetGla_1.0, USCS name: BGI HetGla_1.0	HetGla_female_1.0, UCSC name: hetGla2
Submitter	Beijing Genomics Institute	Broad Institute
Scaffolds	39267	4229
Scaffold N50	1.6 Mb	20 Mb
Number of contigs	273990	114653
Contig N50	21750	47778
Predicted genes	30743	30876
Predicted proteins	41963	34892
Size (Mb)	2643.96	2618.2

[1]Naked mole-rat genome assemblies (build 1 and build 2) were downloaded from the National Centre for Biotechnology Informatics (NCBI) FTP server: ftp://ftp.ncbi.nlm.nih.gov/genomes/Heterocephalus_glaber/.

[2]Build 1 was released by Kim *et al.*[84].

[3]Build 2 was released by Keane *et al.*[97].

2.5 Determining sequence-specific biological functions of probe targets complementary to the naked mole-rat genome

Since no comprehensive database of known exons or genes exists for the naked mole-rat and since there is potential for exonic regions in the mouse to be conserved in the naked mole-rat, all probe sequences complementary to the naked mole-rat genome were examined for functional attributes using the existing annotated mouse genome. Annotations for mouse genes and exons based on NCBI's build 37 of the mouse genome were downloaded from the archives (release 67) of an open-source genomic database called Ensembl BioMart (http://www.ensembl.org/info/website/archives/index.html). Probes on the array were designed to interrogate these genes and exons annotated by the Ensembl group.

Annotations for the stringent probe list and annotations for the gene and exon lists were compiled to generate BED files for use on BEDTools. All probe annotations were assessed for overlap with the annotations of mouse genes and exons on their respective chromosomes using the "intersect" function. All probe sequences were also mapped along NCBI's build 37 of the mouse genome (UCSC:mm9) to ascertain their distances in nucleotides to proximal genes, upstream or downstream, using a function called "closest". Output from BEDTools was converted to Excel format for examination of the overlapping and proximal genes and exons.

Mus genes overlapping with and proximal to IGP and SNP probe sequences found complementary to the naked mole-rat genomes were analyzed on various platforms for biological attributes. The Database for Annotation, Visualization and Discovery (DAVID) toolset was used for grouping overlapping genes into biological clusters based on functional classification and annotations [National Cancer Institute at Frederick, Frederick, MD][98]. The Functional Annotation tool was used to identify gene ontology (GO) term enrichment for the genes. Three default GO categories (GOTERM_BP_FAT, GOTERM_CC_FAT) were used to identify the most relevant GO terms for each set of genes overlapping IGPs and SNP probes.

Diseases and biological functions that were overrepresented in genes of interest were determined. Specifically, genes in each list were grouped into disease and biological function networks using QIAGEN's Ingenuity Pathway Analysis' Core Analysis [IPA®, QIAGEN Redwood City, CA]. Focus Genes (or Focus Molecules) are those genes from the two gene lists that pass filters and have potential to be linked to other genes as part of a gene network. Direct and indirect relationships with a maximum of 35 Focus Molecules per network were included. Pseudogenes were also included. Molecule relationships with endogenous chemicals were excluded. Networks included for analysis were those networks that have a score of two or higher with a score of two reflecting a 99% confidence of not being generated by random chance alone. This score is based on a p-value of 0.05 from a Fisher's Exact Test.

The Ensembl genes that are targeted by the MDGA probes can be associated with variation. The predicted effect of the variants is listed as a SNP class function (i.e., missense change). All genes found to overlap with SNP probes and predicted to have nonsynonymous effects were manually examined using Ensembl's gene search function and researching the Mouse Genome Informatics (MGI) international database (accessible at http://www.informatics.jax.org/).

2.6 Analysis of variance was used to determine differences in DNase I accessibility between tissues

Analysis of Variance (ANOVA) was used to statistically compare the percentages of closed DNA intervals by tissue type – cerebellum, liver, spleen, and tail – and by chromosome. Specifically, a single factor ANOVA was used to determine whether any tissue or chromosome showed a significantly higher or lower percentage of closed DNA. The single factor ANOVA was carried out on a program called Statistics Package for the Social Sciences (SPSS). The significance level was set to 0.05 which is typical for most studies in the field. The *F* Test value outputted by the single factor ANOVA shows whether or not significant differences exist between group means.

Chapter 3 Results

3.1 A number of probe annotations required computational correction or removal prior to SNP genotyping and CNV calling

From the original IGP annotation list, a total of 316 IGP annotations were computationally removed (from further analysis steps) and annotations for 242 IGPs were corrected (Table 3.1). A total of 268 probes that were found to align to a minimum of two loci in the mouse genome were deemed as duplicate probes and discarded. Another 48 probe sequences were not found in the mouse genome at all, or had annotations that could not be manually corrected. The unverified probe sequences were unable to be mapped to the mouse reference genome and therefore discarded since hybridization between probe sequence and target DNA in the mouse is unlikely. All probes were confirmed to be the 25 nucleotides in length that was specified for DNA to probe hybridization.

From the previously filtered SNP probe list, a total of 2088 SNP probes (261 probe sets) were further excluded due to inconsistencies in their annotations for SNP allele or SNP allele position (Table 3.1)[1,3]. A stringent probe list containing validated SNP probes targets 492,952 unique SNP loci, down from the original 623,124 unique SNP loci[1–3]. Further filtering of the probes, as described by Locke *et al*. was extensive and reduced the original list of IGPs targeting 597,758 loci to a final filtered list of stringent probes that target 435,167 unique exonic loci[1]. Collectively, the stringent probe list assays 90.6% of all protein-coding Mus genes and 68% of all Mus genes (Table 3.2). Stringent IGPs in particular are able to query 89% of all Mus protein-coding genes and 63% of all Mus genes. Stringent SNP probes assay 68% of all protein-coding Mus genes and 49% of all Mus genes.

Table 3.1 Annotations for invariant genomic probes (IGPs) and single nucleotide polymorphism (SNP) probes used for analyses were either corrected or removed based on exclusion criteria

Probe	Exclusion criteria for	Number of	Annotation	Number of

type	stringent probe list	probe annotations affected	filtering step taken	probe sets (unique loci) affected
IGP	Probe sequence was not in correct chromosomal position[1]	242	Correction	40
IGP	Probe sequence was not unique in the mouse genome[1]	268	Removal	48
IGP	Probe did not have complete annotation for CNV detection[2]	48	Removal	8
IGP	Probe sequence was not 25 nucleotides in length	0	Removal	0
SNP	Probe sequence did not contain SNP allele at correct location	2088	Removal	261

[1]Probe sequence was not complementary without mismatch to a single location in the mouse reference genome build 37 or probe sequence was mapped to the incorrect genomic position (National Centre for Biotechnology Informatics (NCBI), mm9:Ensembl).

[2]Probe sequence could not be run through BLAST and/or could not be mapped to the mouse reference genome build 37 (National Centre for Biotechnology Informatics (NCBI), mm9:Ensembl).

Table 3.2 Proportion of all Mus genes associated with stringent probes on the Mouse Diversity Genotyping Array

	Gene descriptor[1]	Number of Mus genes[2]	Number of Mus genes targeted by all probes	Percent Mus genes targeted by all probes[3]	Number of Mus genes targeted by IGPs	Percent Mus genes targeted by IGPs	Number of Mus genes targeted by SNP probes	Percent Mus genes targeted by SNP probes
Gene Status	Known[4]	30416	23069	76%	21666	71%	16771	55%
	Novel[5]	6647	2447	37%	1840	28%	1350	20%
	Putative[6]	928	477	51%	307	33%	314	34%
	Total	**37991**	**25993**	**68%**	**23813**	**63%**	**18435**	**49%**
Gene type	Protein coding	22707	20579	91%	20187	89%	15438	68%
	Pseudogene	5474	1222	22%	579	11%	733	13%
	lincRNA	2057	1496	73%	780	38%	1241	60%
	miRNA	1639	304	19%	299	18%	6	<1%
	snoRNA	1560	517	33%	486	31%	35	2%
	snRNA	1429	252	18%	249	17%	4	<1%
	Antisense	1381	1024	74%	746	54%	726	53%
	miscRNA	491	53	11%	36	7%	17	4%

	Gene descriptor[1]	Number of Mus genes[2]	Number of Mus genes targeted by all probes	Percent Mus genes targeted by all probes[3]	Number of Mus genes targeted by IGPs	Percent Mus genes targeted by IGPs	Number of Mus genes targeted by SNP probes	Percent Mus genes targeted by SNP probes
Gene type	IGV	355	177	50%	173	49%	11	3%
	rRNA	338	72	21%	68	20%	6	2%
	Processed transcript	299	225	75%	168	57%	171	57%
	IGJ gene	88	2	2%	2	2%	0	<1%
	Sense intronic	78	31	40%	5	6%	27	35%
	IGD gene	25	0	<1%	0	<1%	0	<1%
	MT tRNA	22	0	<1%	0	<1%	0	<1%
	IGC gene	13	12	92%	12	92%	5	39%
	Non-coding	12	12	100%	10	83%	7	58%
	Polymorphic pseudogene	8	6	75%	6	75%	2	25%
	Sense overlapping	8	3	38%	2	25%	3	38%
	3 prime overlapping ncrna	3	3	10%	3	100%	2	67%
	ncRNA host	2	2	100%	2	100%	0	<1%

	Gene descriptor[1]	Number of Mus genes[2]	Number of Mus genes targeted by all probes	Percent Mus genes targeted by all probes[3]	Number of Mus genes targeted by IGPs	Percent Mus genes targeted by IGPs	Number of Mus genes targeted by SNP probes	Percent Mus genes targeted by SNP probes
Gene type	MT rRNA	2	1	50%	0	<1%	1	50%
	Total	**37991**	**25993**	**68%**	**23813**	**63%**	**18435**	**49%**

[1]A complete list of Mus genes found on build 37 of the mouse genome (mm9: Enseml 67) was downloaded from archives on Biomart's Ensembl: http://www.ensembl.org/info/website/archives/index.html. Gene types include: lincRNA (long intergenic non-coding RNA), miRNA (microRNA), snoRNA (small nucleolar RNA), snRNA (small nuclear RNA), miscRNA (micscellaneous RNA), IGV (immunoglobulin variable gene), rRNA (ribosomal RNA), IGJ (immunoglobulin J gene), IGD (immunoglobulin D gene), MT tRNA (mitochondrial transfer RNA), IGC (immunoglobulin constant gene), ncRNA (non-coding RNA), and MT rRNA (mitochondrial ribosomal RNA).

[2]Counts are based on the complete set of unique Ensembl Gene IDs representing all genes in the mouse genome (build 37).

[3]All stringent probes were mapped to genes on build 37 of the mouse genome (mm9: Enseml 67). Stringent probes are Mouse Diversity Genotyping Array probes that met all inclusion design criteria[1].

[4]Known genes have an official gene name, symbol, and function.

[5]Genes classed as novel are those protein coding genes that do not have an available official gene name and symbol.

[6]Putative refers to a segment of DNA that is believed to be a gene based on its open reading frame; however gene function is

unknown.

3.2 Copy number deletions did not show a bias for complete overlap with closed DNA and duplications were detected in regions of closed DNA

3.2.1 DNase I accessibility measures across tissues show similar ratios of open to closed DNA

The highest percentage of open DNA was found to be in the embryonic mesoderm tissue with an open to closed ratio of 1.85 followed by adult spleen (1.63), liver (1.30), and cerebellum (1.43) (Table 3.3). Assessing total open and closed regions of DNA by chromosome revealed specific patterns (Table 3.4). Chromosome 19 contained the lowest percentage of closed DNA and chromosomes 1 and 7 contained the highest percentage of closed DNA relative to all chromosomes across all tissues. Chromosome 1 had the highest percentage of closed DNA, particularly in the mesoderm and spleen tissues. And chromosome 7 had the second highest amount of closed DNA with spleen showing the highest relative percentage of closed DNA. Percentages of total closed DNA by chromosome increased in a similar and consistent pattern from chromosome 1 to 7 and decreased thereafter with chromosome 19 having the lowest percentage of closed DNA. Despite these relative differences in percentage of closed DNA by chromosome and by tissue type, no significant differences between the tissues were found (p-value of 0.09 > 0.05 based on a single factor ANOVA).

Table 3.3 Percentage of total DNase I accessibility and inaccessibility across the autosomal genome by tissue type[1]

Mouse tissue[2]	Age	Percent open	Percent closed	Open/Closed ratio[3]
Mesoderm	Embryonic day 11.5	65[±]	35	1.85
Cerebellum	Adult 8 weeks	59	41	1.43
Liver	Adult 8 weeks	57	43	1.3
Spleen	Adult 8 weeks	62	38	1.63

[1]DNase I sensitivity signal annotation tracks were downloaded from the University of California, Santa Cruz (UCSC) Genome Browser: http://genome.ucsc.edu/cgi-bin/hgFileUi?db=mm9&g=wgEncodeUwDNase[66].

[2]DNase I sensitivity measures for mouse cerebellum, liver, and spleen were chosen for their comparability to Hill sample tissues used for CNV reference detection. Mesoderm was selected as the closest proxy to mouse tail samples used for CNV detection.

[3]Open to closed ratio is the total percentage of open, accessible DNA divided by the total percentage of closed, inaccessible DNA across the genome.

[±]A darker shade of green or red indicates a relatively higher percentage of open or closed DNA, respectively.

Table 3.4 Distribution of all closed regions as a percentage of all base pairs across the autosomal genome[1]

Chromosome	Mesoderm[2] Embryonic day 11.5	Cerebellum[2] Adult 8 weeks	Liver[2] Adult 8 weeks	Spleen[2] Adult 8 weeks
1	2.89[±]	3.19[±]	3.41	3.08
2	2.4	2.85	3.06	2.65
3	2.4	2.59	2.74	2.57
4	2.23	2.61	2.76	2.42
5	2.1	2.58	2.71	2.32
6	2.14	2.41	2.56	2.29
7	2.5	3.12	3.04	2.64
8	1.86	2.17	2.32	2.06
9	1.55	2.05	2.1	1.74
10	1.79	2.02	2.21	1.94
11	1.38	1.95	1.99	1.53
12	1.87	2.09	2.24	1.99
13	1.74	2.02	2.15	1.85
14	2.05	2.18	2.34	2.16
15	1.41	1.62	1.79	1.53
16	1.41	1.58	1.7	1.53
17	1.35	1.68	1.7	1.44
18	1.25	1.44	1.58	1.38
19	0.8	1.05	1.08	0.87
Total genome	**0.35**	**0.41**	**0.43**	**0.38**

[1]DNase I sensitivity signal annotation tracks were downloaded from the University of California, Santa Cruz (UCSC) Genome Browser: http://genome.ucsc.edu/cgi-bin/hgFileUi?db=mm9&g=wgEncodeUwDNase[66].

[2]DNase I sensitivity measures for mouse cerebellum, liver, and spleen were chosen for their comparability to Hill sample tissues used for CNV detection. Mesoderm was selected as the closest proxy to mouse tail samples used for CNV detection.

[±]A darker shade of red indicates a relatively higher percentage of closed DNA.

3.2.2 A disproportionate number of deletions in closed DNA was not observed

CNV calls from the Locke and Hill sample sets had revealed more total deletions than duplications (Tables 2.1 and 2.2). There were also higher counts of deletions partially overlapping with reported regions of closed DNA though very few deletions were completely contained in closed DNA. The number of deletions and duplications as a proportion of total deletions and duplications show similar degrees of overlap with closed DNA (ranging from 1% to 99% base pair coverage of each CNV). Almost all CNVs from the Locke sample set (99%) and all CNVs from the Hill sample set (100%) were overlapping with closed DNA to some degree however most of the CNVs show minimal overlap with closed DNA (less than 50% base pair overlap). Deletions are more frequently found in smaller base pair overlaps with closed DNA (less than 50% base pair overlap) (Table 3.5). State 0 CNVs overlap less with closed DNA compared to CNVs of state 1. Similarly, duplications more frequently overlap closed DNA with smaller base pair overlaps (less than 50% base pair overlap) (Table 3.6). More duplications (state 3 CNVs and state 4 CNVs) overlap minimally with closed DNA than all deletions.

About 21-24% percent of all tail, cerebellum, and liver CNVs and 8% of spleen CNVs were overlapping closed DNA with more than 50% base pair coverage (all deletions and duplications). Across all duplications and deletions, average percent overlap with closed DNA was similar and consistent across all tissues. Between the two sample sets, Hill CNVs show a higher base pair overlap with closed DNA than Lock CNVs. When examining all Hill sample set CNVs, deletions and duplications show no partial overlap with closed DNA above 75% base pair coverage and any overlaps above 75% occur only as complete coverage with closed DNA (100% base pair coverage). Within the Locke sample set for tail CNVs, 25% of deletions show more than 50% base pair overlap with closed DNA compared to only 15% for duplications. About 18% of Hill duplications overlap closed DNA (with more than 50% base pair coverage) compared to 0-1% for deletions found in the three tissues.

Table 3.5 Extent of overlap between deletions and regions of closed DNA

Sample set	Sample subset (number of samples)	Tissue type	State	Total number of deletions	Number of deletions by percentage of base pair overlap with closed DNA[1]					
					0	>0-<25	>26-<50	>51-<75	>75-<99	100
Locke[2]	Classical mice (n=120)	Tail	0	424	3	109	209	103	0	0
			1	867	19	387	301	107	50	3
			Total deletions	**1291**	**22**	**496**	**510**	**210**	**50**	**3**
	Wild caught mice (n=23)		0	231	10	45	108	64	4	0
			1	491	1	150	240	85	7	8
			Total deletions	**722**	**11**	**195**	**348**	**149**	**11**	**8**
	Total Deletions (n=331)	**Tail**		**5656**	**85**	**1580**	**2553**	**1275**	**126**	**37**
Hill[3]	Cerebellum (n=12)	Cerebellum	0	28	0	2	15	11	0	0
			1	69	0	15	40	13	0	1
			Total deletions	**97**	**0**	**17**	**55**	**24**	**0**	**1**
	Liver (n=5)	Liver	0	10	0	1	5	4	0	0
			1	10	0	1	8	1	0	0
			Total deletions	**20**	**0**	**2**	**13**	**5**	**0**	**0**
	Spleen (n=10)	Spleen	0	18	0	0	17	1	0	0
			1	62	0	33	29	0	0	0
			Total deletions	**80**	**0**	**33**	**46**	**1**	**0**	**0**
	Total (n=27)	**Three tissues**		**197**	**0**	**52**	**114**	**30**	**0**	**1**

[1]DNase I sensitivity signal annotation tracks were downloaded from the University of California, Santa Cruz (UCSC) Genome

Browser: http://genome.ucsc.edu/cgi-bin/hgFileUi?db=mm9&g=wgEncodeUwDNase.

[2]Copy number variant calls were published in Locke *et al*. and discovered using publically available Mouse Diversity Genotyping Array data[1]. Mouse Diversity Genotyping Array CEL files were downloaded from the Center for Genome Dynamics at the Jackson Laboratory (http://cgd.jax.org/datasets/diversityarray/CELfiles.shtml).

[3]CNV calls were obtained from the Hill Laboratory (Milojevic, unpublished).

Table 3.6 Extent of overlap between duplications and regions of closed DNA

Sample set	Sample subset (number of samples)	Tissue type	State	Total number of Duplications	Number of Duplications by percentage of base pair overlap with closed DNA[1]					
					0	>0-<25	>26-<50	>51-<75	>75-<99	100
Locke[2]	Classical laboratory mice (n=120)	Tail	3	887	7	537	237	70	14	22
			4	646	1	263	231	61	78	12
			Total Duplications	**1533**	**8**	**800**	**468**	**131**	**92**	**34**
	Wild caught mice (n=23)		3	109	0	62	33	13	1	0
			4	138	0	59	62	17	0	0
			Total Duplications	**247**	**0**	**121**	**95**	**30**	**1**	**0**
	Total Duplications (n=331)	**Tail**		**3978**	**20**	**2041**	**1283**	**402**	**187**	**45**
Hill[3]	Cerebellum (n=12)	Cerebellum	3	53	0	7	38	5	0	3
			4	34	0	6	17	8	0	3
			Total Duplications	**87**	**0**	**13**	**55**	**13**	**0**	**6**
	Liver (n=5)	Liver	3	8	0	1	7	0	0	0
			4	3	0	0	1	0	0	2
			Total Duplications	**11**	**0**	**1**	**8**	**0**	**0**	**2**
	Spleen (n=10)	Spleen	3	50	0	5	39	4	0	2
			4	18	0	2	12	1	0	3
			Total Duplications	**68**	**0**	**7**	**51**	**5**	**0**	**5**
	Total (n=27)	**Three tissues**		**166**	**0**	**21**	**114**	**18**	**0**	**13**

[1]DNase I sensitivity signal annotation tracks were downloaded from the University of California, Santa Cruz (UCSC) Genome

Browser: http://genome.ucsc.edu/cgi-bin/hgFileUi?db=mm9&g=wgEncodeUwDNase.

[2]Copy number variant calls were published in Locke *et al*. and discovered using publically available Mouse Diversity Genotyping Array data[1]. Mouse Diversity Genotyping Array CEL files were downloaded from the Center for Genome Dynamics at the Jackson Laboratory (http://cgd.jax.org/datasets/diversityarray/CELfiles.shtml).

[3]CNV calls were obtained from the Hill Laboratory (Milojevic, unpublished).

A very small number of deletions and duplications were found to have all base pairs to be overlapping with segments of closed DNA, referred to as 100% base pair overlap (Table 3.7). Only 0.85% of CNVs (82 of 9,634) from the Locke sample set were found to lie completely in closed DNA. A higher percentage of Hill CNVs – 3.86% – was found to overlap completely with closed DNA. Deletions do not completely overlap closed DNA more than duplications across all sample sets. In fact, no state 0 CNVs completely overlap with closed DNA and mostly duplications were found to completely overlap with closed DNA across all tissues. The Locke sample set showed 82 CNVs completely overlap closed DNA, of which 37 are deletions (state 1) and 45 are duplications (states 3 and 4). Across the Hill sample set**,** only the cerebellum tissue was found to have one deletion of state 1 that completely overlapped closed DNA. All other Hill CNVs completely overlapping closed DNA were duplications, with more state 4 than state 3 duplications. A disproportionate number of deletions in closed DNA was not found.

3.2.3 No CNVs were found to completely overlap with intervals of closed DNA that were unique to each tissue

While the percentages of genome-wide intervals of closed DNA were not significantly different between the tissues, total tissue-specific percentages of closed DNA however, varied significantly from tissue to tissue (Table 3.8). Tissue-specific percentages of closed DNA are defined as regions of the genome that are closed at specific positions and are not found to be closed at those positions in the other tissues. This allows for the mapping of closed DNA regions that are found in one tissue but not in other tissues at the same genomic positions. The liver, followed by the cerebellum, consisted of the highest percentage of tissue-specific closed DNA while the mesoderm contained the highest percentage of tissue-specific open DNA across all chromosomes (p-value of $3.10^{-35} < 0.05$ based on a single factor ANOVA).

CNVs across all tissues showed minimal overlap with tissue-specific closed and open regions of the genome (Table 3.9). Deletions and duplications were not found to overlap (more than 50% base pair coverage) with tissue-specific closed DNA across any sample sets regardless of tissue type. Deletions and duplications were not found to completely

Table 3.7 Extent of CNVs found to lie completely in closed DNA (100% base pair overlap)

Sample set	Sample subset (number of samples)	Tissue type	Total CNVs	CNVs in closed DNA[1]	Number of CNVs in closed DNA	Number of CNVs in closed DNA by state			
						0	1	3	4
Locke[2]	Classical laboratory mice (n=120)	Tail	2824	1.31%	37	0	3	22	12
	Wild caught mice (n=23)	Tail	969	0.83%	8	0	8	0	0
	Total Jackson (n=351)	**Tail**	**9634**	**0.85%**	**82**	**0**	**37**	**27**	**18**
Hill[3]	Cerebellum (n=12)	Cerebellum	184	3.80%	7	0	1	3	3
	Liver (n=5)	Liver	148	1.35%	2	0	0	0	2
	Spleen (n=10)	Spleen	31	16.13%	5	0	0	2	3
	Total Hill (n=27)	**Three tissues**	**363**	**3.86%**	**14**	**0**	**1**	**5**	**8**

[1]DNase I sensitivity signal annotation tracks were downloaded from the University of California, Santa Cruz (UCSC) Genome Browser: http://genome.ucsc.edu/cgi-bin/hgFileUi?db=mm9&g=wgEncodeUwDNase.

[2]Copy number variant calls were published in Locke *et al*. and discovered using publically available Mouse Diversity Genotyping Array data[1]. Mouse Diversity Genotyping Array CEL files were downloaded from the Center for Genome Dynamics at the Jackson Laboratory (http://cgd.jax.org/datasets/diversityarray/CELfiles.shtml).

[3]CNV calls were obtained from the Hill Laboratory (Milojevic, unpublished).

Table 3.8 Distribution of tissue-specific intervals of closed DNA as a percentage of total closed DNA[1,2]

Chromosome	Mesoderm[3] Embryonic day 11.5	Cerebellum[3] Adult 8 weeks	Liver[3] Adult 8 weeks	Spleen[3] Adult 8 weeks
1	6.54[±]	8.84	10.19	7.68
2	6.07	9.45	10.96	7.63
3	6.68	8.58	9.72	7.93
4	5.86	9.17	10.38	7.22
5	5.58	9.63	10.76	7.1
6	6.41	9.07	10.24	7.68
7	4.67	9.63	9.22	5.73
8	5.77	9.06	10.62	7.51
9	5.54	10.5	11.16	7.2
10	6.38	8.87	10.72	7.69
11	5.3	11.22	11.81	6.6
12	5.93	8.54	10.08	7.2
13	6.02	9.01	10.56	7.13
14	6.18	8.02	9.48	7.35
15	6.15	9.06	11.1	7.46
16	6.43	8.88	10.34	7.78
17	5.48	9.83	10.31	6.56
18	6.12	8.81	10.85	7.78
19	5.31	10.38	10.87	6.48

[1]Tissue-specific intervals of closed DNA refers to intervals of closed, inaccessible DNA not found in any other tissue; that is, the intervals of closed DNA were found to be unique to that tissue (tissue-specific).

[2]DNase I sensitivity signal annotation tracks were downloaded from the University of California, Santa Cruz (UCSC) Genome

Browser: http://genome.ucsc.edu/cgi-bin/hgFileUi?db=mm9&g=wgEncodeUwDNase. Percentages of closed DNA were derived from DNase I sensitivity signal annotation tracks.

[±]A darker shade of red indicates a relatively higher percentage of closed DNA.

overlap (100% base pair coverage) with any tissue-specific regions of the genome.

3.3 A subset of probes on the Mouse Diversity Genotyping Array are predicted to cross-hybridize to naked mole-rat DNA

Two sets of probe lists were generated based on alignment/complementarity to the two existing naked mole-rat genome assemblies, build 1 (male) and build 2 (female) (Table 2.4). Probe sequences of both IGP and SNP probes were found to complement without mismatch to 26,546 loci in build 1 (male) and 27,154 loci in build 2 (female) of the naked mole-rat genome assemblies. Sequence data for the naked mole-rat genome assemblies has not yet been mapped to chromosomes and therefore sequence matches between the MDGA probes and naked mole-rat builds include autosomal, sex, and mitochondrion DNA. Probe sequences aligning to more than one locus in the naked mole-rat genome were deemed as duplicate target probes and excluded from lists of probes aligning to the naked mole-rat genomes. A total of 6303 duplicate IGP and 80 SNP probes complemented without mismatch to more than one genomic locus and were excluded from the probe lists. Removal of 7748 IGP and another 175 SNP probes that did not meet original Yang *et al*. design specifications further reduced the number of probes found to complement to a single unique locus in the naked mole-rat genomes[37]. This results in 665 SNP probes and 16,542 IGPs that are predicted to cross-hybridize to a single copy target in build 1 of the naked mole genome and 673 SNP probes and 16,080 IGPs in build 2. Ultimately 17,207 and 16,753 single copy loci in build 1 and build 2, respectively, are targeted by MDGA probes. The probes represent 3.78% of stringent IGPs and less than one percent of stringent SNP probes.

Table 3.9 Extent of overlap between CNVs and tissue-specific closed intervals of DNA

Regions	Sample set	Tissue type (number of samples)	Total number of CNVs	Number of CNVs in tissue-specific closed DNA	Number of CNVs by percent base pair overlap with tissue-specific closed DNA				
					<0-<25	<25-<50	<50-<75	<75-<100	100
closed DNA	Locke[2] Hill[3]	Tail (n=351)	9634	9240	9155	85	-	-	-
		Cerebellum (n=12)	184	175	173	2	-	-	-
		Liver (n=5)	31	29	29	-	-	-	-
		Spleen (n=10)	148	142	142	-	-	-	-
open DNA	Locke[2] Hill[3]	Tail (n=351)	9634	8684	8647	37	-	-	-
		Cerebellum (n=12)	184	164	164	-	-	-	-
		Liver (n=5)	31	26	26	-	-	-	-
		Spleen (n=10)	148	137	137	-	-	-	-

[1]DNase I sensitivity signal annotation tracks were downloaded from the University of California, Santa Cruz (UCSC) Genome Browser: http://genome.ucsc.edu/cgi-bin/hgFileUi?db=mm9&g=wgEncodeUwDNase.

[2]Copy number variant calls were published in Locke *et al*. and discovered using publically available Mouse Diversity Genotyping Array data[1]. Mouse Diversity Genotyping Array CEL files were downloaded from the Center for Genome Dynamics at the Jackson Laboratory (http://cgd.jax.org/datasets/diversityarray/CELfiles.shtml).

[3]CNV calls were obtained from the Hill Laboratory (Milojevic, unpublished).

Table 3.10 Approximately 17,000 single copy loci in the naked mole-rat genomes are predicted to bind to stringent probes on the Mouse Diversity Genotyping Array

Probe[2]	Number of complementary loci in the naked mole-rat genomes[1]	
	Build 1	**Build 2**
SNP	665	673
IGP	16542	16080
Total	**17207**	**16753**

[1]Number of unique loci targeted by complementary stringent probe sequences to the naked mole-rat genomes. Naked mole-rat genome assemblies (build 1 and build 2) were downloaded from the National Centre for Biotechnology Informatics (NCBI) FTP server: ftp://ftp.ncbi.nlm.nih.gov/genomes/Heterocephalus_glaber/. Stringent probes refer to probes that meet inclusion criteria for appropriate probe design[2].

[2]Probes on the Mouse Diversity Genotyping Array were designed to target either the exons of genes for CNV calling (IGPs) or two potential alleles (A or B) at known SNP loci for SNP genotyping (SNP probes)[37].

Counts of full and partial IGP sets complementary to the naked mole-rat genome were determined to examine the extent of potential exon sequence homology. A full IGP set consists of six 25-mer probes complementing three unique loci – proximal, medial, and distal regions –on both DNA strands of the naked mole-rat genome (Figure 1A). Of the stringent list of IGPs, 260 full IGP sets were found to be complementary to both builds, build 1 and build 2 (Table 3.11). Another 15 full IGP sets were complementary to build 2 only. Partial IGP sets were those IGP sets with one or two unique probes (and not the full three) that were complementary to either naked mole-rat genome build. A higher number of partial IGP sets were found, with 1,401 IGP sets targeting two unique 25-mer regions and 12,802 IGP sets targeting one unique 25-mer region of an exon.

All SNP probes aligning to the naked mole-rat genomes belong to a unique probe set (and therefore target a unique SNP locus) that targeted either Mus SNP allele A or B. A total of 686 SNP probe sequences belonging to unique SNP probe sets – 418 SNP probes targeting Mus allele A and 268 SNP probes targeting Mus allele B – were found to align to the naked mole-rat genomes across the two genomes. Of these SNP probe sequence matches, the majority were shared between the two naked mole-rat builds albeit at different genomic positions.

BLAST of SNP probes (excluding offset probes but representing 100% of MDGA SNP loci) resulted in the discovery of one SNP probe set to align to a naked mole-rat genome. SNP probe sequences targeting both SNP allele A and B at the same locus in the naked mole-rat were not found. The probe sequences containing Mus SNP allele A and B (of one SNP probe set) aligned to two different loci in the naked mole-rat. The two different loci were not in close proximity to each other although both of the sequence matches were on the mitochondrial DNA of both species (chromosome MT on both builds).

3.3.1 Stringent probes complementary to the naked mole-rat genomes are associated with genic regions in the mouse

The stringent IGPs complementary to the naked mole-rat genomes target 25,245 unique Mus exons and their associated 7,063 unique Mus genes. Approximately six percent of

Table 3.11 A total of 275 full invariant genomic probe (IGP) sets complement both naked mole-rat genome builds[1,2]

	1/3 probe set[3]	2/3 probe set[3]	Full probe set[1]
Build 1 only	202	11	0
Build 2 only	415	55	15
Builds 1 & 2	12185	1335	260
All builds	**12802**	**1401**	**275**

[1]A full invariant genomic probe set requires three loci – a proximal, medial, and distal region – of a Mus exon to be covered by three unique probes (3/3 probes).

[2]Complementary IGPs are stringent IGPs that were found to complement without mismatch to the naked mole-rat genomes. Naked mole-rat genome assemblies (build 1 and build 2) were downloaded from the National Centre for Biotechnology Informatics (NCBI) FTP server: ftp://ftp.ncbi.nlm.nih.gov/genomes/Heterocephalus_glaber/. Stringent IGPs meet inclusion criteria for appropriate probe design[1].

[3]A partial invariant genomic probe set refers to only one or two unique loci being targeted (through sequence complementarity) by one or two unique probes (1/3 or 2/3 probes).

all Mus exons known to exist in the mouse show sequence complementarity to some extent in the naked mole-rat genomes. Table 3.12a details the 7,325 unique and known, putative, or novel Mus genes targeted by the IGPs. Complementary sequences from 18% of all Mus genes and 6% of all Mus exons (build 37; Ensembl 67) can be targeted by the MDGA probes in experimental naked mole-rat samples (Table 3.12b). Approximately 29% of all protein coding Mus genes are associated with IGPs predicted to bind to naked mole-rat DNA.

Conservation of exons between the mouse and the naked mole-rat genomes can be examined using a stringent list of the 260 full IGP sets that have a relatively high likelihood of detecting genes (all three exons of a gene have potential to be interrogated in the naked mole-rat). The additional 15 full probe sets were excluded because of complementarity to only one build. The 260 probe sets target exons coded by 239 unique genes. The genes show functional enrichment for general gene ontology (GO) categories that are involved in the regulation of biological processes and developmental processes as well as specific functions with the top three being transcription, regulation of transcription, and positive regulation of transcription. Pathway analysis revealed four major gene networks involving 30, 25, 23, and 21 genes respectively, from an initial 239 genes (Table 3.13). The genes involved in the four networks are all involved in general cellular functioning. Top networks were deemed as such if scores were higher than two since a score of two reflects at least a 99% confidence of not being generated by random chance alone. Scores are based on a p-value of less than 0.05 from a Fisher's Exact Test.

For the SNP probes complementary to the naked mole-rat genomes, 57% (389 of 688) were found to lie to some extent within genic regions and encompass 358 unique genes. The top general GO terms associated with the genes are cellular component organization, development, behavior, and regulation of biological process. The top three specific GO terms are related to transcription: DNA-templated transcription, positive regulation of transcription from RNA polymerase II promoter, and regulation of DNA-templated transcription. The top five networks associated with the 358 genes involved 49, 26, 23, 21, and 20 genes respectively. The genes in the five networks are all involved in various

biological functions with a relatively strong association with developmental processes (Table 3.14). Only five SNP probe sequences (of all complementary SNP probes) were found to have predicted nonsynonymous effects, specifically missense changes, in the mouse. These genes are of functional relevance and are detailed in Table 3.15.

The SNP probe set pair found in the naked mole-rat genome was found on chromosome MT in build 2 of the naked mole-rat genome however the SNP probe sequence for allele A and allele B aligned to two unique genomic loci in the naked mole-rat. The SNP probe sequence lies on the highly conserved Mus 16S ribosomal RNA (16 rRNA) gene. The 16S rRNA is associated with four GO annotations that are involved in two major GO classifications: biological processes (cellular component organization and protein metabolic process) and cellular components (mitochondrion, non-membrane-bounded organelle, and organelle lumen).

Redundancy in gene coverage by the complementary probes can reveal an overrepresentation of specific Mus genes and exons in the naked mole-rat. Much redundancy exists in genes targeted by IGPs as outlined in Table 3.12a, where total counts of gene coverage are compared to unique counts. Some genes are overlapped by unique IGPs multiple times with up to 36 unique IGP sequences targeting one gene (intentional by design). All unique and overrepresented genes are known Mus protein coding genes. Genes being targeted by more than ten unique IGP sequences show functional enrichment for cellular component organization (top GO category). Eighty-five percent of the 389 SNP probes that overlap known Mus genes are unique genes; overrepresented genes or genes that are covered by more than one unique SNP probe sequence are observed to different extents across the two naked mole-rat builds. The overrepresented genes are covered by unique SNP probe sequences either twice, thrice, or four times. The overrepresented genes show enrichment in genes related to developmental processes (33 unique genes).

A number of stringent IGP and SNP probes that interrogate one specific genomic locus in the mouse was found to be complementary to more than one genomic locus in the naked

mole-rat. A total of 2284 unique IGPs and 55 SNP probe sequences containing either SNP allele A or B complement two or more unique loci in the naked mole-rat genomes.

Table 3.12a Counts of all Mus genes covered by stringent IGPs predicted to bind to a naked mole-rat genome

	Gene descriptor[1]	Total gene count[2]	Unique gene count[3]
Gene Status	Known[4]	16192	6774
	Novel[5]	289	165
	Putative[6]	100	35
	Total	**16581**	**6974**
Gene type	Protein coding	15926	6591
	Antisense	204	117
	lincRNA	191	106
	miRNA	153	103
	Pseudogene	27	16
	Processed transcript	31	15
	snoRNA	16	13
	Non-coding	24	7
	snRNA	2	2
	Sense overlapping	1	1
	Polymorphic pseudogene	1	1
	3 prime overlapping ncRNA	4	1
	IGV	1	1

Total	**16581**	**6974**

[1]All complementary stringent IGPs were mapped to genes on build 37 of the mouse genome (mm9: Enseml 67). Gene list was downloaded from archives on Biomart's Ensembl: http://www.ensembl.org/info/website/archives/index.html. Stringent probes are Mouse Diversity Genotyping Array probes that meet all inclusion design criteria[1]. Gene types include: lincRNA (long intergenic non-coding RNA), miRNA (microRNA), snoRNA (small nucleolar RNA), snRNA (small nuclear RNA), ncRNA (non-coding RNA), and IGV (immunoglobulin variable gene).

[2]Total counts include overrepresented genes where overrepresentation refers to many unique IGP sequences covering the same gene.

[3]Unique counts are based on unique Ensembl Gene IDs.

[4]Known genes have an official gene name, symbol, and function.

[5]Genes classed as novel are those protein coding genes that do not have an available official gene name and symbol.

[6]Putative refers to a segment of DNA that is believed to be a gene based on its open reading frame; however gene function is unknown.

Table 3.12b Proportion of all Mus genes associated with stringent IGPs predicted to bind to a naked mole-rat genome

	Gene descriptor[1]	Number of Mus genes[2]	Number of Mus genes targeted by complementary IGPs[3]	Percentage of all Mus genes targeted by complementary IGPs
Gene Status	Known[4]	30416	6774	22%
	Novel[5]	6647	165	2%
	Putative[6]	928	35	4%
	Total	**37991**	**6974**	**18%**
Gene type	Protein coding	22707	6591	29%
	Pseudogene	5474	16	<1%
	lincRNA	2057	106	5%
	miRNA	1639	103	6%
	snoRNA	1560	13	1%
	snRNA	1429	2	<1%
	Antisense	1381	117	8%
	miscRNA	491	-	-
	IGV	355	1	<1%
	rRNA	338	-	-
	Processed transcript	299	15	5%
	IGJ	88	-	-
	Sense intronic	78	-	-

	Gene descriptor[1]	Number of Mus genes[2]	Number of Mus genes targeted by complementary IGPs[3]	Percentage of all Mus genes targeted by complementary IGPs
Gene type	IGD gene	25	-	-
	MT tRNA	22	-	-
	IGC gene	13	-	-
	Non-coding	12	7	58%
	Polymorphic pseudogene	8	1	13%
	Sense overlapping	8	1	13%
	3 prime overlapping ncrna	3	1	33%
	ncRNA host	2	-	-
	MT rRNA	2	-	-
	Total	**37991**	**6974**	**18%**

[1]A complete list of Mus genes found on build 37 of the mouse genome (mm9: Enseml 67) can be downloaded from archives on Biomart's Ensembl: http://www.ensembl.org/info/website/archives/index.html. Gene types include: lincRNA (long intergenic non-coding RNA), miRNA (microRNA), snoRNA (small nucleolar RNA), snRNA (small nuclear RNA), miscRNA (micscellaneous RNA), IGV (immunoglobulin variable gene), rRNA (ribosomal RNA), IGJ (immunoglobulin J gene), IGD (immunoglobulin D gene), MT tRNA (mitochondrial transfer RNA), IGC (immunoglobulin constant gene), ncRNA (non-coding RNA), and MT rRNA (mitochondrial ribosomal RNA).

[2]Counts are based on the complete set of unique Ensembl Gene IDs representing all

genes in the mouse genome (build 37).

[3]All complementary stringent IGPs were mapped to genes on build 37 of the mouse genome (mm9: Enseml 67). Stringent probes are Mouse Diversity Genotyping Array probes that meet all inclusion design criteria[1]. Naked mole-rat genome assemblies (build 1 and build 2) were downloaded from the National Centre for Biotechnology Informatics (NCBI) FTP server: ftp://ftp.ncbi.nlm.nih.gov/genomes/Heterocephalus_glaber/.

[4]Known genes have an official gene name, symbol, and function.

[5]Genes classed as novel are those protein coding genes that do not have an available official gene name and symbol.

[6]Putative refers to a segment of DNA that is believed to be a gene based on its open reading frame; however gene function is unknown.

Table 3.13 Full IGP sets complementing both naked mole-rat genomes are enriched for gene networks involved in basic cellular processes and development[1]

Score[2]	Number of genes involved[3]	Top functions
35	30	Cell Death and Survival, Cell-To-Cell Signaling and Interaction, Nervous System Development and Function
27	25	Gene Expression, Cell Cycle, DNA Replication, Recombination, and Repair
24	23	Cellular Development, Embryonic Development, Organismal Development
21	21	Cellular Assembly and Organization, Organismal Survival, Cellular Movement

[1]Naked mole-rat genome assemblies (build 1 and build 2) were downloaded from the National Centre for Biotechnology Informatics (NCBI) FTP server: ftp://ftp.ncbi.nlm.nih.gov/genomes/Heterocephalus_glaber/.

[2]Score is based on a p-value of less than 0.05 from a Fisher's Exact Test. A score of 2 reflects at least a 99% confidence of not being generated by random chance alone (Ingenuity Pathway Analysis). Top networks are ranked according to the score that directly informs of the likelihood of genes being associated with each other in with gene networks.

[3]All complementary stringent IGPs were mapped to genes on build 37 of the mouse genome (mm9: Ensembl 67). A complete list of Mus genes found on build 37 of the mouse genome (mm9: Enseml 67) were downloaded from archives on Biomart's Ensembl: http://www.ensembl.org/info/website/archives/index.html. Stringent probes are Mouse Diversity Genotyping Array probes that meet all inclusion design criteria[1].

Table 3.14 Gene networks for development are associated with the 358 genes targeted by SNP probes complementing the naked mole-rat genomes

Score[2]	Number of genes involved[3]	Top functions
49	67	Gene Expression, Cellular Development, Organismal Development
26	25	Nervous System Development and Function, Organismal Survival, Cell Morphology
23	21	Amino Acid Metabolism, Small Molecule Biochemistry, Metabolic Disease
21	18	Cell-To-Cell Signaling and Interaction, Nervous System Development and Function, Cellular Assembly and Organization
20	17	Nutritional Disease, Embryonic Development, Organismal Development

[1]Naked mole-rat genome assemblies (build 1 and build 2) were downloaded from the National Centre for Biotechnology Informatics (NCBI) FTP server: ftp://ftp.ncbi.nlm.nih.gov/genomes/Heterocephalus_glaber/.

[2]Score is based on a p-value of less than 0.05 from a Fisher's Exact Test. A score of 2 reflects at least a 99% confidence of not being generated by random chance alone (Ingenuity Pathway Analysis). Top networks are ranked according to the score that directly informs of the likelihood of genes being associated with each other in with gene networks.

[3]All complementary stringent IGPs were mapped to genes on build 37 of the mouse genome (mm9: Ensembl 67). A complete list of Mus genes found on build 37 of the mouse genome (mm9: Enseml 67) were downloaded from archives on Biomart's Ensembl: http://www.ensembl.org/info/website/archives/index.html. Stringent probes are Mouse Diversity Genotyping Array probes that meet all inclusion design criteria[1].

Table 3.15 Five Mus genes with SNPs having known nonsynonymous effects are targeted by five SNP probes predicted to bind to a naked mole-rat genome[1]

Ensembl gene ID[2]	Symbol	Gene or protein name	Chromosome	Location
ENS*MUS*G00000033671	*Cep350*	centrosomal protein 350	1	155844964 - 155973255
ENS*MUS*G00000026042	*Col5a2*	collagen, type V, alpha 2	1	45374321 - 45503282
ENS*MUS*G00000075210	*Olfr1012*	olfactory receptor 1012	2	85759439 - 85760374
ENS*MUS*G00000004508	*Gab2*	growth factor receptor bound protein 2-associated protein 2	7	97081586 - 97308946
ENS*MUS*G00000025195	*Dnmbp*	dynamin binding protein	19	43846821 - 43940191

[1]Naked mole-rat genome assemblies (build 1 and build 2) were downloaded from the National Centre for Biotechnology Informatics (NCBI) FTP server: ftp://ftp.ncbi.nlm.nih.gov/genomes/Heterocephalus_glaber/.

[2]SNP probes complementary to the naked mole-rat genomes that were associated with genes having nonsynonymous effects were determined from an updated variant annotation file based on build 38 (mm10) (Qi, unpublished). Ensembl gene IDs of interest were used as a query in the Mouse Genome Informatics (MGI) database of known mouse genes (http://www.informatics.jax.org/marker/). Only stringent SNP probes were included for this analysis with stringent referring to Mouse Diversity Genotyping Array probes that meet all inclusion design criteria[1].

Since these probe sequences were found to recur in the naked mole-rat genomes, associated Mus genes may be of biological significance. About 1300 Mus genes are targeted by the probes predicted to bind to multiple loci in the naked mole-rat genome. These genes are involved in general cellular component organization, cellular metabolic processes, and regulation of biological processes.

Chapter 4 Discussion

4.1 Correction or removal of incorrect probe annotations improves accuracy in variant detection

A false positive is a type I error where the array calls a CNV in the genome in a region where it does not actually exist. False calling of CNVs, specifically copy number losses, can occur when sample DNA fails to hybridize to complementary probe sequences due to probes not meeting design criteria or incomplete sample preparation that does not remove bound protein. Probe annotations that are associated with incorrect sequence or positional information can lead to the calling of a variant that does not truly exist in the sample i.e., not a true biological variant. False negatives in array data refer to type II errors where true biological variation is not detected by the array. If a probe's annotated location is incorrect and is subsequently used for variant calling, the variant may potentially be mapped to the wrong genomic locus. False negatives can also arise from low resolution, widely spaced gaps between probes (uneven distribution), and not having enough probes (low probe density). CNVs that exist in regions of closed, inaccessible DNA may be unavailable for hybridization to the array and lead to underreporting of true biological CNVs. An increase in false negatives can be a problem for deletions (by increasing the number of state 0 deletions and not reporting the true copy number) and duplications that exist in these regions of closed, inaccessible.

IGPs that do not target a unique locus in the mouse genome are another source of error in array data. CNV calling is based on probes that target a unique DNA segment found only once for a haploid genome. Having probe sequences that complement without mismatch to multiple target sequences across the genome i.e., probes not targeting a unique, single locus and sequence in the genome results in algorithms starting off with the wrong assumption that there is only one IGP sequence targeting one unique locus in a haploid genome. If probes do not meet design criteria for uniqueness, it is not possible to correctly interpret fluorescent intensities from the probes and subsequent comparisons of copy number between genomes are not meaningful. Identical probe sequences annotated to multiple loci across the genome can also affect how the fluorescent intensities from

those probes are interpreted because the algorithm will not able to attribute the fluorescent intensities to the correct locus.

It is important to be aware that a particular DNA segment can exist in only one physical genomic location and that allowing for the particular DNA segment to bind to duplicate probe sequences with differing annotated genomic locations leads to erroneous CNV calls. A CNV call may be mapped to a genomic region that does not harbor a true biological CNV resulting in a false positive, or type I error. Duplicate probe sequences also lead to the likelihood of missed CNV calls (especially duplications) due to competition of probe sets for hybridization to the limited sample DNA fragments. A limited amount of sample DNA available for hybridization to two sets of identical probe sequences leads to competition between the two sets of probes and ultimately results in a reduced or nonexistent observable fluorescent intensity at one or both probe sets. Both false positive and false negative calls can increase when duplicate probe set annotations are not filtered out prior to CNV calling.

IGPs annotated to incorrect genomic locations result in the calling of copy number states at incorrect locations and can lead to the fragmenting of CNV calls. A minimum of three consecutive probes of the same copy number state covering a region is required for a putative CNV to be called[1]. Even when sample DNA harboring the CNV binds to probes, if a single probe is annotated to an incorrect genomic location, the entire CNV may not be called or a larger CNV may not be detected because the probe annotated to the incorrect location will not be included in the group of consecutive probes (and associated copy number states) that are necessary for the entire CNV to be called. Furthermore, fragmenting of CNV calls can happen in two other ways: 1) a probe that is annotated to the incorrect location may lead to the probe being included in the calling of another CNV when it should not be and 2) the probe being mapped to the wrong location can break up a CNV call at another location by preventing consecutive probes of the same copy number state to be called as a CNV. Fragmentation of CNV calls under- or overestimates the number of true biological CNVs and underestimates the size of CNVs. Since copy number state must be called upon consecutively within a certain distance, incorrectly

annotated probes can also contribute to a false negative, or type II error where a true biological CNV is undetected because of extraneous CNV state calls in the region of the CNV.

Compromised hybridization success due to poor SNP probe design adversely affects the genotype assignment of a SNP probe intensity by resulting in the increased reporting of 'no calls'. A 'no call' is likely to occur when the SNP genotyping algorithm is unable to resolve an individual fluorescent intensity into one of the SNP genotype calls – AA, AB, and BB. A 'no call' indicates that there was an inconsistency or ambiguity in the fluorescent intensity and that it did not meet sufficient criteria to be characteristic of one of the three SNP genotypes. In other words, the particular sample did not achieve the statistical threshold required for genotyping. False negative and false positive errors for SNP genotyping are also affected by incorrect annotations of SNP alleles or SNP allele positions. The removal of probe annotations being unmappable to the mouse reference genome and annotations having the incorrect SNP allele at the specified SNP locus is an important step. This step ensures that probe sequences designed for the mouse are in fact detecting the target genome and that the correct SNP at the specified SNP allele is being genotyped.

Extensive filtering of MDGA probes resulted in the removal of probes targeting well over 200,000 loci in the mouse. Filtering refers to the removal or correction of probe annotations that are incorrect or don't meet probe design criteria such that only accurate probe annotations are used for SNP genotyping and CNV calling. Since the physical probes on the array cannot be fixed or removed, filtering of probe annotations is necessary. The resulting filtered list of probes is called a stringent list of probes with stringent referring to all design criteria being met. The resulting capabilities of the array to interrogate the mouse genome are demonstrated through the stringent probe lists that assay 91% of all protein-coding Mus genes and 68% of all Mus genes. The post-filtered, stringent probes provide an unbiased coverage of protein-coding Mus genes. It is important to note that since the IGPs are designed to target exons, the array has a bias for detecting CNVs involving exonic regions. The post-filtered, stringent IGPs maintain their

ability to provide coverage of most Mus genes for genic CNV with coverage of 89% of all protein-coding genes and 63% of all Mus genes. Stringent SNP probes assay 68% of all protein-coding Mus genes and 49% of all Mus genes. The lower coverage of genes by SNP probes is expected because not all SNP probes were designed to target the exons of genes and can exist in intergenic regions of the genome.

4.2 DNase I accessibility landscapes are predicted to be useful as a measure of target DNA accessibility for hybridization

4.2.1 Embryonic tissues have relatively more open landscapes of DNase I accessibility

Embryonic mesoderm tissue having more a relatively more accessible, open genomic landscape than genomic landscapes of adult cerebellum, liver, and spleen tissues is consistent with literature demonstrating that embryonic genomes (in this case, embryonic day 11.5) are more transcriptionally active than genomes of adults[5,53,59,99]. Furthermore, tissue-specific intervals of open DNA – intervals of open, accessible DNA found to be unique to the tissue – were found to be highest in the mesoderm by a significant amount. The cerebellum contained the one of the highest levels of total closed DNA (and especially tissue-specific intervals of closed DNA) across the autosome, consistent with literature showing tight regulation of transcription in the brain more than other tissues[53,56,58]. Specifically, adult neurons are known to show high levels of DNA methylation, or the silencing of genes, which is consistent with a more closed genomic landscape in adult brain cells[55,58].

4.2.2 A lack of correlation between deletions across all sample sets with closed DNA suggests effective protein removal from protein-bound DNA

The issue of ineffective protein removal in array experiments was previously examined by van Heesch, whereby a longer proteinase K digestion was shown to increase DNA yield and improved variant calling from brain tissue but not other tissues[4]. This is consistent with brain tissue samples having a relatively closed genome due to the higher levels of DNA-bound proteins[55,56,58]. The same study also demonstrated that a longer

proteinase K digestion of samples derived from diverse tissue and cell types can improve DNA content uniformity[4]. Though my results show that CNV calls from the cerebellum and liver show a relatively higher percent overlap with closed DNA, the difference is minimal and therefore indicative of minimal tissue-specific chromatin effects. Any significant differences in overlap with closed DNA between tissues or between mouse strains would indicate ineffectiveness of a DNA isolation protocol designed for only one tissue type. However, it seems ineffective DNA isolation did not specifically affect any one tissue because deletions and duplications were found in predicted regions of closed DNA.

CNVs across all tissues showed minimal overlap with tissue-specific intervals of closed DNA (unique to each tissue) and importantly, neither deletions nor duplications were found to completely overlap with any of the tissue-specific closed regions of the genome. A modified DNA extraction protocol with an extended proteinase K digestion as an added precaution was used for Hill tissue samples and seems to have reduced tissue-specific chromatin effects that can adversely affect DNA yield and CNV calling.

Results show that there was no evidence consistent with a hypothesis of protein-bound DNA being unavailable for hybridization and increasing CNV deletions. This is because an overrepresentation of deletions and underrepresentation of duplications in regions of closed DNA was not found. Minimal overlap of closed DNA across all CNVs is expected despite effective removal of proteins due to the fact that both CNV boundaries and DNase I accessibility intervals are best estimates made by the CNV calling algorithm and the DNase I sequencing technology. Being able to detect duplications predicted to lie in closed regions of the genome indicates that protein-bound DNA did not seem to increase the percentage of false discovery. In fact, duplications (and not deletions as hypothesized) being observed to coincide most with closed regions of DNA provides stronger evidence for effective removal of proteins. There are, however, several caveats to this conclusion.

DNase I sensitivity data for each tissue were obtained from a database and are therefore a proxy for the tissues used for CNV detection. DNase I sensitivity assays were not performed on the same tissues used for CNV analyses. DNase I sensitivity data for each

tissue were chosen based on similarity to the tissues used for CNV analyses. While similar, each tissue for which DNase I sensitivity data are derived from vary slightly in terms of mouse strain, age, and germ layer from the tissues used for CNV analyses[66]. Whereas Hill Laboratory samples from cerebellum, liver, and spleen are derived from adult mice ranging in age from 4.4 months to 11.4 months with a genetic background of either CBA/CaJ (8 mice) or B6 (2 mice), the DNase I sensitivity data are derived from the cerebellum, liver, and spleen tissues of only B6 adult mice at the age of 8 weeks. These tissues are relatively comparable due to similarity in mouse strain (highly inbred) but may not be completely comparable in terms of age though this difference is only up to 4 months. Furthermore, the mesoderm tissue from which DNase I sensitivity data were obtained are from CD-1 mice at embryonic day 11.5 compared to the adult Jackson tail samples (derived from both the mesoderm and ectoderm) from a variety of mouse strains[100]. CD-1 mice are not representative of the classical laboratory mouse strains and other strains used in the study. An embryonic genome is not representative of an adult genome since genomes are much more transcriptionally active during embryogenesis than during adulthood[53,66,101,102]. For this reason, the tissues (and their DNase I data) used as a proxy for this study may not be completely appropriate in terms of both Mus strain and age.

Direct comparisons of tissue and cell-type-specific DNase I sensitivity data to CNV data from the same cell and tissue type would be much more reliable for examining chromatin accessibility in CNV studies, however in the absence of such data and feasible alternative measures, DNase I accessibility measures from similar tissues and cell types has potential for use as a proxy[103,104]. It is unfortunately not possible to compare DNase I sensitivity data from a tail sample to CNV calls derived from the tail because whole genome DNase I sensitivity data were available. Performing DNase I sensitivity assays in-house on the same tissue being studied for CNVs would allow for the most quantifiable and reliable comparison between detected CNVs and regions open and closed DNA. Carrying out such a post-hoc analysis is reasonable and informative in the absence of validation. It is, however, difficult to accurately estimate and quantify the protein-bound landscape of the DNA used for the CNV analysis using DNase I accessibility data from other tissues

acting as proxies. To accurately quantify the ability of DNase I accessibility data from proxy tissues to estimate the landscape of closed DNA in tissues of interest, DNase I sensitivity measures from the tissues of interest and from the proxy tissues need to be directly compared and thus necessitates empirical testing of the results from this analysis.

Despite having statistical thresholds for genotyping and marker requirements for CNV calling, microarray data are not without technical error. Previous studies have compared CNV calls from various CNV calling algorithms and CNV detection technologies[31,44,105,106]. These studies demonstrate the need for validation of putative CNV calls since there is no single CNV algorithm or detection platform that can target the full extent of CNVs throughout the genome. For example, array-based platforms are a cost-effective method for CNV discovery but are limited in their ability to detect CNVs of smaller size. CNV algorithms also differ significantly in the number and size of CNVs that are called[31,32,107]. Ultimately, confirmation of the presence and state of a CNV is typically achieved by digital droplet PCR (ddPCR), NGS, or other high-throughput sequencing methods[108]. When validation is not possible, it is recommended to use a second algorithm to produce more informative results [44]. The use of a second algorithm can eliminate false positives that were undetected by either software and increases confidence in replicated CNV calls. Using a secondary method such as aCGH can also confirm CNV calls by providing replication in data or by discovering CNVs missed by the SNP array.

4.3 A subset of Mouse Diversity Genotyping Array probes is predicted to hybridize to naked mole-rat target DNA

4.3.1 Potential segmental duplications may be revealed by probe sequences aligning to multiple loci in the naked mole-rat genome

A number of unique probe sequences in the mouse were found to align to more than one locus (recurring) in the naked mole-rat genomes. These probe sequences exist as a single copy (i.e., unique locus) in the mouse genome but do not complement to a single, unique locus in the naked mole-rat genomes and were excluded from the final stringent probe lists for potential cross-hybridization, as the use of duplicate probes in a microarray study

results in competition for DNA and adversely affects the detection of variants. However, the analysis of probe sequences that align to more than one region in the naked mole-rat genome can give insight to potentially important genes and functional attributes. It is possible recurring Mus sequences in the naked mole-rat genomes may show enrichment for gene functions in cellular component organization, cellular metabolic processes, and regulation of biological processes.

Probe sequences existing as a single copy in the mouse genome but aligning to more than one locus in the naked mole-rat genome can also provide insight to potential (gene) duplications. Duplication events of DNA segments can lead to a particular DNA segment to exist in two (or more) different positions within the same genome. Gene duplications can affect the structure of entire genomes and underlie genome evolution. Gene duplication is believed to play a major role in evolution whereby increases in gene copy number can have effects on protein function, dosage effects, and fitness of an organism[109–111]. Specific probe sequences existing in multiple copies (in close proximity) along a naked mole-rat chromosome may allow for the study of gene duplications, particularly segmental duplications. An analysis of duplication events in the naked mole-rat genome is possible in the future when the sequences that comprise the naked mole-rat reference genomes are mapped out to chromosomes.

Segmental duplications – segments of highly homologous duplicated DNA (> 1 kb) –are directly related to the rate of evolution in species and are therefore of interest[16,112]. An SD analysis found that the naked mole-rat had the lowest percentage of SD (3.20) compared to the mouse (4.70), rat (3.30), and human (3.59) genomes[15,17,84]. This indicates a relatively slow rate of evolution in the naked mole-rat. Additionally, more than 90% of the naked mole-rat having synteny to human, mouse, and rat genomes suggests a relatively low rate of naked mole-rat genome rearrangements after diverging from the murid common ancestor[84]. Based on the low rate of SD in the naked mole-rat, detection of potential duplication events by MDGA probes may be less likely although some conservation in the order of Mus exonic sequences is expected due to the low rate of genome rearrangements since diverging from the common ancestor of the mouse[84].

Recurring IGP sequences were found to exist in the naked mole-rat, however due to the short length of the sequences, the chance for random repeat matches is higher than for a longer probe sequence[113]. For that reason, it is more informative to consider full probe sets (275) (targeting the proximal, medial, and distal regions of a Mus exon) that aligned to the naked mole-rat genomes.

Evidence for recurring exonic sequences in the naked mole-rat was provided through BLAST results. However, BLAST results indicated only that a probe sequence aligned to more than one region in the naked mole-rat and not how many times each recurring sequence occurs or at what genomic positions each sequence aligns to. A more thorough analysis would include the full BLAST output that contains alignment positions for the recurring sequences, as well as the number of times that each sequence aligns to a different genomic locus across the naked mole-rat genome. It is important to note a major limitation to this analysis is the lack of naked mole-rat sequence data by chromosome. Sequence data is mapped out to scaffolds but not yet to chromosomes making a side by side comparison of chromosomal sequence data unfeasible. Having undetermined chromosomes also limits predictions for potential segmental duplication events or conserved exons. To determine potential segmental duplication events, a future study could examine the proximity of IGP sequence matches along a naked mole-rat chromosome for each IGP set. This could would be used to confirm whether IGPs are indeed conserved in the same order in proximity to each other and may potentially reveal proximal, medial, and distal regions of potential exons conserved in the naked mole-rat.

4.3.2 Exon conservation between mouse and naked mole-rat genomes is difficult to explore through complementary invariant genomic probes

DNA regions are typically conserved across species due to phenotypic or functional relevance[114]. Coding genes in particular are subject to positive selective advantage or negative selection if mutated[114]. It is, therefore, of interest to determine the extent of sequence similarity and possibility of homology between the Mus IGP sequences and the naked mole-rat genomes. Conservation of IGP targeted exons is useful for identifying potential coding regions and in this analysis, for identifying potentially functional

sequence fragments in the naked mole-rat. Because there is no database for known naked mole-rat genes, the nature of exonic regions aligning to the naked mole-rat genome was examined using known Mus exonic and genic databases. A similar approach was carried out by Hoffman *et al.* whereby a small subset of known canine genes nearest to cross-amplified SNPs in the seal were examined for potentially conserved genes and gene fucntions[75]. A small subset of all Mus (1300) genes representing 18% of all Mus genes and 29% of Mus protein coding genes is queried by stringent IGPs. The small number is expected due to the many years of divergence.

Because only 3.78% of stringent IGPs complement naked mole-rat genomes, it is difficult to predict any significant cross-hybridization of IGPs for the study of CNV. Additionally, a low percentage of complementary IGP sets is expected since the extent of sequence conservation, or homology between species generally decreases with increasing years of divergence. Gene order is also less and less conserved with increasing evolutionary distance[115–119]. Ultimately, the number of IGP sets is expected to be low because increased phylogenetic distance is associated with increased sequence dissimilarity due to mutation, recombination, exon shuffling, genetic drift, and other factors in response to varying selection forces such habitat and mating.

Conversely, some conservation in the order of Mus exonic sequences is still expected due to the low rate of genome rearrangements since diverging from the common ancestor of the mouse[84]. Gene order is less conserved with increasing phylogenetic distance but some associations can exist[120,121]. For example, genes with similar function or expression cluster more commonly than other genes[122]. High levels of inbreeding in naked mole-rat populations also contribute to significantly lower levels of genomic recombination and exon shuffling which can increase the likelihood of linkage groups remaining in a state of disequilibrium over time. Since linkage groups are directly related to the genetic structure of a population, linkage groups may reflect inbreeding of a population. It is already known that the low level of nucleotide diversity in the naked mole-rat is reflective of its population genetic structure and practices of inbreeding[123].

Conservation of exons and genes also does not decrease linearly with sequence divergence and certain structures are known to be highly conserved despite phylogenetic distance[120–122,124]. Even when protein sequence similarity is low, certain gene structures can be very similar between populations and species[124]. The degree to which exon and gene structure is conserved within diverse protein domains is complicated. One study found gene structure conservation despite low protein sequence similarity[125]. Gene content and order were found to be highly conserved and nonrandom in order when relative gene orientation, intergenic distance, and functional relationships were taken into account[119–122]. It is known that a characteristic of mammalian protein-coding genes is the high evolutionary conservation of exon-intron structure even with a large number of years of divergence[126,127]. IGP sequence BLAST results revealed 275 complete IGP sets of which 260 IGP sets were shared between both naked mole-rat genomes. The alignment of all three of the proximal, medial, and distal probe sequences of 260 exonic regions suggests higher confidence in potential conservation of the genes these probe sets target. A significantly higher number (1401) of partial IGP sets – with two of the three exonic probe sequences aligning to the naked mole-rat genomes – reveals that examining exon conservation between the mouse and naked mole-rat is limited by the small number of complementary IGPs.

When examining sequence conservation between species, it is important to consider the definition of sequence identity and the statistical chance for sequence identity. Sequence identity from a bioinformatics perspective refers to the percentage of aligned nucleotides between two DNA sequences and establishes the likelihood for sequence homology based on evolution from a common ancestor. Only 100% sequence identity was considered for this study, that is BLAST results were considered a perfect match only if there was no mismatch between all 25 nucleotides of the probe sequence and the complementary sequence in the naked mole-rat genome. Secondly, the likelihood of sequence homology is complicated because not all sequences show homology with equal probability. The shorter a nucleotide sequence is, the higher the probability that sequence homology can be found by sheer chance[113]. It is, therefore, more meaningful to look at partial and full sets of IGPs that aligned to the naked mole-rat genomes. It is important to

do so because IGP sequence lengths are relatively short. Partial and full sets of complementary IGPs are also more informative because partial and full sets of IGPs aligning to the naked mole-rat genomes correlate with a higher probability of true homology in exonic sequences with more meaningful associated functions. It is especially meaningful if probe sequences for a Mus exon are found close together along naked mole-rat chromosomes though this analysis is not currently possible due to limitations of the naked mole-rat genome scaffolds.

Potential exon and/or gene conservation is best examined by looking at the 260 complete IGP sets (all three exonic sequences per gene) predicted to bind to both builds of the naked mole-rat genome. The IGP sets overlapping with Mus genes were associated with basic cell functioning networks and enrichment for genes involved in transcription and regulation of transcription. The findings are similar to those of a cross-hybridization study carried out by Hoffman *et al.* whereby genes nearest to polymorphic SNPs were found to be significantly enriched for functional annotations relating to energy metabolism, a basic function across diverse domains of life[75]. The results of this study are consistent with the existing literature and similarly suggest a possible bias towards conserved regions of the genome.

4.3.3 A very small number of SNP probe sequences aligned to the distantly-related naked mole-rat genomes

The very small number of unique SNP probe sequences (688) found to complement without mismatch to the naked mole-rat genomes is expected for three reasons: 1) sequence homology decreases with evolutionary divergence 2) the number of loci that remain polymorphic decreases exponentially with phylogenetic distance before leveling off after around five million years of divergence and 3) SNP probe sequences are not designed based on invariant genomic regions and instead, on highly variable regions that consist of polymorphic SNPs[70,113].

While the Mus SNP allele A and allele B from one SNP probe set aligned to the naked mole-rat genome, they did not target the same locus in the naked mole-rat. However, the SNP probe sequence containing Mus allele A was found at virtually the same position in

the naked mole-rat. The mitochondrially encoded gene 16S rRNA associated with this Mus SNP contains some of the most conserved DNA sequences across domains of life[128]. Highly conserved regions are typically required for basic cellular functions and development throughout evolution and are considered to have functional value. Correspondingly, the extremely highly conserved 16S rRNA is associated with four GO annotations that are involved in two major GO classifications: biological processes (cellular component organization and protein metabolic process) and cellular components (mitochondria, non-membrane-bounded organelle, and organelle lumen).

Associated GO terms and gene networks for genes overlapping complementary SNP probes are all involved in various biological functions with a strong association to developmental processes, identical to results of the cross-amplification study carried out by Hoffman *et al.*[75]. The results of this gene analysis are expected based on the fact that highly conserved sequences are typically enriched in fundamental cell processes[75].

Since selection acts on phenotypes arising from the variation found in genes, it is informative to examine the predicted effects of SNPs since these effects may have functional relevance to the naked mole-rat. Coding regions of genes containing SNPs that have nonsynonymous effects are of interest because the SNP allele itself can change the sequence and lead to a change in the protein product of the gene. Non-synonymous SNPs are believed to have the highest impact on phenotype because of the potential for change in biological function[129].

4.3.4 A measure of SNP diversity is limited for this study because sample size is too low

A major limitation of this study to detect polymorphic SNP loci is the fact that there are currently only two naked mole-rat builds for which entire genomes have been sequenced. An adequate sample size is required to be able to detect polymorphic sites across the genomes of individuals. A higher sample size is necessary to be able to measure allele frequencies for the study of genetic diversity in a population. SNPs are generally assigned a minor allele frequency which refers to the frequency at which the second most common SNP allele occurs in a population. For example, the HapMap project database contains

SNPs with a minor allele frequency of 0.05 or greater only. It is not possible to determine the presence of alleles at particular SNP sites in the naked mole-rat because a sample size of two makes it impossible for a minor allele frequency of 0.05 or greater to be calculated for each polymorphic site[130]. To examine SNP diversity, or the extent of polymorphism within and between naked mole-rat populations, a greater sample size and panel of SNP markers is necessary.

A caveat to the study of SNP diversity in naked mole-rat colonies is the genetic structure and population of the naked mole-rat. Sequencing studies of the naked mole-rat genomes determined a significantly lower number of SNPs than in mouse and rat populations and estimated nucleotide diversity (mean per nucleotide heterozygosity) in naked mole-rat is lower (similar to humans)[84]. This is indicative of both a relatively small sample size and a high level of inbreeding that is consistent with low SNP diversity[84,97]. High levels of observed genetic similarity within wild naked mole-rat colonies does in fact support the close relatedness through inbreeding that is typically prevalent within the eusocial naked mole-rat colonies and supports the fact that naked mole-rat colonies are known to have the highest inbreeding coefficient for wild mammals[131,132]. Thus, it makes sense that both a small sample size and an inbred population would result in low SNP diversity and my results are consistent with previous studies on diversity[84,97,133].

An issue with cross-species hybridization studies for SNP discovery is ascertainment bias. Since known SNPs are used for SNP probe design, an overrepresentation of common SNPs is expected for all microarray studies. The consequences of this type of bias are underrepresentation of less common SNP allele frequencies and unequal coverage of SNPs across the genome. Furthermore, since natural selection acts on SNPs that are not neutral, cross-hybridizing SNPs in the naked mole-rat are likely biased for neutral SNPs that may not have functional or evolutionary value[134]. SNP discovery is also biased when sample size is small. Typically a minimum sample size is required so that the sample allele frequencies are good estimates of the true population allele frequencies, including less common SNPs[135]. A previous study determined that both sample size and the genomic region surveyed directly influence SNP features and population genetic

estimates and so it is important to consider that polymorphism at specific loci is undetectable without a high enough sample size[136]. An empirical study carried out at the population level would be useful in determining the frequency of an allele at polymorphic sites.

A higher frequency of type I errors, or false positives, is typically observed when there is a small sample of polymorphic loci. In microarray cross-species amplification studies, a few thousand random SNPs have been demonstrated to be sufficient to reliably estimate SNP diversity in populations[137]. Since no SNPs were found through blasting of SNP probe sequences, either a much higher sample size is required or a much larger panel of unbiased SNP markers need to be used. A large enough number of loci on the initial array may allow for the generation of a panel of markers that would otherwise not exist despite many years of divergence[73].

A large number of Mus probe targets (about 17,000) have been identified to potentially cross-hybridize to single-copy target sequences in two naked mole-rat genomes and can be used for molecular typing. Molecular typing is typically used to identify different types (i.e., determine relatedness) of organisms within a species[138,139]. Molecular typing using the cross-hybridizing MDGA probes can be defined as a call for hybridization or a no call where there is no hybridization at each of the previously determined loci (about 650 SNP probe loci and just under 17,000 IGP loci). The cross-hybridizing MDGA probes have potential for use in genotyping DNA samples from different colonies and subspecies of naked mole-rats whereby a genotype for each naked mole-rat would be generated based on sample-specific patterns of hybridization (i.e., a call or no call at each of the probe sites). Molecular typing of naked mole-rats can be used to assess genetic variation in samples within a colony or geographic location and between colonies or geographic locations.

Considerations for future cross-species hybridization studies

The generated probe list of stringent, complementary, and unique probes have potential to interrogate specific evolutionarily conserved loci in experimental naked mole-rat DNA. This is not of interest for CNV calling but for comparative genomics studies. The cross-

species application of probes to discover conserved sequences of various types (i.e., genes, exons, CNVs, SNPs) is not novel. A 2006 study identified conserved genetic sequences and genes that were all involved in early development across three evolutionary distant species[140]. The first primate sequences and also putative sites of CNV were identified in various great apes through the use of probe sequences designed on the human genome[79–83]. Cross-species application of microarrays have become increasingly useful over the years for identifying conserved sequences, potential functions of these sequences, and sequence variation[18,67,71,74–78]. Cross-species hybridization of the MDGA may be a first step to generating microarray data on experimental naked mole-rat DNA for population genetics and comparative genomics studies, though results are expected to be limited. I recommend cross-species application of the MDGA to more closely related species, particularly within the genus Mus.

While cross-species hybridization has been applied to closely related species, the approach has not been applied to divergent species largely due to a lack of appropriate probes and a lack of standard hybridization conditions needed for hybridization success that can increase rates of error. With increasing evolutionary distance, structural rearrangements or shuffling across the genome can make probes on an array unsuitable for targeting variation across species. A low number of complementary SNP probes (and a drastic decrease in polymorphic SNPs with increasing years of divergence) as well as the requirement of a minimum of three consecutive probes in the correct order for CNV calling (with increased structural rearrangements and shuffling with evolutionary divergence) means that MDGA probes may not be useful in distantly related species.

4.4 Conclusions

Bioinformatics methods can allow for the examination of sources of error that lead to false discovery. First, the MDGA probe filtering results contribute to a published list of stringent probes that has been demonstrated to reduce false positives and false negatives in array data[1–3]. Second, a post-hoc assessment for array CNV calls can be a relatively quick and low-cost method to increase confidence in the effective preparation of DNA from different tissue and cell types; however, the results of this analysis need to be tested

empirically within the same biological samples used for CNV analysis. Lastly, MDGA probe alignment to the naked mole-rat serves as a preliminary analysis of the ability of MDGA SNP probes for measuring genetic diversity and IGPs for exploring exon and gene conservation. The theoretical utility of the MDGA for application to the naked mole-rat samples can now be validated empirically. Specifically, the ability of the MDGA to type, or distinguish naked mole-rats across colonies and populations can be tested.

References

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

2. Fadista, J., and Bendixen, C. (2012). Genomic position mapping discrepancies of commercial SNP chips. *PLoS ONE 7*, 1–5.

3. Eitutis, S. (2013). 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. Unpublished.

4. van Heesch, S., Mokry, M., Boskova, V., Junker, W., Mehon, R., Toonen, P., de Bruijn, E., Shull, J.D., Aitman, T.J., Cuppen, E., et al. (2013). Systematic biases in DNA copy number originate from isolation procedures. *Genome Biology 14*, R33.

5. The Encode Project Consortium (2013). An integrated encyclopedia of DNA elements in the human genome. *Nature 489*, 57–74.

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

7. The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature 437*, 1299–1320.

8. Lindblad-Toh, K., Winchester, E., Daly, M.J., Wang, D.G., Hirschhorn, J.N., Laviolette, J.P., Ardlie, K., Reich, D.E., Robinson, E., Sklar, P., et al. (2000). Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genetics 24*, 381–386.

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

10. Leaché, A.D., and Oaks, J.R. (2017). The utility of single nucleotide polymorphism (SNP) data in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics 48*, 69–84.

11. Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics 36*, 949–951.

12. MacDonald, J.R., Ziman, R., Yuen, R.K.C., Feuk, L., and Scherer, S.W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research 42*, 986–992.

13. Ohta, T. (1988). Time for acquiring a new gene by duplication. *Proceedings of the National Academy of Sciences 85*, 3509–3512.

14. Ohno, S. (1970). Evolution by gene duplication (New York: Springer-Verlag). 160 pp.

15. Dumas, L., Kim, Y.H., Karimpour-Fard, A., Cox, M., Hopkins, J., Pollack, J.R., and Sikela, J.M. (2007). Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Research 17*, 1266–1277.

16. Bailey, J. a, and Eichler, E.E. (2006). Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature Reviews 7*, 552–564.

17. Tuzun, E., Bailey, J., and Eichler, E. (2004). Recent segmental duplications in the working draft assembly of the Brown Norway Rat. *Genome Research 14*, 493–506.

18. Marques-bonet, T., Kidd, J.M., Ventura, M., Graves, T.A., Cheng, Z., Hillier, L.W., Jiang, Z., Baker, C., Malfavon-borja, R., A, L., et al. (2009). A burst of segmental duplications in the African Great Ape ancestor. *Nature 457*, 877–881.

19. Yoshiki, A., and Moriwaki, K. (2006). Mouse phenome research: implications of genetic background. *Ilar Journal 47*, 94–102.

20. The Jackson Laboratory (2018). Mouse Database. at < http://www.informatics.jax.org/inbred_strains/mouse/STRAINS.shtml>

21. Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature 420*, 520–562.

22. Phifer-Rixey, M., and Nachman, M.W. (2015). Insights into mammalian biology from the wild house mouse *Mus musculus*. *ELife 4*, 1–13.

23. Brudno, M., Poliakov, A., Salamov, A., Cooper, G.M., Sidow, A., Rubin, E.M., Solovyev, V., Batzoglou, S., and Dubchak, I. (2004). Automated whole-genome multiple alignment of rat, mouse, and human. *Genome Research 14*, 685–692.

24. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. *Nature 444*, 444–454.

25. Agam, A., Yalcin, B., Bhomra, A., Cubin, M., Webber, C., Holmes, C., Flint, J., and Mott, R. (2010). Elusive copy number variation in the mouse genome. *PloS ONE 5*, e12839.

26. Beliveau, B.J., Joyce, E.F., Apostolopoulos, N., Yilmaz, F., Fonseka, C.Y., McCole, R.B., Chang, Y., Li, J.B., Senaratne, T.N., Williams, B.R., et al. (2012). Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *Proceedings of the National Academy of Sciences 109*, 21301–21306.

27. Sanger, F., Nicklen, S., and Coulson, a R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences 74*, 5463–5467.

28. Altshuler, D., Lander, E., and Ambrogio, L. (2010). A map of human genome variation from population scale sequencing. *Nature 476*, 1061–1073.

29. Shen, W., Paxton, C.N., Szankasi, P., Longhurst, M., Schumacher, J.A., Frizzell, K.A., Sorrells, S.M., Clayton, A.L., Jattani, R.P., Patel, J.L., et al. (2018). Detection of genome-wide copy number variants in myeloid malignancies using next-generation sequencing. *Journal of Clinical Pathology 71*, 372-378.

30. Eckel-Passow, J.E., Atkinson, E.J., Maharjan, S., Kardia, S.L.R., and de Andrade, M. (2011). Software comparison for evaluating genomic copy number variation for Affymetrix 6.0 SNP array platform. *BMC Bioinformatics 12*, 220.

31. Xu, L., Hou, Y., Bickhart, D., Song, J., and Liu, G. (2013). Comparative analysis of CNV calling algorithms: literature survey and a case study using bovine high-density SNP data. *Microarrays 2*, 171–185.

32. Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., Lionel, A.C., Thiruvahindrapuram, B., Macdonald, J.R., Mills, R., et al. (2011). Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature Biotechnology 29*, 512–520.

33. Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., et al. (2012). Origins and functional impact of copy number variation in the human genome. *Nature 464*, 704–712.

34. Fadista, J., Thomsen, B., Holm, L.-E., and Bendixen, C. (2010). Copy number variation in the bovine genome. *BMC Genomics 11*, 284.

35. Berglund, J., Nevalainen, E.M., Molin, A.-M.M., Perloski, M., André, C., Zody, M.C., Sharpe, T., Hitte, C., Lindblad-Toh, K., Lohi, H., et al. (2012). Novel origins of copy number variation in the dog genome. *Genome Biology 13*, R73.

36. Wetmur, J.G. (1991). DNA probes: applications of the principles of nucleic acid hybridization. *Critical Reviews in Biochemistry and Molecular Biology 26*, 227–259.

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

38. Southern, E.M., Maskos, U., and Elder, J.K. (1992). Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. *Genomics 13*, 1008–1017.

39. Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., and Fodor, S.P.A. (1996). Accessing genetic information with high-density DNA arrays. *Science 274*, 610–614.

40. Sakharkar, M.K., Perumal, B.S., Sakharkar, K.R., and Kangueane, P. (2005). An analysis on gene architecture in human and mouse genomes. *In Silico Biology 5*, 347–365.

41. Guénet, J.L. (2005). The mouse genome. *Genome Research 15*, 1729–1740.

42. Zhang, D., Chen, S., and Yin, P. (2014). Optimizing the specificity of nucleic acid hybridization. *Nature Chemistry 4*, 208–214.

43. Affymetrix Inc (2007). BRLMM-P : a Genotype Calling Method for the SNP 5.0 Array. at <https://tools.thermofisher.com/content/sfs/brochures/brlmmp_whitepaper.pdf>

44. Haraksingh, R.R., Abyzov, A., and Urban, A.E. (2017). Comprehensive performance comparison of high-resolution array platforms for genome-wide copy number variation (CNV) analysis in humans. *BMC Genomics 18*, 321.

45. Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F. a, Hakonarson, H., and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research 17*, 1665–1674.

46. Broude, N.E. (2002). Stem-loop oligonucleotides: a robust tool for molecular biology and biotechnology. *Trends in Biotechnology 20*, 249–256.

47. Darby, B.J., Jones, K.L., Wheeler, D., and Herman, M.A. (2011). Normalization and centering of array-based heterologous genome hybridization based on divergent control probes. *BMC Bioinformatics 12*, 183.

48. Bar-Or, C., Czosnek, H., and Koltai, H. (2007). Cross-species microarray hybridizations: a developing tool for studying species diversity. *Trends in Genetics 23*, 200–207.

49. Saiki, R., Gelfand, D., Stoffel, S., Scharf, S., Higuchi, R., Horn, G., KB, M., and Erlich, H. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science 239*, 487–491.

50. Dhami, P., Coffey, A.J., Abbs, S., Vermeesch, J.R., Dumanski, J.P., Woodward, K.J., Andrews, R.M., Langford, C., and Vetrie, D. (2005). Exon array CGH: detection of copy-number changes at the resolution of individual exons in the human genome. *The American Journal of Human Genetics 76*, 750–762.

51. Hesselberth, J.R., Chen, X., Zhang, Z., Sabo, P.J., Sandstrom, R., Reynolds, A.P., Thurman, R.E., Neph, S., Kuehn, M.S., Noble, W.S., et al. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods 6*, 283–289.

52. Gross, D.S., and Garrard, W.T. (1988). Nuclease hypersensitive sites in chromatin. *Annual Review of Biochemistry 57*, 159–197.

53. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. *Nature 489*, 75–82.

54. Illingworth, R., Kerr, A., DeSousa, D., Jørgensen, H., Ellis, P., Stalker, J., Jackson, D., Clee, C., Plumb, R., Rogers, J., et al. (2008). A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biology 6*, 0037–0051.

55. Vierstra, J., Rynes, E., Sandstrom, R., Zhang, M., Canfield, T., Hansen, R.S., Stehling-Sun, S., Sabo, P.J., Byron, R., Humbert, R., et al. (2014). Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science 346*, 1007–1012.

56. Graff, J., Kim, D., Dobbin, M.M., and Tsai, L. (2011). Epigenetic regulation of gene expression in physiological and pathological brain processes. *Physiological Reviews 91*, 603–649.

57. Bonasio, R., Agrawal, A.A., Laforsch, C., Tollrian, R., Allfrey, V.G., Faulkner, R., Mirsky, A.E., Allis, C.D., Jenuwein, T., Reinberg, D., et al. (2015). The expanding epigenetic landscape of non-model organisms. *Journal of Experimental Biology 218*, 60–63.

58. Mo, A., Mukamel, E.A., Davis, F.P., Luo, C., Henry, G.L., Picard, S., Urich, M.A., Nery, J.R., Sejnowski, T.J., Lister, R., et al. (2015). Epigenomic signatures of neuronal diversity in the mammalian brain. *Neuron 86*, 1369–1384.

59. Felsenfeld, G., Boyes, J., Chung, J., Clark, D., and Studitsky, V. (1996). Chromatin structure and gene expression. *Proceedings of the National Academy of Sciences 93*, 9384–9388.

60. Van Holde, K.E. (1989). Chromatin. *Journal of Molecular Regconition 19*, 227–228.

61. Cedar, H., and Bergman, Y. (2012). Programming of DNA methylation patterns. *Annual Reviews Biochemistry 81*, 97–117.

62. Wu, J., Huang, B., Chen, H., Yin, Q., Liu, Y., Xiang, Y., Zhang, B., Liu, B., Wang, Q., Xia, W., et al. (2016). The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature 534*, 652–657.

63. Wykes, S.M., and Krawetz, S.A. (2003). The structural organization of sperm chromatin. *Journal of Biological Chemistry 278*, 29471–29477.

64. Morell, P., and Quarles, R. (1999). Characteristic composition of myelin (Philadelphia: Lippincott-Raven). at < https://www.ncbi.nlm.nih.gov/books/NBK28221/>

65. Promega Inc (2017). Proteinase K. at <https://www.promega.com/-/media/files/resources/protocols/product-information-sheets/n/proteinase-k-protocol.pdf>

66. The Mouse ENCODE Consortium (2014). A comparative Encyclopedia of DNA Elements in the Mouse Genome. *Nature 515*, 355–364.

67. Miller, J.M., Poissant, J., Kijas, J.W., and Coltman, D.W. (2011). A genome-wide set of SNPs detects population substructure and long range linkage disequilibrium in wild sheep. *Molecular Ecology Resources 11*, 314–322.

68. Skinner, B.M., Al Mutery, A., Smith, D., Vollker, M., Hojjat, N., Raja, S., Trim, S., Houde, P., Boecklen, W.J., and Griffin, D.K. (2014). Global patterns of apparent copy number variation in birds revealed by cross-species comparative genomic hybridization. *Chromosome Research 22*, 59–70.

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

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

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

diversity, and phylogeny studies. *PLoS Genetics 8*, e10002451.

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

73. Haynes, G.D., and Latch, E.K. (2012). Identification of novel single nucleotide polymorphisms (SNPs) in deer (*odocoileus* spp.) using the BovineSNP50 Beadchip. *PLoS ONE 7*, e36536.

74. Decker, J.E., Pires, J.C., Conant, G.C., McKay, S.D., Heaton, M.P., Chen, K., Cooper, A., Vilkki, J., Seabury, C.M., Caetano, A.R., et al. (2009). Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. *Proceedings of the National Academy of Sciences 106*, 18644–18649.

75. Hoffman, J.I., Thorne, M.A.S., McEwing, R., Forcada, J., and Ogden, R. (2013). Cross-amplification and validation of SNPs conserved over 44 million years between seals and dogs. *PLoS ONE 8*, 1–10.

76. Fontanesi, L., Martelli, P.L., Beretti, F., Riggio, V., Dall'Olio, S., Colombo, M., Casadio, R., Russo, V., and Portolano, B. (2010). An initial comparative map of copy number variations in the goat (*Capra hircus*) genome. *BMC Genomics 11*, 639.

77. Fontanesi, L., Beretti, F., Martelli, P.L., Colombo, M., Dall'olio, S., Occidente, M., Portolano, B., Casadio, R., Matassino, D., and Russo, V. (2011). A first comparative map of copy number variations in the sheep genome. *Genomics 97*, 158–165.

78. Griffin, D.K., Robertson, L.B., Tempest, H.G., Vignal, A., Fillon, V., Crooijmans, R.P.M.A., Groenen, M.A.M., Deryusheva, S., Gaginskaya, E., Carré, W., et al. (2008). Whole genome comparative studies between chicken and turkey and their implications for avian genome evolution. *BMC Genomics 9*, 168.

79. Gazave, E., Darré, F., Morcillo-suarez, C., Ventura, M., Catacchio, C.R., Alkan, C., Ebeling, M., Küng, E., See, A., Petit-marty, N., et al. (2011). Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. *Genome Research 21*, 1626–1639.

80. Locke, D.P., Eichler, E.E., Segraves, R., Carbone, L., Archidiacono, N., and Albertson, D.G. (2003). Large-scale variation among human and great apes determined by array comparative genomic hybridization. *Genome Research 13*, 347–357.

81. Perry, G.H., Yang, F., Marques-Bonet, T., Murphy, C., Fitzgerald, T., Lee, A.S., Hyland, C., Stone, A.C., Hurles, M.E., Tyler-Smith, C., et al. (2008). Copy number variation and evolution in humans and chimpanzees. *Genome Research 18*, 1698–1710.

82. Perry, G.H., Tchinda, J., McGrath, S.D., Zhang, J., Picker, S.R., Cáceres, A.M., Iafrate, a J., Tyler-Smith, C., Scherer, S.W., Eichler, E.E., et al. (2006). Hotspots for

copy number variation in chimpanzees and humans. *Proceedings of the National Academy of Sciences 103*, 8006–8011.

83. Lee, A.S., Gutiérrez-Arcelus, M., Perry, G.H., Vallender, E.J., Johnson, W.E., Miller, G.M., Korbel, J.O., and Lee, C. (2008). Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. *Human Molecular Genetics 17*, 1127–1136.

84. Kim, E.B., Fang, X., Fushan, A.A., Huang, Z., Lobanov, A. V, Han, L., Marino, S.M., Sun, X., Turanov, A.A., Yang, P., et al. (2011). Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature 479*, 223–227.

85. Buffenstein, R. (2005). The naked mole-rat: a new long-living model for human aging research. *Journal of Gerontology 60A*, 1369–1377.

86. Buffenstein, R. (2008). Negligible senescence in the longest living rodent, the naked mole-rat: insights from a successfully aging species. *Journal of Comparative Physiology 178*, 439–445.

87. Edrey, Y.H., Hanes, M., Pinto, M., Mele, J., and Buffenstein, R. (2008). Succesful aging and sustained good health in the naked mole rat: a long-lived mammalian model for biogerontology and biomedical research. *Advances in Experimental Medicine and Biology 644*, 1–5.

88. Ikeno, Y., Hubbard, G.B., Lee, S., Richardson, A., Strong, R., Diaz, V., and Nelson, J.F. (2005). Housing density does not influence the longevity effect of calorie restriction. *Journal of Gerontology 60A*, 1510–1517.

89. Liang, S., Mele, J., Wu, Y., Buffenstein, R., and Hornsby, P.J. (2010). Resistance to experimental tumorigenesis in cells of a long-lived mammal, the naked mole-rat (*Heterocephalus glaber*). *Aging Cell 9*, 626–635.

90. Jarvis, J. (1981). Eusociality in a mammal: cooperative breeding in naked mole-rat colonies. *Science 212*, 571–573.

91. MacRae, S.L., Zhang, Q., Lemetre, C., Seim, I., Calder, R.B., Hoeijmakers, J., Suh, Y., Gladyshev, V.N., Seluanov, A., Gorbunova, V., et al. (2015). Comparative analysis of genome maintenance genes in naked mole rat, mouse, and human. *Aging Cell 14*, 288–291.

92. Bens, M., Sahm, A., Groth, M., Jahn, N., Morhart, M., Holtze, S., Hildebrandt, T.B., Platzer, M., and Szafranski, K. (2016). FRAMA : from RNA-seq data to annotated mRNA assemblies. *BMC Genomics 17*, 1–12.

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

94. Milojevic, M. Thesis in preparation. Unpublished.

95. Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., et al. (2012). BEDOPS: High-performance genomic feature operations. *Bioinformatics 28*, 1919–1920.

96. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics 26*, 841–842.

97. Keane, M., Craig, T., Alföldi, J., Berlin, A.M., Johnson, J., Seluanov, A., Gorbunova, V., Di Palma, F., Lindblad-Toh, K., Church, G.M., et al. (2014). The Naked Mole Rat Genome Resource: facilitating analyses of cancer and longevity-related adaptations. *Bioinformatics 30*, 3558–3560.

98. Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols 4*, 44–57.

99. Vanneste, E., Voet, T., Le Caignec, C., Ampe, M., Konings, P., Melotte, C., Debrock, S., Amyere, M., Vikkula, M., Schuit, F., et al. (2009). Chromosome instability is common in human cleavage-stage embryos. *Nature Medicine 15*, 577–583.

100. Wilson, V., and Beddington, R.S.P. (1996). Cell fate and morphogenetic movement in the late mouse primitive streak. *Mechanisms of Development 55*, 79–89.

101. Frank, S.A. (2010). Evolution in health and medicine Sackler colloquium: Somatic evolutionary genomics: mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration. *Proceedings of the National Academy of Sciences 107 Suppl*, 1725–1730.

102. Wu, X., and Sharp, P.A. (2013). XDivergent transcription: A driving force for new gene origination? *Cell 155*, 990–996.

103. Natarajan, A., Yardimci, G.G., Sheffield, N.C., and Frazer, K. a (2012). Predicting cell-type − specific gene expression from regions of open chromatin the genome. *Genome Research 22*, 1711–1722.

104. Wilken, M.S., Brzezinski, J.A., La Torre, A., Siebenthall, K., Thurman, R., Sabo, P., Sandstrom, R.S., Vierstra, J., Canfield, T.K., Hansen, R.S., et al. (2015). DNase I hypersensitivity analysis of the mouse brain and retina identifies region-specific regulatory elements. *Epigenetics and Chromatin 8*, 1–17.

105. Winchester, L., Yau, C., and Ragoussis, J. (2009). Comparing CNV detection methods for SNP arrays. *Briefings in Functional Genomics & Proteomics 8*, 353–366.

106. Zhang, X., Du, R., Li, S., Zhang, F., Jin, L., and Wang, H. (2014). Evaluation of copy number variation detection for a SNP array platform. *BMC Bioinformatics 15*, 50.

107. Wishart, A.E. (2014). Somatic copy number mosaicism contributes to genomic diversity in *Mus musculus*. Unpublished.

108. Li, W., and Olivier, M. (2013). Current analysis platforms and methods for detecting copy number variation. *Physiological Genomics 45*, 1–16.

109. Katju, V., and Bergthorsson, U. (2013). Copy-number changes in evolution: Rates, fitness effects and adaptive significance. *Frontiers in Genetics 4*, 1–12.

110. Kondrashov, F.A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings of the Royal Society B: Biological Sciences 279*, 5048–5057.

111. Bergthorsson, U., Andersson, D.I., and Roth, J.R. (2007). Ohno's dilemma: evolution of new genes under continuous selection. *Proceedings of the National Academy of Sciences 104*, 17004–17009.

112. Sharp, A.J., Locke, D.P., McGrath, S.D., Cheng, Z., Bailey, J.A., Vallente, R.U., Pertz, L.M., Clark, R.A., Schwartz, S., Segraves, R., et al. (2005). Segmental duplications and copy-number variation in the human genome. *The American Journal of Human Genetics 77*, 78–88.

113. Pearson, W.R. (2013). An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics 42*, 3.1.1-3.1.8.

114. Loewe, L., and Hill, W.G. (2010). The population genetics of mutations: good, bad and indifferent. *Philosophical Transactions of the Royal Society of Biological Sciences 365*, 1153–1167.

115. Zhao, S., Shetty, J., Hou, L., Delcher, A., Zhu, B., Osoegawa, K., de Jong, P., Nierman, W.C., Strausberg, R.L., and Fraser, C.M. (2004). Human, mouse, and rat genome large-scale rearrangements: stability versus speciation. *Genome Research 14*, 1851–1860.

116. Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. (2003). Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences 100*, 11484–11489.

117. Pevzner, P., and Tesler, G. (2003). Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proceedings of the National Academy of Sciences 100*, 7672–7677.

118. Pevzner, P., and Tesler, G. (2003). Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Research 13*, 37–45.

119. Mural, R.J., Adams, M.D., Myers, E.W., Smith, H.O., Miklos, G.L.G., Wides, R., Halpern, A., Li, P.W., Sutton, G., Nadeau, J., et al. (2002). A comparison of whole-

genome shotgun-derived mouse chromosome 16 and the human genome. *Science 296*, 1661–1671.

120. Hurst, L.D., Pál, C., and Lercher, M.J. (2004). The evolutionary dynamics of eukaryotic gene order. *Nature Reviews Genetics 5*, 299–310.

121. Dávila López, M., Marťne Guerra, J.J., and Samuelsson, T. (2010). Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. *PLoS ONE 5*, e10654.

122. Lercher, M.J., Urrutia, A.O., and Hurst, L.D. (2002). Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature Genetics 31*, 180–183.

123. Bennett, N.C., and Faulkes, C.G. (2000). African mole-rats: ecology and eusociality. (Cambridge: Cambridge University Press). 273 pp.

124. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. *Science 304*, 1321–1325.

125. Betts, M.J., Guigó, R., Agarwal, P., and Russell, R.B. (2001). Exon structure conservation despite low sequence similarity: A relic of dramatic events in evolution? *EMBO Journal 20*, 5354–5360.

126. Roy, S.W., Fedorov, A., and Gilbert, W. (2003). Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proceedings of the National Academy of Sciences 100*, 7158–7162.

127. Chernikova, D., Managadze, D., Glazko, G., Makalowski, W., and Rogozin, I. (2016). Conservation of the exon-intron structure oflong intergenic non-coding RNA genes in eutherian mammals. *Life 6*, 27.

128. Isenbarger, T.A., Carr, C.E., Johnson, S.S., Finney, M., Church, G.M., Gilbert, W., Zuber, M.T., and Ruvkun, G. (2008). The most conserved genome segments for life detection on earth and other planets. *Origins of Life and Evolution of Biospheres 38*, 517–533.

129. Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Research 30*, 3894–3900.

130. Visscher, P.M. (2007). Variation of estimates of SNP and haplotype diversity and linkage disequilibrium in samples from the same population due to experimental and evolutionary sample size. *Annals of Human Genetics 71*, 119–126.

131. Faulkes, C.G., Abbott, D.H., and Mellor, A.L. (1990). Investigation of genetic diversity in wild colonies of naked mole-rats (*Heterocephalus gaber*) by DNA

fingerprinting. *Journal of Zoology 221*, 87–97.

132. Reeve, H.K., Westneat, D.F., Noon, W.A., Sherman, P.W., and Aquadro, C.F. (1990). DNA "fingerprinting" reveals high levels of inbreeding in colonies of the eusocial naked mole-rat. *Proceedings of the National Academy of Sciences 87*, 2496–2500.

133. Laurie, C.C., Laurie, C.A., Rice, K., Doheny, K.F., Zelnick, L.R., McHugh, C.P., Ling, H., Hetrick, K.N., Pugh, E.W., Amos, C., et al. (2012). Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nature Genetics 44*, 642–650.

134. Lachance, J., and Tishkoff, S.A. (2013). SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *Bioessays 35*, 780–786.

135. B-Rao, C. (2001). Sample size considerations in genetic polymorphism studies. *Human Heredity 52*, 191–200.

136. Trask, J.A.S., Malhi, R.S., Kanthaswamy, S., Johnson, J., Garnica, W.T., Malladi, V.S., and Smith, D.G. (2011). The effect of SNP discovery method and sample size on estimation of population genetic data for Chinese and Indian rhesus macaques (*Macaca mulatta*). *Primates 52*, 129–138.

137. Fischer, M.C., Rellstab, C., Leuzinger, M., Roumet, M., Gugerli, F., Shimizu, K.K., Holderegger, R., and Widmer, A. (2017). Estimating genomic diversity and population differentiation – an empirical comparison of microsatellite and SNP variation in *Arabidopsis halleri*. *BMC Genomics 18*, 69.

138. Donelli, G., Vuotto, C., and Mastromarino, P. (2013). Phenotyping and genotyping are both essential to identify and classify a probiotic microorganism. *Microbial Ecology in Health & Disease 24*, 20105.

139. Nielsen, E.V.A.M., Engberg, J., Fussing, V., Petersen, L., Brogren, C., and On, S.L.W. (2000). Evaluation of phenotypic and genotypic methods for subtyping *Campylobacter jejuni* isolates from humans, poultry, and cattle. *Journal of Clinical Microbiology 38*, 3800–3810.

140. Vallée, M., Robert, C., Méthot, S., Palin, M.-F., and Sirard, M.-A. (2006). Cross-species hybridizations on a multi-species cDNA microarray to identify evolutionarily conserved genes expressed in oocytes. *BMC Genomics 7*, 113.

Curriculum Vitae

Name: Nisha Patel

Post-secondary Education and Degrees:
Western University
London, Ontario, Canada
2012-2018 M.Sc.

University of Toronto
Toronto, Ontario, Canada
2007-2012 B.Sc.

Related Work Experience
Teaching Assistant
Western University
2012-2015

Publications:

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