Evaluation of two mouse models of high genetic variation for suitability to test a heterozygote instability hypothesis

Hailie Pavanel, The University of Western Ontario

Supervisor: Hill, Kathleen A, The University of Western Ontario
Co-Supervisor: Turley, Eva, The University of Western Ontario
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Biology
© Hailie Pavanel 2021

Follow this and additional works at: https://ir.lib.uwo.ca/etd

Part of the Computational Biology Commons, Genetics Commons, and the Genomics Commons

Recommended Citation

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlsadmin@uwo.ca.
Abstract
Characterization of genetic variation underlying complex phenotypes is incomplete yet critical to understanding mutational mechanisms and phenotypes. Heterozygote Instability (HI) is a new, poorly understood source of mutations needing models for mechanistic study. Two models ideal for characterizing HI-associated mutational mechanisms are outbred mice and mouse basal cell carcinoma (BCC). Both have discontinuous landscapes of heterozygosity essential to assess HI-induced mutations. Here, heterozygosity and copy number variants (CNVs) in two outbred mouse stocks are characterized with 1690, and 3935 autosomal CNVs detected. A positive correlation exists between chromosomal heterozygosity and CNV occurrence ($R^2 = 0.14$ and 0.09), and 41 and 22% of CNVs co-localized with heterozygosity. Genetic variation in human BCC is documented to target characterization in mouse BCC toward filling an identified knowledge gap. An outbred BCC mouse model permits HI hypothesis testing in contexts of meiosis, mitosis, replication, and recombination in gametes, stem cells, and cancer cells.

Keywords: Genetic variants, germline mutations, somatic mutations, outbred mice, basal cell carcinoma, mouse models, Heterozygote Instability, mutation landscape, single nucleotide polymorphisms, copy number variants
Summary for Lay Audience

Mouse models are helpful to study complicated or elusive human genetic phenomena. DNA sequence differences are referred to as genetic variants and include differences in single bases and deletions or duplications of larger regions of sequences, called copy number variants (CNV). Genetic variants contribute to diseases and complex traits that have both genetic and environmental factors. Typically the chromosomes inherited from each parent are not identical, with the inheritance of single-base differences at the same location, referred to as heterozygosity. Recent research discovered that localized regions of heterozygosity increased mutations in plants. However, this phenomenon, known as Heterozygote Instability (HI), has yet to be investigated in mammals. This thesis explores HI in two mouse models of high heterozygosity by analyzing inherited variation in outbred mice and acquired variation in mouse models of skin cancer, basal cell carcinoma (BCC). Researchers have bred outbred mice to mimic the high genetic diversity found among humans. However, the genetic variants in their DNA are less characterized than the currently favoured laboratory inbred mouse, precluding exploration of mechanisms of mutation associated with heterozygosity. In the outbred mouse stocks, heterozygosity and CNVs must be characterized to study their co-localization and assess if regional variation in heterozygosity is associated with CNVs. This thesis characterized heterozygosity and CNVs in two stocks of outbred mice and detected CNVs associated with regional variation in heterozygosity. Additionally, the high heterozygosity in human skin and human BCC permits the study of mutagenesis in the context of HI. A systematic review of studies researching mutations in human tumors and BCC mouse models highlighted the striking lack of mutations characterized in mouse models of BCC. This disparity limits understanding of BCC mouse models, their traits and their applicability to study HI and complex traits. Experimental designs were presented to assess and detect the extent of genetic variants to determine whether mouse models are good mimics of human BCC.
These mouse models of high genetic variation offer a means to elucidate the mechanisms of mutation that contribute to an elusive and complicated form of genomic instability, HI, relevant to complex human traits, including cancer formation.
Co-Authorship Statement

The work presented in Chapter 2 contains material from a primary research manuscript in preparation for submission to *Genetics*. This publication is co-authored with Dr. Kathleen Allen Hill. I performed all of the bioinformatics approaches presented in this chapter with assistance in applying statistical analyses by Dr. Bin Luo, Dr. Charmaine Dean, and Dr. Reg Kulperger. Further assistance was provided by Nicholas A Boehler, Steven Villani and Dr. Camila de Souza in creating R scripts for automation of the bioinformatics pipeline and statistical analysis.

Sections of Chapter 2 pertaining to the breeding schemes, genetic backgrounds, and genotypes of CD-1 and NMRI mice are contributions to a manuscript in preparation for submission to *Genetics* presenting Machine Learning with Digital Signal Processing (ML-DSP) as a tool to provide classifications for mice based on classifying mouse SNP genotypes. This manuscript is co-authored by Nicholas A Boehler, Dr. Gurjit Singh Randhawa, Rachel D Kelly, myself, Maximillian PM Soltysiak, Dr. Lila Kari, and Dr. Kathleen Allen Hill.

A section of Chapter 2 pertaining to the application of the J statistic is contributed as a case study to a manuscript in preparation for publication in *Bioinformatics* presenting the J statistic as a tool for assessment of the spatial distribution of mutations on the genomic landscape. This manuscript is co-authored by Dr. Bin Luo, Nicholas A Boehler, myself, Freda Qi, Dr. Charmaine Dean, Dr. Kathleen Allen Hill, and Dr. Reg Kulperger.

The systematic literature review presented in Chapter 3 contains material from a mini review manuscript in preparation for publication in *Frontiers in Genetics*. This publication is co-authored with Dr. Kathleen Allen Hill and Dr. Eva Turley. I performed all of the work presented in Chapter 3, with direction and guidance from Dr. Kathleen Allen Hill and Dr. Eva Turley.
Acknowledgements

I am grateful for the support of many people, without them this thesis may have never reached completion. First and foremost, I would like to thank my supervisors, Dr. Kathleen Hill and Dr. Eva Turley, for their unwavering support, guidance, and feedback, which without this thesis would not have been possible. I would like to thank Dr. Kathleen Hill for sharing her unwavering support, patience and insightful critiques that have helped me grow as an individual and researcher. I would like to thank Dr. Eva Turley for her support, assistance and wealth of knowledge as I learned more about cancer biology and BCC. I am inspired daily by both of you and your passion and persistence for research.

I would also like to thank my advisors Dr. Robert Cumming and Dr. Sashko Damjanovski, for their invaluable input throughout this project. Your valuable inputs and insights were crucial to the completion of this thesis.

I would also like to acknowledge and thank a number of people for their expert assistance throughout my research project. Thank you to Dr. Bin Luo, Dr. Camila de Souza, Dr. Conny Toelg and Dr. Gurjit Singh Randhawa for their technical and statistical support. I would also like to thank Rachel Kelly and Nicholas Boehler for creating such a welcoming and enjoyable environment in the lab, their countless research suggestions and coding help, and support throughout all milestones in this degree. Additionally, I would like to thank other past and present Hill and Turley lab members for their reviewing assistance and suggestions throughout presentations, my proposal and my thesis.

Next, I thank my parents for their unwavering support, love, and enthusiasm throughout graduate school. Finally, I would like to thank my friends and brothers for their support, patience, and laughs. I would not have been able to complete this thesis without all of their support, for which I will always be grateful.
# Contents

Abstract ii  
Summary for Lay Audience iii  
Co-Authorship Statement v  
Acknowledgements vi  
List of Tables xii  
List of Figures xiii  
List of Appendices xv  
List of Abbreviations and Nomenclature xvii  

1 Introduction 1  
1.1 Genome variation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1  
1.2 Heterozygote instability . . . . . . . . . . . . . . . . . . . . . . . . . . 2  
1.3 Genetic variants in complex phenotypes . . . . . . . . . . . . . . . . . . 4  
1.3.1 Functionally relevant mutations . . . . . . . . . . . . . . . . . . . . 4  
1.3.2 Single nucleotide variants and polymorphisms (SNVs and SNPs) . 6  
1.3.3 Copy number variants (CNVs) . . . . . . . . . . . . . . . . . . . . . 6  
1.3.4 Driver gene mutations . . . . . . . . . . . . . . . . . . . . . . . . . 7  
1.4 Genetic variation in mouse models of human genetic variation and disease 8
1.5 Missing genetic variation in mouse models .................................. 11
1.6 Methods to characterize missing genetic variation ......................... 12
  1.6.1 Detection of genome-wide genetic variation ............................ 12
  1.6.2 Confirming previously detected genetic variation ...................... 14
1.7 Thesis goal and specific objectives ........................................... 14
References ................................................................................ 19

2 Characterization of genetic variants in discontinuous landscapes of
heterozygosity in two outbred mouse stocks .................................. 25
  2.1 Introduction ............................................................................. 25
    2.1.1 Research goal, central hypothesis, and specific objectives ........ 34
  2.2 Materials and methods ............................................................ 36
    2.2.1 Samples and SNP genotype and CNV calling ........................... 36
    2.2.2 Genetic distance determination and phylogenetic analyses ......... 37
    2.2.3 Analysis of the landscape of heterozygosity at SNP loci .......... 37
    2.2.4 Detection of recurrent CNV regions and unique CNVs ............ 38
    2.2.5 Pathway and GO term characterization of genes overlapped or
        encompassed by recurrent and unique CNVs .......................... 38
    2.2.6 Analysis of spatial association between heterozygous SNP genotypes
        and CNVs .............................................................................. 39
    2.2.7 Statistical Analysis ............................................................. 39
  2.3 Results .................................................................................... 40
    2.3.1 Levels of genomic and chromosomal heterozygosity at SNP loci
        differ in CD-1 and NMRI mice ................................................. 40
    2.3.2 CD-1 and NMRI mice have similar SNP and CNV genetic distance
        values and display relatedness similarly on Neighbour-Joining trees 41
2.3.3 CD-1 mice have a higher average of CNVs per mouse than NMRI mice, with more gains and longer CNVs yet a smaller percentage of CNVs that contain or share an overlapping region with a gene.

2.3.4 Recurrent CNVs were identified on all autosomes and the X chromosome in CD-1 and NMRI mice, with identification of CNVRs unique to one stock and CNVRs shared between the two stocks.

2.3.5 Enrichment of immune-related and metabolic pathways and terms in genes encompassed by or overlapping a CNV and CNVRs in CD-1 and NMRI mice.

2.3.6 Mutation cluster detection and heterozygous SNP genotypes-CNV association.

2.4 Discussion.

2.4.1 Levels of heterozygosity at SNP loci in CD-1 and NMRI mice more closely represent heterozygosity in the human genome.

2.4.2 SNP genetic distance better represents relatedness between mice.

2.4.3 CD-1 and NMRI mice CNV profiles reported here are comparable to reported trends in humans, and in inbred, wildtype and wild-derived mice.

2.4.4 Cataloguing unique and recurrent CNVs and copy number variable regions.

2.4.5 Enrichment of pathways and terms that respond to a changing external environment in genes encompassed by or overlapping a CNV and CNVRs.

2.4.6 Discontinuous clusters of heterozygosity show proximal spatial association with CNVs.

2.4.7 Study limitations.

2.4.8 Future directions.
3 Toward characterization of genetic variation in mouse models of basal cell carcinoma

3.1 Introduction

3.1.1 Genetic variation is high and dynamic in the epidermis

3.1.2 High genetic variation in cancer

3.1.3 BCC as a case for the highest genetic variation

3.1.4 Mouse models of BCC

3.1.5 Similarities and differences between human and mouse skin physiology

3.1.6 Available mouse models of basal cell carcinoma provide the means to analyze HI in rapidly dividing somatic cells

3.1.7 Research goal and specific objectives

3.2 Materials and methods

3.2.1 Systematic literature review

3.2.2 Comparison of reported mutations in human and mouse BCC tumors

3.2.3 Characterization of reported mutations in human and mouse BCC tumors

3.2.4 Experimental design and bioinformatics pipelines

3.3 Results

3.3.1 Summary of study characteristics

3.3.2 Disparity in reported mutation burden in human tumors and mouse models

3.3.3 Analysis of pathway and gene ontology enrichment in reported genes with mutations in BCC tumors

3.3.4 Multiple experimental designs provided for characterization of somatic variation in BCC
3.4 Discussion ......................................................... 110
  3.4.1 Reported mutations in BCC and a disparity between human and mouse models ............................................. 110
  3.4.2 Enriched pathways and terms reflect known pathways involved in human BCC tumorigenesis and should be confirmed in mouse models of BCC ............................................. 113
  3.4.3 Multiple experimental designs provide options to detect genetic variation in mouse models of BCC depending on the scope of the research question and budget ..................................... 114
  3.4.4 With characterization, mouse models of BCC are promising proxies of high dynamic genetic variation to study HI in somatic cells ............................................. 116
  3.4.5 Study limitations .................................................. 117
  3.5 Conclusion .......................................................... 117

References .............................................................. 119

4 Conclusions and future directions 134

References .............................................................. 138

Appendices .............................................................. 139

Curriculum Vitae ......................................................... 143
# List of Tables

| Table 2.1 | SNP genotyping results by outbred mouse stock. | 40 |
| Table 2.2 | Average heterozygosity at SNP loci per autosome. | 41 |
| Table 2.3 | Autosomal CNV calls by mouse stock and copy number state. | 44 |
| Table 2.4 | Number of unique and recurrent CNVs on the autosomes. | 47 |
| Table 2.5 | Number of unique and recurrent CNVs on the X chromosome. | 47 |
| Table 2.6 | Most common recurrent CNVs and CNVRs in CD-1 mice | 53 |
| Table 2.7 | Most common recurrent CNVs and CNVRs in NMRI mice | 54 |
| Table 2.8 | Most common recurrent CNVs and CNVRs in both CD-1 and NMRI mice | 56 |
| Table 3.1 | Studies reporting genic mutations detected in human BCC tumors | 96 |
| Table 3.2 | Studies reporting genic mutations that lead to BCC tumors in mouse models. | 100 |
| Table 3.3 | List of genes with reported mutations in human BCC tumors. | 101 |
| Table 3.4 | Summary of potential methods for mutation discovery and confirmation in mouse models of BCC. | 107 |
## List of Figures

| Figure 2.1 | Breeding history of CD-1 and NMRI mouse stocks. | 29 |
| Figure 2.2 | Methods workflow from sample extraction to data analysis. | 31 |
| Figure 2.3 | Hotspot Detector for Copy Number Variants (HD-CNV) creates a heatmap karyotype of recurrent and unique CNVs. | 32 |
| Figure 2.4 | Rainfall plots portray the landscape of heterozygosity at SNP loci | 33 |
| Figure 2.5 | Rainbow plots display the potential spatial association between heterozygous SNPs and CNVs. | 33 |
| Figure 2.6 | J function is a statistical tool used to analyze the spatial distribution of heterozygous SNPs relative to CNVs. | 34 |
| Figure 2.7 | Unscaled Neighbour-Joining tree for SNP and CNV diversity in the CD-1 stock. | 42 |
| Figure 2.8 | Unscaled Neighbour-Joining tree for SNP and CNV diversity in the NMRI stock. | 43 |
| Figure 2.9 | Distribution of the average number of CNVs on autosomal chromosomes and chromosomal heterozygosity at SNP loci in CD-1 and NMRI mice | 45 |
| Figure 2.10 | Length of CNV gains and losses across the autosomes in both outbred mouse stocks | 46 |
| Figure 2.11 | HD-CNV graph files visualized as a karyotype of recurrent and unique CNV events on the autosomes for both stocks and the X Chromosome in male mouse samples | 48 |
Figure 2.12  Distribution of CNV events across the autosomes and X chromosome  50
Figure 2.13  Statistical analysis of the occurrence of proximal associations of CNVs to heterozygous SNP loci per autosome and autosomal heterozygosity. ................................. 58
Figure 2.14  Visualization of heterozygous SNP genotypes across the autosomes through chromosomal position and distance between consecutive heterozygous SNPs. ................................. 59
Figure 2.15  The discontinuous landscapes of heterozygosity at SNP loci on chromosome 17 in six different mice show the inter-animal and inter-stock differences in landscapes, ideal for studying HI. ................................. 60
Figure 2.16  Visualization of the proximal association of heterozygous SNP genotypes and CNVs on the autosomal chromosomes in CD-1 and NMRI mice ................................. 61
Figure 3.1  Dynamics of genetic variation and structure in the skin. ................................. 84
Figure 3.2  Hedgehog (Hh) pathway: Key oncogenic pathway in BCC ................................. 87
Figure 3.3  Reported mutations in PTCH1. ................................. 101
Figure 3.4  Characterization of functional enrichment in genes mutated in human BCC tumors. ................................. 106
Figure 3.5  Bioinformatics pipelines to analyze and visualize WES and RNA-sequencing data ................................. 110
List of Appendices

Appendix 2A: 378 CD-1 and NMRI outbred mouse CEL files and sample IDs from The Center for Genome Dynamics at The Jackson Laboratory.  139

Appendix 2B: CNV and SNP genetic distance matrices for all pairwise comparisons of 378 sample genotypes.  139

Appendix 2C: Copy number variant landscape plot (CNV-LP) script.  139

Appendix 2D: Autosomal CNVs detected for 378 samples from both outbred mouse stocks.  139

Appendix 2E: X chromosome CNVs detected for male CD-1 and NMRI mice.  139

Appendix 2F: Summary of X chromosome CNV calls in male CD-1 and NMRI mice by copy number state.  139

Appendix 2G: Copy number variant landscape plot (CNV-LP) data for CD-1 and NMRI mice.  139

Appendix 2H: Genes affected by CNVs in CD-1 and NMRI mice.  139

Appendix 2I: GO terms enriched in genes affected by CNVs in CD-1 and NMRI outbred mouse stocks.  139

Appendix 2J: KEGG pathways enriched in genes affected by CNVs in CD-1 and NMRI outbred mouse stocks.  139

Appendix 2K: PANTHER pathways enriched in genes affected by CNVs in CD-1 and NMRI outbred mouse stocks.  139
Appendix 2L: Rainfall, rainbow and J statistic plots depicting heterozygous SNP loci, spatial association between heterozygous SNP loci and CNVs and the statistical significance of the spatial association for each chromosome in every CD-1 and NMRI mouse sample. 140

Appendix 3A: Mutations in genes mutated in human BCC tumors identified by a systematic literature review. 140

Appendix 3B: PANTHER pathways enriched in genes with reported mutations in human BCC identified by the systematic literature review. 140

Appendix 3C: KEGG pathways enriched in genes with reported mutations in human BCC identified by the systematic literature review. 140

Appendix 3D: GO terms enriched in genes with reported mutations in human BCC identified by the systematic literature review. 140

Appendix 3E: Targeted next-generation sequencing summary for comparison to a BCC-specific targeted exome panel. 141
List of Abbreviations and Nomenclature

BAF  B allele frequency
BCC  Basal cell carcinoma
bp   Base pair(s)
CD-1 Caesarean derived-1
CEL  Cell intensity file [file extension]
CN   Copy number
CNV  Copy number variant
CNV-LP Copy number variant landscape plot
CNVR Copy number variable region
DAVID Database for Annotation, Visualization and Integrated Discovery
ddPCR Digital droplet polymerase chain reaction
DNA  Deoxyribonucleic acid
GO   Gene ontology
HD-CNV Hotspot detector for copy number variants
Hh   Hedgehog
IGP  Invariant genome probe
Kb   Kilobase(s)
LRR  Log R ratio
Mb   Megabase(s)
MDGA Mouse Diversity Genotyping Array
NBCCS Nevoid basal cell carcinoma syndrome
Gene nomenclature in mice and humans differ; gene symbols in mice are italicized and written in sentence case (e.g., Pch1), whereas gene symbols in humans are italicized and written with uppercase letters (e.g., PTCH1).

Protein nomenclature in mice and humans is the same; protein symbols are not italicized and written with uppercase letters (e.g., PTCH1).
Chapter 1.

Introduction

1.1 Genome variation

Genetic variants are responsible for differences within and between populations and can contribute to phenotypic diversity. Examples include diversity in gene regulation, complex phenotypes, varied clinical outcomes, varied responses to drugs, and diseases, including cancer, Alzheimer’s disease, and schizophrenia [1–3]. A complex phenotype results from the combined interactions between multiple genetic variants and environmental factors. In comparison, a single gene disorder results from a small number of variants with large effect sizes [4]. A difference in a DNA sequence at a given location or gene between individuals is known as a genetic variant. Genetic variants consist of a combination of polymorphisms and mutations. Polymorphisms are differences between individuals in single nucleotides present in greater than 1% of the population, which may or may not affect phenotype. In contrast, mutations have occurrences below this percentage and can have more profound effects [1, 5].

Mutations can either arise in germ cells that result in inherited mutations in the constitutive zygote genome or arise post-zygotically as acquired mutations in somatic and germ cells. Germline mutations occur in an individual’s gametes, and they are passed on to the next generation. In contrast, somatic mutations are present in other cells of the body and are not inherited [6]. A de novo mutation is the first occurrence of a mutation, and it can either be a germline or somatic mutation, or both, depending on when and where the mutation occurs [6, 7]. If the mutation occurs before fertilization, in either the egg cell or the sperm cell as it is passed on to the offspring, it is a de novo germline
mutation. If the mutation occurs later in development in cells other than the gametes, it is a \textit{de novo} somatic mutation \cite{6, 7}. Mutations can be categorized as point mutations, which are changes at one nucleotide, or structural variants, including insertions, deletions, inversions, duplications, or translocations \cite{6}. To better understand how genetic variants affect complex phenotypes, mutational mechanisms associated with genetic variation must also be defined. The mutation rate is not constant along the chromosomal sequence, and while multiple mechanisms are known, all mechanisms underlying this are not fully understood \cite{8}. Variation in mutation rate can be sequence-dependent, and a novel sequence-dependent mutational mechanism was identified at heterozygous locations across the genome.

\section*{1.2 Heterozygote instability}

Yang \textit{et al} (2015), for the first time, demonstrated a connection between increased heterozygosity and elevated mutation rates. Heterozygosity is defined as the occurrence of two different alleles at the same location between homologous chromosomes, and mutation rate is defined as the number of \textit{de novo} mutations that occur in an individual or gene over time. The connection between heterozygosity and \textit{de novo} mutations highlighted a new potential mechanism contributing to spontaneous mutagenesis. Through investigating single nucleotide polymorphisms (SNPs) and small insertions and deletions in \textit{Arabidopsis} \cite{9}, Yang \textit{et al} (2015) found that during meiosis, plants with high levels of heterozygosity have higher mutation rates than plants with lower levels. The mutations were occurring in close proximity to heterozygous loci and crossover events \cite{9}. This phenomenon was tested in peach plants by crossing either intra-species plants or inter-species plants, and Xie \textit{et al} (2016) found that hybrid plants had higher mutation rates, comparable to findings in \textit{Arabidopsis} \cite{10}. Both studies suggest this phenomenon occurs during meiosis. It is not known if this is germ cell-specific or also operative in somatic cells. HI has not been studied
in situations of high cell division, for example, during development or carcinogenesis. The hypothesis that heterozygosity contributes to \textit{de novo} mutagenesis is called the Heterozygote Instability (HI) hypothesis \cite{11}. HI has been computationally demonstrated with local heterozygosity and microsatellites in 1163 human genome sequences from the 1000 Genome Project \cite{12} but has yet to be explored with insertions and deletions or larger structural variants. There are three tenets of the Heterozygous Instability hypothesis. First, there must be heterozygosity present in the genome. Second, the heterozygosity must be non-randomly distributed throughout the genome, with clusters of heterozygosity separated by clusters of homozygosity. Third, there must be \textit{de novo} copy number variants (CNVs) or other structural variants present in the genome.

The phenomenon of heterozygosity increasing \textit{de novo} mutations observed in \textit{Arabidopsis} and again in peach plants has not yet been reported in mammals. Point mutations, small insertions and deletions, and larger structural variants (duplications and deletions of at least 500 base pairs), such as CNVs occur in the genomes of both plants and mammals \cite{13–16}. Other similar mutational mechanisms exist in plants and mammals, including an increased level of mutations at repetitive DNA sequence and near transposable elements \cite{16}, and UV-induced mutagenesis \cite{14}. Comparable mutations and mechanisms indicate that similar mutational mechanisms, like HI, may occur in mammals and should be further studied to shed light on the mechanisms operating. Additionally, while evidence suggests HI operates during meiosis \cite{9, 10}, no research has been performed to analyze HI during mitosis in somatic cells. Thus, two opportunities to further study HI are in the genome of mammals, looking at both germ and somatic cells.

This thesis evaluates two mouse models of high genetic variation, outbred mice and mouse models of basal cell carcinoma for the study of HI in germ and somatic cells, respectively, and provides novel data and workflows towards the characterization of both models. Characterization of the landscape or spatial distribution of genetic variation in this thesis is defined as the detection, examination and documentation of heterozygous SNP
loci or heterozygosity and CNVs. Characterization of heterozygosity at SNP loci involves detection of the genomic location of heterozygous SNP loci and areas of localized high levels of heterozygosity called clusters, and areas of localized high levels of homozygosity called deserts. Additionally, CNVs are characterized through many attributes, including location, length, copy number state, recurrence in multiple samples, and genic content. After characterizing the SNP landscape and the CNV landscape, the two landscapes can be analyzed simultaneously to see where they occur across the genome and if the occurrence of CNVs is proximal to clusters of heterozygosity.

1.3 Genetic variants in complex phenotypes

1.3.1 Functionally relevant mutations

Genetic variants occur throughout the genome. To better understand complex phenotypes and diseases, research focuses on characterizing and understanding the mutations that have functional implications. Point mutations can be silent, nonsense, or missense mutations, with silent mutations not causing a change in an amino acid due to redundancy in the genetic code. The redundancy means most amino acids are specified by more than one codon, allowing for a nucleotide variation without changing the encoded amino acid. Nonsense mutations replace the original amino acid with a premature stop codon, causing a truncated protein. Missense mutations lead to replacing the original amino acid with a different amino acid, where this can be conserved or non-conserved [1, 5, 6]. Conservative mutations occur when the substituted amino acid has the same properties as the original amino acid, either hydrophobic, hydrophilic, polar, non-polar, or charged. Non-conservative mutations occur when the substituted amino acid does not have the same properties as the original amino acid [6]. Silent and conservative mutations will not usually contribute to a change in phenotype and thus are functionally silent, whereas nonsense, missense, and non-conservative mutations all potentially affect function, either
Structural variants can cause frameshift mutations. Frameshift mutations are due to insertions or deletions of nucleotides that disrupt the reading frame of a gene, consequently changing the amino acids downstream of the mutation [6]. Types of structural variants include duplications, deletions, copy number variants, insertions, transversions and translocations. When structural variants encompass or overlap genes, they may affect gene dosage and thus have possible functional impacts due to either an increase or decrease in the amount of protein produced [17]. Mutation hotspots are genomic locations where mutations occur more frequently than expected in multiple individuals or within specific cancer types, referred to as recurrent mutations [5, 18]. Mutation hotspots are relevant to structural mutations and further study of the varied mutation rate along the chromosomal sequence.

For a mutation to affect a phenotype, termed a functional variant, the mutation must alter the translated protein product’s ability to function correctly, alter the protein RNA’s expression level, or occur in a functional region. [6, 19]. Mutations could occur within the coding region and directly affect the folded protein structure, or affect the regulation of a gene by occurring in gene regulatory regions, resulting in silencing or overexpressing proteins, whether in the wrong tissue or at the wrong time [6, 19]. Regulatory regions could include enhancers, silencers, transcription start sites, splice donor or acceptor sites, ribosome binding sites, termination signals, or polymerase binding sites [19]. Abnormal proteins in cancer can cause uncontrolled growth and cell division or suppress other genes that would function to fix mutations and stop growth. In cancer cells, proto-oncogenes and tumor suppressor genes can be mutated, resulting in uncontrolled growth and proliferation [5, 18]. The mutation landscape is important to understand phenotypic impact because knowing the genomic location of mutations is crucial to understanding possible phenotypic changes. Additionally, the clustering of mutations is important because as the number of mutations increases close to one another, the likelihood of functional impact increases.
1.3.2 Single nucleotide variants and polymorphisms (SNVs and SNPs)

Single nucleotide variants (SNVs) or point mutations are the most common genetic variant occurring throughout the genome. A SNV is a one base pair substitution [1, 20, 21]. Single nucleotide polymorphisms (SNPs) are differences in a single nucleotide between individuals, where the variant must be present in at least more than 1% of the population [20]. SNP refers to the frequency in the population or inheritance, whereas SNV does not [6]. In a diploid organism, genotypes at SNP loci can be homozygous, with the same nucleotide at the SNP locus between a pair of parental chromosomes, or heterozygous, with two different nucleotides at the SNP loci. SNPs can occur in any DNA location [1, 20], and arises through single point mutations, most frequently originating from errors in replication [2]. When SNPs arise within a gene, they can be neutral or can affect phenotype and fitness positively or potentially negatively, as in the case of disease. Additionally, SNPs occurring in regulatory regions of genes can potentially affect gene expression with different fitness effects [21]. SNPs have been used as markers of genetic variation to study contributions of such variation to diseases or complex traits [1]. SNPs contribute to cancer, type II diabetes and other diseases [1, 2].

1.3.3 Copy number variants (CNVs)

Structural variants are large genomic alterations, approximately 1 Kb in size, including insertions, deletions, inversions, duplications and translocations [22]. Copy number variants (CNV) are a subtype of structural variants. CNVs are large (500 bp – 1 Mb) segments of DNA found in variable amounts between individuals and populations [1, 3, 23, 24]. CNVs arise due to deletions or duplications of segments of DNA caused by errors in DNA replication and recombination, during non-allelic homologous recombination, fork stalling and template switching, non-homologous end joining due to double-stranded
breaks, mobile element insertion, and retrotransposition [17, 22, 25, 26]. Polymerases are prone to stalling at regions that are difficult to replicate, including repeat regions and areas of high sequence similarity [17, 25]. This occurrence is referred to as replication stress and can lead to stalled replication forks. The polymerase will switch to another active replication fork to bypass the area, leading to possible misalignment and CNV formation [25]. These errors lead to a gain or loss of genetic information, and CNVs can disrupt gene dosage if they contain a gene [1, 3, 23, 24, 26]. Copy number variable regions (CNVR) are CNVs occurring at the same genomic location in multiple samples, either due to shared inheritance or mutation hotspots [26, 27]. CNVs have been found globally across the genome [27]. Research has shown that CNVs play roles in diseases including Hemophilia A and Thalassemia [24], and in complex traits including autism, Alzheimer’s disease, and schizophrenia [3, 24]. Sequence analysis of CNVs showed genes and regulatory sequences frequently overlapped or encompassed within CNVs, which confirms CNV’s relevance to phenotypic variation and complex phenotypes [26].

1.3.4 Driver gene mutations

Specific mutations can confer advantageous or deleterious functions within a cell or organism. Cells will positively select for or favour the acceptance of mutations that confer a positive advantage, while negatively selecting or favouring the loss of mutations with deleterious effects to the cell [18]. This selection will lead to a subset of mutant cells, or clones, that have a selective advantage over cells without the advantageous mutation. From a health perspective, what phenotypes are advantageous and deleterious to a normal cell differs from that of a cancer cell. In cancer, driver mutations confer a growth advantage, leading to tumorigenesis and proliferation. Driver mutations confer advantageous phenotypes, subject to positive selection and increased fitness [5, 18]. In contrast, passenger mutations confer less or no apparent growth advantage and thus tend to have no positive or negative selective advantage. Driver genes confer a crucial growth
advantage to a cell when mutated, leading to further genomic instability, proliferation, and tumorogenesis [5, 28]. Methods used to differentiate driver and passenger mutations include frequency- and function-based methods [28], or analyzing structural changes in the translated protein associated with mutations under positive selection [5]. Driver genes can be classified as proto-oncogenes or tumor suppressors, depending on whether they are activating or inactivating mutations. When mutated, proto-oncogenes are either activated or activate a downstream gene, providing a growth advantage to the cell. Tumor suppressors are inactivated, allowing uncontrolled growth and proliferation [28]. Mutations in proto-oncogenes and tumor suppressor genes produce phenotypes that are advantageous and positively selected for in cancer cells [18]. Identification of driver genes has been the focus of many studies, but researchers must be able to classify driver genes and passenger mutations to understand their complex dynamics in tumorigenic potential [18].

1.4 Genetic variation in mouse models of human genetic variation and disease

Mus musculus, more commonly referred to as the house mouse or the laboratory mouse, is an important model organism commonly used in genetic research studying complex phenotypes and genetic variation. For the past century, human biology and disease research used mice as a model organism because of their short generation time, large litter size, and the ability to manipulate their genome easily [29]. For example, different breeding schemes can control the level of genetic heterogeneity between individual mice [30]. In 2002, the mouse genome was sequenced, characterized and made available to the public. The characterization found that 99% of mice genes have a homolog in the human genome, making them ideal organisms for experiments modelling the human genome [31]. Mouse models can be classified depending on their origin, breeding scheme, and levels of diversity in the genome. There are seven primary classifications of laboratory mice:
Chapter 1. Introduction

(1) inbred, (2) wild-caught, (3) wild-derived, (4) outbred, (5) recombinant inbred, (6) collaborative cross, and (7) diversity outbred mice.

Inbred mice (1) have been frequently used to model the human genome in experiments because of their genomic stability [32]. Inbred mice are laboratory mice consecutively brother-sister mated for more than 20 generations, resulting in mice homozygous at approximately 99% of all loci, with little to no variation between mice [29, 32]. Additionally, inbred mice are well characterized, and as of 2011, there are genomic sequences of 17 inbred strains of mice available [30]. Inbred mice are beneficial because little to no variation in mice minimizes variability in experiments [29]. One disadvantage is that complete inbreeding in mammals is not common in nature and, therefore, does not represent the levels of heterozygosity and genetic variation seen in humans well.

Wild-caught mice (2) are live-trapped and transported to a laboratory for research, whereas wild-derived mice (3) have been bred in the laboratory for multiple generations. Both wild-caught and wild-derived mice have higher levels of genetic diversity than inbred mice and together are a diverse group, but wild-derived mice are inbred within the laboratory [32]. Once in the laboratory, wild-caught and wild-derived mice will become very similar unless new mice are periodically introduced into the stocks [33]. Both groups of mice originated in a natural environment and are generally exposed to more antigens than the average laboratory mouse, making them ideal candidates for immunology studies [33, 34]. A disadvantage to wild-caught and wild-derived mice is that the genomes of individual mice are not well defined. Thus, information on the heterozygous landscape of each mouse is unique and undefined. This undefined landscape creates difficulty making predictions and reproducing results for confirmation, considering the uniqueness of individual mice.

Outbred mice (4) are laboratory mice that are bred to be a more genetically diverse group and to have higher levels of heterozygosity in their genomes. Breeding mice with different genetic backgrounds leads to recombination events resulting in mosaic genomes
with regions of heterozygosity and increased diversity. Outbred mice have a distinct combination of alleles due to unique recombination patterns and have more considerable variation in SNPs [35, 36]. Compared to inbred mice, outbred mice tend to have more offspring [35, 37], due to inbreeding depression which leads to a decrease in the ability to have an abundance of offspring in inbred mice [38, 39]. Inbreeding can lead to an increase of deleterious recessive alleles within a population or a decrease in favourable heterozygote combinations [38, 39]. Overall, outbred mice may be better subjects for most biomedical research due to their higher levels of genome variation, which better represents the human genome [37].

Recombinant inbred mice (5) are bred from two distinct inbred lines to create progeny that are a mosaic of the two parental inbred lines. Recombinant inbred mice are commonly used for genetic mapping [40]. The recombinant inbred collaborative cross strain (6) of mice are bred from eight diverse inbred mouse strains (C57BL/6J, A/J, NZO/HiLtJ, 129SI/SvImJ, NOD/ShiLtJ, CAST/EiJ, PWK/PJ), including three with wild origins to increase diversity [41–43]. The parental mice are intercrossed and then inbred over three generations to establish collaborative cross strains, which leads to greater allelic diversity and more SNPs [41, 42]. The diversity outbred mice (7) are designed to be unique mice, thus being the most genetically diverse stock of mice used in research. Mice from the collaborative cross strains are randomly mated to establish diversity outbred mice [43, 44]. Both strains display a range of observable phenotypes, have a high degree of heterozygosity, and unique alleles. Collaborative cross and diversity outbred mice are commonly used for mapping quantitative trait loci and phenotyping studies [42–44].

In addition to classifying mouse models dependent on their genetic background and levels of genomic diversity, mouse models can be classified depending on the phenotype they are used to study, whether proxies for a complex trait or a disease. Despite this variation in available mouse models, there is not a model that fits all research needs, and current models may need to be further characterized to suit a research question better.
1.5 Missing genetic variation in mouse models

While previous research has well-characterized SNPs and CNVs in commonly used classical and inbred mouse models, SNPs and CNVs are currently undetected in many other mouse models used to study complex phenotypes and diseases. The development of new technologies has provided the methods to detect variants in these models [22]. Seventeen strains of inbred mice have had SNPs, insertions, deletions, structural variants and functional variants detected and analyzed in their genomes [30]. Forty-one strains of inbred mice have a characterized CNV landscape, including detected CNVs affecting genes related to human disease [45]. While outbred mice contain higher levels of genetic variation and are used to study complex phenotypes and disease, research has not characterized them to the same extent as inbred mice. One study assessed 66 strains of commercially available outbred mice for use in genome-wide association studies [46]. However, it did not focus on cataloging the genetic variation and genes affected in these strains. In addition to outbred mice, the landscape of genetic variation in other specific mouse models of complex phenotypes needs to be characterized to the same extent as inbred mice. This characterization would improve understanding of complex phenotypes, variation in phenotype, clinical outcomes, and drug response. Considering the critical role mouse models play in research, it is crucial to learn about the genetic variation in the mouse genome.

Missing heritability refers to the knowledge gap of variants and their contributions to phenotypes [47]. Similar to the missing heritability of complex traits and diseases in the human genome, there are variants in mouse models that need to be defined to accurately study and understand complex traits and diseases in mouse models. Determining the base levels of genetic variation for outbred mouse strains and disease mouse models will help elucidate the phenomenon of Heterozygote Instability. Methods and approaches to detect genetic variation in mouse models will be further explained below.
1.6 Methods to characterize missing genetic variation

It is essential to detect variants and mutations in the genome to elucidate the genetics of complex phenotypes and diseases. DNA microarrays and next-generation sequencing techniques are the leading genomic methods used to detect genetic variation and mutations. While Sanger sequencing and digital droplet PCR are the leading techniques to confirm previously detected variants. Both detection and confirmation methods will be discussed below. Multiple factors must be considered to choose the most appropriate genetic variation detection method. The specific hypothesis or question to be answered must be analyzed to determine the scope of the project. The best method will differ depending on whether the project evaluates one gene, multiple genes, the whole genome, or a specific set of disease-related genes and depends on the information known before this research. Methods differ depending on whether the research project is looking to discover new variants or confirm previously detected variants. Finally, there are budgetary considerations.

1.6.1 Detection of genome-wide genetic variation

DNA microarrays are a method used to detect SNPs and CNVs at a low cost. Microarrays allow a rapid, simple analytical pipeline to analyze the data that provide high-resolution analysis of SNP and CNV genetic variation genome-wide. The Mouse Diversity Genotyping Array (MDGA) is an Affymetrix® microarray used for genotyping mice and assessing genetic diversity. The MDGA was modelled after the Genome-Wide Human SNP Array 6.0 and designed based on the C57BL/6J mouse strain reference genome [48]. The MDGA uses 493,290 SNP loci, and 421,905 Invariant Genomic Probe (IGP) loci to detect SNP and CNV genotype calls within a mouse sample [48, 49]. The MDGA has eight SNP
probes designed to detect either the reference/major (A) or alternative/minor (B) allele. A duplicate probe for both the sense and antisense strand creates redundancy and controls for position effects. Duplicate probes are placed at different locations on the microarray. IGP probe sets contain six probes per exon — a proximal, medial, and distal probe for the sense and antisense strand. Calling three consecutive probes increases the accuracy of CNV calls [48].

Next-generation sequencing (NGS) is a massively parallel sequencing method used in genomics and transcriptomics. Millions of small fragments of DNA are sequenced simultaneously so that each DNA nucleotide is sequenced multiple times. This repeat sequencing increases the read-depth and sensitivity to improve the accurate detection of variants. The analytical pipelines used are more complex than pipelines used for microarrays but provide more sequence information and increased genome coverage. NGS can be used to sequence the entire genome or a selected subset. Applications of NGS include whole-genome sequencing (WGS), whole-exome sequencing (WES), targeted-exome sequencing (TES) and RNA-sequencing. WGS detects a species’ entire DNA sequence and allows the discovery of genetic aberrations, including SNPs, deletions, insertions and CNVs, in coding and non-coding regions. WES detects all protein-coding genes in the genome by selecting only the subset of DNA, known as exons, which encodes proteins. WES detects the sequence of the exons plus 150 nucleotides into the neighbouring introns and splice junctions. The majority of disease-causing mutations occur in exons, and thus, WES is an excellent method for detecting and discovering disease-causing variants. TES detects the sequence of a select number of genes or selected coding regions within genes of interest. TES can be custom designed or purchased with preselected content. Disease panels focusing on a select set of genes or gene regions with known associations with a disease are available. TES is a cost-effective and efficient approach to detect and confirm driver gene mutations and \textit{de novo} passenger mutations in a chosen subset of genes. RNA-sequencing measures the levels of mRNA molecules expressed, known as
the transcriptome, from the genes of an organism. RNA-sequencing allows analysis of how expression levels differ in different tissues or with a specific disease or treatment. Mutations that alter RNA levels can also be detected through RNA-sequencing when both control and experimental samples are present [50]. WGS provides the highest coverage of the genome and the most significant potential to detect new variants. However, WES, RNA-sequencing and TES allow for less expensive, higher depth analysis of regions of interest and detection of possible disease-causing variants [51]. Two things that drive the selection of the research approach are the scope of the research question and budget. TES is the choice for large-scale targeted mutation analysis. WES is the low-cost option for broad discovery, and RNA-sequencing is comparable to WES in terms of discovery but limited to the biosample transcriptome. WES gives the best option to study HI in functionally relevant areas of the genome.

1.6.2 Confirming previously detected genetic variation

Sanger sequencing and digital droplet PCR (ddPCR) are genetic methods used to confirm previously detected genetic variation and mutations in the genome. Sanger sequencing is used to validate variants in genes or sections of DNA, including SNPs and small insertions or deletions. ddPCR is a method used for amplification and quantification of DNA, commonly used for validation of CNVs.

1.7 Thesis goal and specific objectives

Given that the HI hypothesis formulated from observations in plants has yet to be formally tested in mammals, and there are similar mutational mechanisms in plants and mammals, mouse models with high levels of heterozygosity are proposed as ideal models for the study of HI. Previous evidence implicates HI during meiosis, but it is uncertain whether HI is only germ-cell specific or if HI occurs in somatic cells as well. Certain mouse
models provide a discontinuous landscape of high heterozygosity suitable for detecting potential evidence of mutagenesis associated with HI in germ and somatic cells and permit exploration of the associated underlying mutational mechanisms. Two mouse models are selected in this thesis and evaluated for their suitability to study HI. The first model is the outbred mouse genetic background generated in two outbred mouse stocks, CD-1 and NMRI, and aims to investigate HI in germ cells. The second model is a $Ptch1^{+/-}$ mouse model of basal cell carcinoma and aims to investigate HI in rapidly dividing somatic cells. Both models require characterization of the heterozygous genomic landscape and a candidate-associated mutation type represented here by copy number variants. Well-characterized proxies will allow for testing the HI hypothesis, elucidation of HI-associated mutational mechanisms, and the genetic variation contributing to the missing heritability in complex phenotypes.

Outbred mice better represent the genetic variation found in the human genome than inbred mice [35, 37], which are well-defined and characterized [30]. The full extent of germine genetic variation in outbred mice is presently not characterized - in particular, the extent and nature of CNVs in the genome are underreported. CNVs are known to impact gene dosage and affect complex phenotypes, and outbred mice are frequently used to study these phenotypes [52]. CD-1 mice are used to study genetics, toxicology, carcinogenesis and aging [53], whereas NMRI mice are used to study general biology, toxicology and pharmacology [54]. Both stocks are used to study toxicology due to the high variation between mice, as inter-animal variation is highly valuable in assessing unique susceptibility and responses. Thus, to use CD-1 and NMRI outbred mice as a valid proxy, these CNVs need to be catalogued. Additionally, the genomes of outbred mice provide a unique opportunity to examine HI, as known breeding schemes suggest the genomes of CD-1 and NMRI mice consist of clusters of high heterozygosity separated by clusters of homozygosity. This discontinuous landscape is ideal for examining HI. Characterizing the CNV landscape of outbred mice and the heterozygosity throughout
their genomes will allow further elucidation of this phenomenon.

Human BCC tumors are well-known models of very high levels of acquired genetic variation and provide an excellent candidate to study the phenomenon of HI in rapidly dividing cells. With UVB exposure, mouse models of BCC develop tumors that accurately represent human tumors [55]. BCC mouse models provide a controlled environment to track somatic mutations and analyze whether HI occurs during replication in rapidly dividing cells. Additionally, BCC presents a unique opportunity to analyze the conundrum of benign cancer with very high levels of mutation, strong cancer driver mutations and rich somatic mutagenesis [56]. This conundrum indicates that the potential of a high mutation burden in cancers needs to be further studied to understand the complex phenotype of cancer initiation, malignancy and metastasis. Similar to the characterization of genetic variation in outbred mice, in order for mouse models of BCC to be good proxies for human BCC and be effective models to study the HI hypothesis and contributions to complex phenotypes, the somatic mutations must be detected and characterized within the mouse model.

Thus, the overall goal of this thesis is to evaluate two mouse models of high genetic variation for suitability to test the HI hypothesis. The specific objectives and the rationale for these objectives are as follows:

**Objective 1:** In Chapter 2, I aim to evaluate the genetic diversity in two stocks of outbred mice, CD-1 and NMRI, as proxies for human genetic diversity and suitability for the study of HI. Additionally, I aim to examine the spatial association of CNVs to localized regions of heterozygosity.

**Aim 1:** To detect the chromosomal heterozygous landscape at SNP loci in two outbred mouse stocks, CD-1 and NMRI.

The mouse diversity genotyping array (MDGA) detects SNPs at 493,290 SNP loci, and statistical analysis pipelines are used to portray the distribution of heterozygous SNP loci. The predicted discontinuous landscapes of heterozygosity differ between outbred
mouse stocks, given their different breeding histories.

**Aim 2:** To detect and characterize the CNV landscape in two outbred mouse stocks, CD-1 and NMRI.

The MDGA detects genome-wide CNVs using 493,290 SNP loci and 421,905 invariant genome probes. Considering the higher level of genetic variation and heterozygosity in outbred mice, it is predicted that there will be multiple CNVs detected per mouse, with features of CNVs, including length, copy number state, genic impact and location, differing between the two outbred mouse stocks. Genic impact involves pathway and gene ontology analysis.

**Aim 3:** To compare the level and nature of genetic variation in outbred stocks to reported levels of heterozygosity and CNVs in inbred mice, wild-caught mice, and wild-derived mice.

The CNV landscape and profile of CNVs detected in Aim 2 are compared between CD-1 and NMRI mice and reported CNV profiles of inbred, wild-caught and wild-derived mice. It is predicted that the two outbred mouse stocks have higher heterozygosity levels and CNVs than inbred mice but less than wild-caught and wild-derived mice.

**Aim 4:** To assess the spatial association between heterozygosity at SNP loci and CNVs as indirect evidence consistent with the Heterozygote Instability hypothesis.

Using the characterized landscapes from Aim 1 and Aim 2, spatial statistical tools are used to analyze the landscape of heterozygosity and the landscape of CNVs per chromosome to detect possible proximal association of CNVs to clusters of heterozygosity. According to the tenets of the HI hypothesis, CNVs are predicted to be more numerous with higher levels of heterozygosity and proximal to greater density of heterozygosity.

**Objective 2:** In Chapter 3, I aim to evaluate mouse models of BCC as proxies for human genetic diversity and suitability for the study of HI with consideration of heterozygosity and somatic mutations arising in the context of tumorigenesis.

**Aim 1:** To complete a systematic literature review of BCC genic mutations to evaluate
the reported genetic variation in human BCC and in mouse models of BCC.

A systematic literature review between the years 1990 and 2020 was performed, and two lists of reported mutations were compiled, one for human BCC and one for mouse models of BCC.

**Aim 2:** To define the knowledge gap between the reported mutations in human BCC tumors and reported mutations in mouse models of BCC.

The lists of mutations compiled from Aim 2 were compared between human and mouse BCC. It is expected that human and mouse BCC share similarities.

**Aim 3:** To design an experimental workflow to detect genetic variation in mouse models of BCC to characterize this model for the study of Heterozygote Instability.

Considering the extent of the knowledge gap between humans and mice identified in Aim 2, multiple experimental designs are presented to characterize the mutation burden in BCC mouse models at different resolutions and sensitivities.
References


Chapter 2.
Characterization of genetic variants in discontinuous landscapes of heterozygosity in two outbred mouse stocks

2.1 Introduction

Genetic variants contribute to differences between individuals and variations in complex phenotypes, health, and clinical outcomes. However, the full extent of genetic contributions to many complex phenotypes, phenotypic diversity, aberrant or atypical phenotypes and fitness are not entirely understood. This gap in knowledge regarding how the genetic variants and their contributions can affect phenotypes is referred to as missing heritability [1] and filling this gap is a focus of research. To better understand how genetic variants affect complex phenotypes, mutational mechanisms associated with genetic variation must also be defined. The mutation rate is not constant along the chromosomal sequence, and while multiple mechanisms are known, all mechanisms underlying this are not fully understood [2]. Variation in mutation rate can be sequence-dependent. For example, in humans and mice, higher rates of C to T substitutions are seen more frequently at CpG sites or CG dinucleotides [3–5]. Additionally, repetitive sequences or sequences with high similarity can cause polymerase slippage or misalignment during DNA replication or DNA repair [5]. Elucidating the mutagens and mechanisms of mutations will create...
opportunities for better understanding sources and the impact of genetic variation in complex phenotypes.

Yang et al (2015) was the first to demonstrate a connection between heterozygosity and mutation rate through investigating SNP heterozygosity and small insertions and deletions in *Arabidopsis* [6]. During meiosis, plants with high levels of heterozygosity have higher mutation rates than plants with lower levels, and the mutations were occurring in close proximity to heterozygous loci and crossover events [6]. The phenomenon of increased heterozygosity connected to increased mutations was then tested in peach plants by crossing either intra-species plants or inter-species plants. Xie et al (2016) found that hybrid plants had higher mutation rates, comparable to results demonstrated in *Arabidopsis* [7]. The specific phenomenon observed in *Arabidopsis* and peach has not yet been reported in mammals. HI has been computationally demonstrated with local heterozygosity and microsatellites in 1163 human genome sequences from the 1000 Genome Project [8]. Similar genetic variants and mutational mechanisms exist in *Arabidopsis* and mammals, including point mutations, small insertions and deletions and CNVs [9–12], and an increased level of mutations at repetitive DNA sequence and near transposable elements [12], and UV-induced mutagenesis [10]. This likeness indicates that similar mutational mechanisms, like HI, may occur in mammals where they are poorly studied. HI could be measured by identifying heterozygosity at single nucleotide polymorphisms (SNPs) across the genome as they are easily assayed using microarrays.

While there are many types of genetic variants, this study uses SNP genotypes to measure heterozygosity and copy number variants (CNV) to measure mutation occurrence. The goal is to characterize heterozygous SNP and CNV landscapes in a constitutive genome and look at their spatial associations. SNP heterozygosity can have discontinuous landscapes with localized variation that ranges from very high to very low levels in different organisms of the same species, which provides an ideal landscape to investigate HI. Research implicates SNPs and CNVs in disease and complex phenotypes, thus characteri-
zation is important [13–17]. CNVs are good candidates for the study of HI induction given their mechanistic association with replication and recombination, known genic impact and impact on complex phenotypes [14, 18, 19]. CNVs occur most frequently due to DNA replication and recombination errors caused by chromosome misalignment, polymerase stalling, or replication stress [14, 18, 19]. CNV formation may occur during non-allelic homologous recombination, fork stalling and template switching, non-homologous end joining due to double-stranded DNA breaks, mobile element insertion, and retrotransposition [14, 18, 19]. A potentially underappreciated mutational mechanism is how heterozygosity acts to increase mutations locally or proximally during DNA replication and recombination.

To fully characterize the CNV landscape in a model, the nature of CNVs must be defined, including location, length, copy number state, recurrence, genes affected, and copy number variable regions (CNVR) – a region of the genome that a CNV overlaps in multiple individuals [17, 20]. Researchers examined CNVRs in humans and found that CNVRs encompassed more of the genome than SNPs, but previously have not been a focus of characterization studies [17]. This finding indicates that CNVs should be defined in each model of interest to study and understand associated mutational mechanisms and contributions to phenotypic variation and complex traits.

Mutation research often uses inbred mice as models, but their genetic background precludes analysis of the impact of heterozygosity. Outbred mice provide the opportunity to examine mutagenesis in a genome landscape of high heterozygosity, with distinct regions of high heterozygosity and regions of high homozygosity, creating discontinuous segments of heterozygosity along the genome. The discontinuity of heterozygosity in the genomes of these mice contains regions of dense clustering of heterozygosity separated by desert regions with low or no heterozygosity, resulting from how outbred mice are bred. A discontinuous landscape of heterozygosity is ideal for examining whether or not heterozygosity is contributing to mutagenesis. This landscape allows first the assessment of CNVs and a possible association with heterozygosity, followed by examining the nature
and distribution of the SNP heterozygosity to infer mutational mechanisms associated with deletion or duplication creating CNVs. In comparison, wild-derived and wild-caught mice have higher levels of genetic diversity than inbred mice and are a diverse group \[21\], with the expectation being that the heterozygous landscape of wild-derived and wild-caught mice are uncharacterized and more random compared to the genomes of outbred mice. Similarly, the recombinant inbred collaborative cross strain of mice has higher levels of genetic variation due to originating from eight diverse inbred mouse strains, three with wild origins \[22–24\]. Finally, diversity outbred mice are bred to be the most genetically diverse group of mice through randomly mating mice from the collaborative cross stock \[24, 25\]. Both collaborative cross and diversity outbred mice display a range of observable phenotypes, have a high degree of heterozygosity, and unique alleles \[22–25\]. The increase of genetic variation seen in the wild-derived, wild-caught, collaborative cross, and diversity outbred strains creates a challenge for predicting the heterozygous genomic landscapes and reproducing genomic locations where HI can be studied. The discontinuous landscape of heterozygosity presented in the genomes of outbred mice is ideal for investigating the mechanistic potential of heterozygosity increasing CNV formation due to a more controlled and predictable environment for hypothesis testing. It is this clustering and spacing of heterozygosity that this chapter will aim to characterize and examine with respect to localized effects on CNV occurrence.

Two commercially available stocks of outbred mice, CD-1 (Caesarian Derived-1) and NMRI (Naval Medical Research Institute) \[26\] present a unique opportunity to examine mutagenesis in two distinct landscapes of clustered heterozygosity. Not only is there variation in the clusters of heterozygosity between CD-1 and NMRI mice, but there is variation in clustering between individual mice within the two stocks. More specifically, there are differences in genetic variation within a mouse between chromosomes and different regions of chromosomes. Thus, CD-1 and NMRI mice are ideal models to study if local heterozygosity affects mutations. Both stocks originated from a group of
200 wild-caught mice with higher levels of heterozygosity in their genome (Figure 2.1). Researchers transported nine mice from Switzerland to the USA, where the mice were inbred. The CD-1 stock was inbred for approximately 15 generations, whereas the NMRI stock was inbred for approximately 51 generations. This inbreeding introduced clusters of homozygosity to the genomes of both stocks of mice. The clusters of heterozygosity are discontinuous throughout the genome, separated by desert regions with low to no heterozygosity. Both CD-1 and NMRI mice were confirmed to have higher levels of variation than inbred mice [27, 28]. Finally, researchers maintain both stocks through outbreeding, which re-introduced heterozygosity and increased genetic variation [26]. This breeding history leads to genomic heterozygosity patterns that make CD-1 and NMRI mice ideal models for testing the spatial distribution of heterozygosity and mutations. Given that they have different levels and spatial distributions of heterozygous loci, their genomes provide unique conditions to see if mutations occur in localized regions of clustered heterozygosity or clustered homozygosity. The variation within the genomes of CD-1 and NMRI mice creates unique genomic contexts that allow for in-depth and varied analysis and comparison of heterozygosity and CNV occurrence between mice, between chromosomes and within chromosomes.

Figure 2.1: Breeding history of CD-1 and NMRI mouse stocks. Types of breeding are highlighted in red text and defined in the inset red box.
Both CD-1 and NMRI mouse stocks have germline genetic variation that has yet to be characterized and must be defined before the novel complex phenotype of HI demonstrated in *Arabidopsis* and peach plants can be evaluated in the genomes of CD-1 and NMRI mice. CD-1 mice have had a small proportion of heterozygosity and CNVs detected in their genome, whereas NMRI mice have had neither characterized. The level of heterozygosity in CD-1 mice was assessed at 3,572 SNPs across the genome to determine the potential of CD-1 mice as models for genetic studies. However, it has not been assessed on a larger scale or with greater resolution [29]. CD-1 mice are found to have genetic variation similar to what is seen in a human founder population, which is a population that originated from a very small number of individuals [29]. Anxiety-associated CNVs have been analyzed and catalogued in CD-1 mice [30]. CNVs have yet to be examined in NMRI mice. CD-1 and NMRI mice are the ideal mouse stocks to study the complex phenotype of HI and begin to fill the knowledge gap surrounding genetic variation and complex phenotypes, in particular through characterizing the CNV and heterozygous SNP landscape.

This chapter will analyze publicly available genotype array data to detect and catalogue heterozygous SNP genotypes and CNVs related to complex phenotypes and examine the spatial distribution of CNVs within the unique background of outbred mice to determine if heterozygosity is associated with CNV occurrence (Figure 2.2). Combined detection and characterization of heterozygous SNP genotype and CNV calls was achieved using the Mouse Diversity Genotyping Array (MDGA) [20, 31]. This array has been used for genotyping both mouse stocks, and the data are publicly available. Analysis of CNV calls was achieved using HD-CNV (Hotspot Detector for Copy Number Variants) [32]. HD-CNV examines CNV calls and detects overlapping regions of DNA to identify CNVs occurring in multiple samples, termed recurrent CNVs or CNVs that occur in one sample only, termed unique CNVs (Figure 2.3) [32]. This combination of technologies is ideal for simultaneous SNP and CNV analysis considering sufficient genomic resolution, cost, efficiency, rapid bioinformatics and publicly available data sets of sufficient sample size.
Chapter 2. Characterization of two stocks of outbred mice

404 CEL files were downloaded from the Center for Genome Dynamics – 101 CD-1 files and 303 NMRI files.

99 CD-1 files and 279 NMRI files passed quality controls

Tissue/DNA extraction, isolation, and preparation
Mouse Diversity Genotyping Array
493,290 SNP loci probes & 421,905 Invariant Genome Probes (IGP)
Fluorescence detection
SNP genotyping
Affymetrix Genotyping Console
CNV calling PennCNV

Data analysis
• Detection and characterization of unique and shared CNVs
• Detection of autosomal heterozygous SNP genotypes
• Analysis of spatial association between CNVs and heterozygous SNPs

Figure 2.2: Methods workflow from sample extraction to data analysis. Steps completed by the Jackson Laboratory are shown by grey arrows and steps completed in-house are shown by purple arrows. Mouse and laptop images retrieved from SMART Servier Medical Art.

The spatial distributions of heterozygous SNP loci and CNVs were analyzed using three plots — rainfall plots, rainbow plots, and J statistic plots [33, 34]. Rainfall plots portray the genomic landscape of heterozygous SNP loci across the genome, allowing the visualization of clusters, defined as heterozygous SNP dense regions, and deserts, defined as heterozygous SNP sparse regions (Figure 2.4). Rainbow plots portray spatial associations between heterozygous SNP loci and CNVs (Figure 2.5). J function statistically analyzes the association between heterozygous SNP loci and CNVs, determining possible relationships and significance (Figure 2.6).
Figure 2.3: **Hotspot Detector for Copy Number Variants (HD-CNV) creates a heatmap karyotype of recurrent and unique CNVs.**

**A.** CNV calls from individual mice are compiled per chromosome and analyzed for sequence overlap. A CNV that shares no sequence overlap is unique. Recurrent CNVs share a 40% or greater sequence overlap. A heatmap displays unique and recurrent CNVs, where red represents a higher number of overlaps, while blue represents a unique CNV. Nodes represent one CNV event, and edges connect nodes if the two CNV events share sequence overlap. Mouse image retrieved from SMART Servier Medical Art.

**B.** A traditional human karyotype. Image retrieved from SMART Servier Medical Art.

**C.** A heatmap karyotype of recurrent and unique CNVs. The size of each circle is relative to the number of CNV events on that chromosome. The Y chromosome was not analyzed due to a lack of probe representation.
Chapter 2. Characterization of two stocks of outbred mice

Figure 2.4: **Rainfall plots portray the landscape of heterozygosity at SNP loci.** Heterozygous SNPs are plotted at their respective genomic location (X-axis) against the distance (Y-axis) between the heterozygous SNP and its nearest previous heterozygous SNP locus neighbour plotted on a logarithmic scale. **A.** If two heterozygous SNPs are farther apart, the point is plotted higher up versus if two heterozygous SNPs are closer together, the point falls lower down on the graph. **B.** Rainfall plots show regions without heterozygous SNPs, termed deserts, versus regions dense with heterozygous SNPs, termed clusters, across the chromosome.

Figure 2.5: **Rainbow plots display the potential spatial association between heterozygous SNPs and CNVs.** **A.** CNVs are plotted at their respective genomic location as red dots along the X-axis. Heterozygous SNPs are plotted at their genomic location as black dots against their distance to the closest CNV. If the heterozygous SNP is close to a CNV, the particular point is plotted lower on the Y-axis, indicating a possible spatial association. **B.** A rainbow plot of a chromosome with a potential spatial association between two CNVs and heterozygous SNPs.
Figure 2.6: **J function** is a statistical tool used to analyze the spatial distribution of heterozygous SNPs relative to CNVs.  

A. Four possible distributions of heterozygous SNPs relative to a CNV: distal — CNV occurs far from heterozygous SNPs, no association, proximal — CNV occurs close to heterozygous SNPs, or both — CNV occurs first close to or far from heterozygous SNPs, and then the chromosomal heterozygous landscape varies between regional clusters of heterozygosity and clusters of homozygosity.  

B. J statistic assesses the significance of the association after running 1000 simulations and produces two confidence bands.  

C. If the observed J function crosses the confidence bands, the association is significant, either the upper confidence band (distal), the lower confidence band (proximal), or crosses one confidence band and then the other (both). If the function does not cross the confidence bands, there is no association between the heterozygous SNPs and CNVs on that particular chromosome.

**2.1.1 Research goal, central hypothesis, and specific objectives**

**Research goal:** The purpose of this study is 1) to detect and characterize SNP and CNV genetic variation in CD-1 and NMRI mice to fill missing heritability potentially related to complex phenotypes, 2) to confirm the discontinuous landscape of heterozygosity at SNP loci, and 3) to analyze the spatial association between CNVs and heterozygous SNP loci to test the Heterozygote Instability hypothesis with germline mutations in outbred mice.
Predictions: CD-1 and NMRI mice will display differences in the content and nature of CNVs compared to each other and inbred mice. Both stocks will show discontinuous heterozygosity at SNP loci, with NMRI mice having less heterozygosity and greater deserts than CD-1 mice. Given the recent observation in *Arabidopsis* and peach plants that there is a connection between heterozygosity and mutation rate, the proximal occurrence of insertions and deletions, and the similar DNA sequence mutational mechanisms in plants and mammals, both stocks will have CNVs close to localized regions of heterozygous SNP loci across the genome. These mouse models provide an opportunity to gauge the association of CNV occurrence with higher levels of chromosomal SNP heterozygosity. Due to the increased levels of clustered heterozygosity in CD-1 mice, there will be an increased number of CNVs per mouse in CD-1 mice compared to NMRI mice.

The specific objectives are:

1. To detect the chromosomal heterozygous landscape at SNP loci within CD-1 and NMRI mice.

2. To detect and characterize the CNV landscape and identify CNVRs shared within and between CD-1 and NMRI mice.

3. To compare the level and nature of genetic variation in CD-1 and NMRI mice to reported variants in inbred mice, wild-caught mice, and wild-derived mice.

4. To explore HI through examining the spatial association of CNVs and heterozygous SNP loci across the genome.
2.2 Materials and methods

2.2.1 Samples and SNP genotype and CNV calling

A total of 404 publicly available CEL files were downloaded – 101 CD-1 and 303 NMRI mouse files, from the Center for Genome Dynamics at the Jackson Laboratory [35]. Two CD-1 and 15 NMRI files were excluded due to not meeting a 97% SNP genotype call rate or issues downloading the files. Accordingly, this study examined the autosomes of 99 CD-1 (43 males and 56 females) and 279 NMRI (all males) mouse CEL files and the X chromosome of all-male mice (Appendix 2A). A list of filtered SNP probes was generated by previous Hill laboratory members and used for SNP genotyping in this chapter [20, 36, 37]. Probes on the Y chromosome represented too few sites, thus there was an insufficient amount of data for analysis. Generally, SNP genotype and CNV calling and analysis follow a previously reported pipeline [20] and are discussed in more detail below.

To detect the heterozygous landscape at SNP loci, SNP genotype calls were generated using Affymetrix Power Tools [38], which uses the BRLMM-P algorithm (Bayesian Robust Linear Model with Mahalanobis distance classifier - Perfect-match probes) [39]. The algorithm clusters calls into three categories, AA (homozygous, major allele), AB (heterozygous) and BB (homozygous, minor allele). Any calls that do not fall within those categories are categorized as a ‘no call.’ To detect and characterize the CNV landscape, CNVs were identified using PennCNV, which applies an HMM algorithm (Hidden Markov Model) to the genotype clustering file produced in the Genotyping Console [20] to produce a Log R Ratio (LRR) and a B allele frequency (BAF) file. Using these files, an in-house script and the reference genome (UCSC:mm9), PennCNV detects the copy number (CN) state – complete loss (state 0), partial loss (state 1) or gain (state 3 and 4) [40]. A separate run was performed to detect the CNVs on the X chromosome by applying the
-chrX option [40]. CNV calls were filtered to contain at least three probe markers with the same intensity, length between 500 bp to 1 Mb, and LRR less than 0.35, and a BAF drift less than 0.01. The detected SNPs and CNVs were used in subsequent methods to characterize the landscape of heterozygous SNPs and CNVs in both CD-1 and NMRI mice.

2.2.2 Genetic distance determination and phylogenetic analyses

To compare the level of genetic diversity in CD-1 and NMRI mice to each other and to human genetic diversity, SNP genetic distance was calculated by totalling the number of loci between pairs of samples that did not share the same genotype call (AA, AB, BB) divided by the total number of SNP loci (Appendix 2B). CNV genetic distance was calculated by totalling the number of CNV genotype differences (CNV loci that differ in copy number (CN) state: 0, 1, 2, or 3+) between pairs of samples divided by the total number of SNP and invariant genomic probe (IGP) loci. A pairwise distance matrix was created [20]. Neighbour-joining trees were created for CD-1 and NMRI mice using the SNP and CNV genetic distance matrices by applying the Ape (Analyses of Phylogenetic and evolution) package on Studio R version 3.4.1. These trees were saved as Newick files and uploaded to FigTree for creating and colouring the Neighbour-joining trees. A select set of samples identified to be closely related through genetic distance analysis were coloured in both the SNP and CNV phylogenetic trees to visualize how related samples are placed throughout each tree. A Mantel test was performed to test the association between the SNP genetic distances and CNV genetic distances, using a P-value cut off less than 0.05.

2.2.3 Analysis of the landscape of heterozygosity at SNP loci

To detect the clustering and landscape of heterozygosity at SNP loci, spatial statistical tools were used to visualize the autosomal genomic landscape of heterozygous SNP
Chapter 2. Characterization of two stocks of outbred mice

2.2.4 Detection of recurrent CNV regions and unique CNVs

Recurrent CNVs are deletions and duplications observed in more than one mouse at the same genomic location. To characterize the CNV landscape in terms of CNVRs, recurrent and unique CNVs were identified using HD-CNV (Hotspot Detector for Copy Number Variants), using 40% reciprocal overlap [32]. CNVRs are identified using HD-CNV. Graph files produced by HD-CNV were input into Gephi, an open source for visualizing networks [42] and formatted using the Fruchterman-Reingold layout. The output images were normalized for size based on the maximum and minimum number of CNVs for all chromosomes and then assembled into a heatmap karyotype. A script (Appendix 2C) was written and used to create CNV Landscape Plots (CNV-LP), which portray the genomic footprint of all CNVs within the genome, showing CNV features including genomic location per chromosome and CN state. A CNV-LP allows visualization of the distribution of recurrent and unique CNVs across the genome and allows the identification of CNVRs or hotspots. Finally, a CNV-LP allows comparisons of CNV landscapes between mice within a stock and between stocks.

2.2.5 Pathway and GO term characterization of genes overlapped or encompassed by recurrent and unique CNVs

To characterize the CNV landscape in terms of genic content and enriched pathways and Gene Ontology (GO) terms, an annotated M. musculus genome (GRCm38.p5) was

Genotypes [33, 34]. Rainfall plots [41] were adapted to plot the inter-locus distance between heterozygous SNP genotypes. Rainfall plots provide a visual representation of regional density in heterozygous SNP genotypes per chromosome, including clusters and deserts. The detected chromosomal landscape of heterozygous SNP loci was then compared to the detected CNV landscape, discussed further below.
downloaded from Ensembl Genome Browser [43, 44]. This annotation file and two files containing all CNV events in CD-1 and NMRI mice were uploaded to Galaxy – an open-source platform for analyzing biomedical data [45]. Intersect intervals under BEDTools were used to determine which CNVs contained or overlapped a gene from the annotated mouse genome [46]. Gene lists were created for all the genes identified separated by CN state. These lists were individually uploaded into Metascape for enrichment analysis [47]. GO Enrichment Analysis was performed using The Gene Ontology Resource [48, 49]. DAVID v6.8 (Database for Annotation, Visualization and Integrated Discovery) was used to find enriched KEGG pathways [50, 51]. Finally, PANTHER (Protein Analysis Through Evolutionary Relationships) 13.1 was used for gene analysis and pathway enrichment [52]. Enriched terms were included using a P-value cut-off of less than 0.05 and were compared to identify similarities and differences within and between stocks.

### 2.2.6 Analysis of spatial association between heterozygous SNP genotypes and CNVs

To investigate for indirect evidence supporting HI in the genomes of CD-1 and NMRI mice, spatial statistical tools were used to analyze the distribution of heterozygous SNP genotypes and CNVs. Possible distributions of CNVs include proximal, distal, varied (both) or with no association to heterozygous SNP genotypes [33, 34]. Rainbow plots visualize possible spatial associations between heterozygous SNP genotypes and CNVs. J statistic is used to determine the distribution and assess the significance of the spatial association.

### 2.2.7 Statistical Analysis

To statistically test detected indirect evidence supporting HI, a MANOVA test was performed to analyze the potential significance of the association between chromosomal
heterozygosity and CNVs with proximally associated heterozygous SNP loci. A Student’s t-test was performed to analyze if the potential association was significantly different in CD-1 and NMRI mice. A Spearman’s rank correlation coefficient analysis was performed to analyze the correlation between autosomal heterozygosity and the occurrence of proximal association per autosome.

2.3 Results

2.3.1 Levels of genomic and chromosomal heterozygosity at SNP loci differ in CD-1 and NMRI mice

To characterize the level of heterozygosity and the heterozygous landscape in CD-1 and NMRI mice, SNP genotyping calls were reported for 493,290 SNP loci. Measures of diversity for each stock were recorded (Table 2.1). The average heterozygosity per chromosome was recorded (Table 2.2), and when chromosomes were analyzed individually, the range of heterozygosity was 7.82-14.20% and 3.03-11.02% in CD-1 and NMRI mice. CD-1 mice compared to NMRI mice have a higher level of genomic heterozygosity per mouse and a higher percent heterozygosity on each autosome. These results will be subsequently used to identify clusters of heterozygosity throughout the genomes of CD-1 and NMRI mice.

Table 2.1: SNP genotyping results by outbred mouse stock.

<table>
<thead>
<tr>
<th>Stock</th>
<th>Samples</th>
<th>Call (%)</th>
<th>No Call (%)</th>
<th>Heterozygosity (%)</th>
<th>BB (%)</th>
<th>BAF (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD-1</td>
<td>99</td>
<td>99.75</td>
<td>0.25</td>
<td>10.68</td>
<td>14.70</td>
<td>40.08</td>
</tr>
<tr>
<td>NMRI</td>
<td>279</td>
<td>99.76</td>
<td>0.24</td>
<td>6.20</td>
<td>16.67</td>
<td>39.54</td>
</tr>
</tbody>
</table>

a Percent of calls that were successfully assigned (AA, AB, BB)
b Percent of calls not made (neither A nor B)
c Percent of calls that are heterozygous (major allele and minor allele)
d Homozygosity of the minor allele
e B Allele Frequency (BAF): Abundance of the minor allele
Table 2.2: Average heterozygosity at SNP loci per autosome.

<table>
<thead>
<tr>
<th>Chr</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD-1</td>
<td>10.90</td>
<td>12.86</td>
<td>10.41</td>
<td>8.89</td>
<td>12.31</td>
<td>10.70</td>
<td>10.06</td>
<td>8.87</td>
<td>9.87</td>
<td>7.84</td>
</tr>
<tr>
<td>NMRI</td>
<td>7.49</td>
<td>9.83</td>
<td>5.85</td>
<td>5.13</td>
<td>7.70</td>
<td>4.34</td>
<td>6.72</td>
<td>4.07</td>
<td>6.22</td>
<td>3.03</td>
</tr>
</tbody>
</table>

2.3.2 CD-1 and NMRI mice have similar SNP and CNV genetic distance values and display relatedness similarly on Neighbour-Joining trees

To compare the level of genetic diversity in CD-1 and NMRI mice to each other and inbred, wild-caught, and wild-derived mice, pairwise comparisons between samples were calculated, and found SNP distance values range from 0.0093–0.18 and 0.0085–0.11 for CD-1 and NMRI mice. Pairwise comparisons between samples found CNV distance values range from 0.00011–0.0076 and 0.0002–0.0099 for CD-1 and NMRI mice. The SNP and CNV genetic distances calculated were used to make Neighbour-Joining trees for CD-1 (Figure 2.7) and NMRI mice (Figures 2.8). Additionally, the subset of closely related samples – coloured in purple or blue, respectively, cluster in the SNP trees and are dispersed throughout the CNV trees (Figure 2.7 and 2.8). A Mantel test determined no significant difference between the SNP and CNV genetic distances, producing P-values of 0.08 and 0.73 for CD-1 and NMRI mice.
Figure 2.7: Unscaled Neighbour-Joining tree for SNP (A) and CNV (B) diversity in the CD-1 stock (n = 99). Pairwise comparisons were used to create a genetic distance matrix using SNP and CNV genotype calls. Scale bars represent genetic distance or difference between samples. A subset of closely related samples was coloured (purple) in both trees to visualize how closely related samples cluster on SNP compared to CNV trees.
Figure 2.8: Unscaled Neighbour-Joining tree for SNP (A) and CNV (B) diversity in the NMRI stock (n = 279). Pairwise comparisons were used to create a genetic distance matrix using SNP and CNV genotype calls. Scale bars represent genetic distance or difference between samples. A subset of closely related samples was coloured (blue) in both trees to visualize how closely related samples cluster on SNP compared to CNV trees.
2.3.3 CD-1 mice have a higher average of CNVs per mouse than NMRI mice, with more gains and longer CNVs yet a smaller percentage of CNVs that contain or share an overlapping region with a gene.

To characterize the CNV landscape in CD-1 and NMRI mice, CNVs were detected in both stocks of mice. CD-1 mice had 1690 autosomal CNVs, with an average of 17 CNV calls per mouse, a minimum of 11 and a maximum of 90 CNVs (Appendix 2D). Comparatively, NMRI mice had 3935 autosomal CNVs, with an average of 14 CNV calls per mouse, a minimum of four CNVs and a maximum of 78 CNVs (Appendix 2E). There was a total of 96 and 378 CNVs in male CD-1 and NMRI mice on the X chromosome, respectively. CNV calls were analyzed on the autosomes of both outbred stocks, compared to the autosomes of classical, wild-derived and wild-caught mice [20] (Table 2.3) and the X chromosome of the male mice in both outbred stocks (Appendix 2F). NMRI mice have a much higher loss-gain ratio than CD-1 mice, indicating a much greater proportion of CN complete and partial losses. A positive correlation between the number of CNVs per autosome and increasing autosome heterozygosity was seen in both stocks (Figure 2.9).

Table 2.3: Autosomal CNV calls by mouse stock and copy number state.

<table>
<thead>
<tr>
<th>Stock</th>
<th>Samples</th>
<th>CNV Calls</th>
<th>Copy Number State&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Loss/Gain&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>CD-1</td>
<td>99</td>
<td>1690</td>
<td>65</td>
<td>(0.65)</td>
</tr>
<tr>
<td>NMRI</td>
<td>279</td>
<td>3935</td>
<td>285</td>
<td>(1.02)</td>
</tr>
<tr>
<td>C57BL/6J&lt;sup&gt;c&lt;/sup&gt;</td>
<td>8</td>
<td>90</td>
<td>0</td>
<td>(0.00)</td>
</tr>
<tr>
<td>WD&lt;sup&gt;d&lt;/sup&gt;</td>
<td>52</td>
<td>2611</td>
<td>1214</td>
<td>(23.34)</td>
</tr>
<tr>
<td>WC&lt;sup&gt;e&lt;/sup&gt;</td>
<td>19</td>
<td>969</td>
<td>231</td>
<td>(12.15)</td>
</tr>
</tbody>
</table>

Values in brackets are normalized by sample count.

<sup>a</sup> Copy number states: full deletion (CN 0), partial deletion (CN 1), and gain (CN 3+)

<sup>b</sup> Total number of losses (CN 0 and 1) divided by number of gains (CN 3+)

<sup>c</sup> Data retrieved from Locke et al., 2015

<sup>d</sup> Wild derived (WD) mice

<sup>e</sup> Wild caught (WC) mice
Chapter 2. Characterization of two stocks of outbred mice

Figure 2.9: Distribution of the average number of CNVs on autosomal chromosomes and chromosomal heterozygosity at SNP loci. Distribution of the average number of CNVs on the autosomal chromosomes of CD-1 mice (n = 99) are shown in purple and for NMRI mice (n = 279) in grey. Linear trendlines (dotted lines) are included for CD-1 and NMRI data and indicate a positive correlation between increasing chromosomal heterozygosity and average number of CNVs per chromosome in CD-1 and NMRI mice.

For the autosomes, the average length of CNVs was 93,643 base pairs and 70,583 base pairs in CD-1 mice and NMRI mice. The median length was 54,193 base pairs and 47,858 base pairs in CD-1 mice and NMRI mice. The average length of CNV gains was 122,793 and 146,575 base pairs, and the average length of CNV losses was 50,559 and 60,893 base pairs for CD-1 and NMRI mice, respectively (Figure 2.10). CN gains in CD-1 mice display a bimodal distribution of lengths, with a large group of smaller CNVs and a smaller group of longer CNVs.

CD-1 mice have more CN gains, whereas NMRI mice have more CN losses. In CD-1 mice, chromosome 17 had the most CNV events, and in NMRI mice, chromosome 3 had the most CNV events. The length of the chromosome did not correlate to the number of CNVs on that chromosome.
Including the X chromosome, CNVs across all samples affect 9.2 and 4.1% of the genome in CD-1 and NMRI mice. In CD-1 and NMRI mice, 80.7 and 89.6% of the CNV events contained or shared an overlapping region with a gene. When broken down into copy number states, CNV events that contained or shared an overlapping region with a gene were 43.3 and 60.7% CN complete losses, 70.8 and 90.2% CN partial losses, and 80.6 and 97.9% CN gains in CD-1 and NMRI mice, respectively.

![Figure 2.10: Length of CNV gains and losses across the autosomes in both outbred mouse stocks. Distribution of lengths of CNV events in CD-1 mice (n = 99) and NMRI mice (n = 279) across the autosomes. CD-1 mice have more shorter CNVs, both CN losses and CN gains. CD-1 gains show a bimodal distribution of lengths.](image)

Overall, the landscape of CNVs differs in CD-1 and NMRI mice in terms of length of CNVs, loss-gain ratio, copy number state and the percentage of CNVs the contain or share an overlapping region with a gene. Both stocks have a positive correlation between increasing autosome heterozygosity and the number of CNVs per autosome, indicating possible evidence for HI. To further investigate HI in the genomes of these mice, specific chromosomal regions with clusters of high levels of heterozygosity must be examined for their spatial distribution in regards to copy number variants.
2.3.4 Recurrent CNVs were identified on all autosomes and the X chromosome in CD-1 and NMRI mice, with identification of CNVRs unique to one stock and CNVRs shared between the two stocks

Towards further characterizing the CNV landscape in the genomes of CD-1 and NMRI mice, HD-CNV was used to detect recurrent CNV regions and unique CNVs on the autosomes (Table 2.4) and X chromosome (Table 2.5). 9.57 and 3.95% of CNVs were unique in CD-1 and NMRI mice, respectively. HD-CNV graph files were portrayed as a karyotype of recurrent and unique CNV events on the autosomes and X chromosomes of the CD-1 mice and NMRI mice. These karyotypes were combined to compare CD-1 and NMRI mice (Figure 2.11).

Table 2.4: Number of unique and recurrent CNVs on the autosomes.

<table>
<thead>
<tr>
<th>Stock</th>
<th>Recurrent</th>
<th>Unique</th>
<th>Unique / recurrent ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD-1</td>
<td>1536</td>
<td>154</td>
<td>0.10</td>
</tr>
<tr>
<td>NMRI</td>
<td>3774</td>
<td>160</td>
<td>0.04</td>
</tr>
</tbody>
</table>

*a* Recurrent CNVs share a 40% or greater sequence overlap between CNV calls in individual mouse samples, as detected by HD-CNV.

*b* Unique CNVs only occur in one mouse.

Table 2.5: Number of unique and recurrent CNVs on the X chromosome.

<table>
<thead>
<tr>
<th>Stock</th>
<th>Recurrent</th>
<th>Unique</th>
<th>Unique / recurrent ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD-1</td>
<td>94</td>
<td>2</td>
<td>0.02</td>
</tr>
<tr>
<td>NMRI</td>
<td>374</td>
<td>4</td>
<td>0.01</td>
</tr>
</tbody>
</table>

*a* Recurrent CNVs share a 40% or greater sequence overlap between CNV calls in individual mouse samples, as detected by HD-CNV.

*b* Unique CNVs only occur in one mouse.
Figure 2.11: HD-CNV graph files visualized as a karyotype of recurrent and unique CNV events on the autosomes in CD-1 (n = 99) and NMRI (n = 279) samples and the X chromosome in male CD-1 (n = 43) and NMRI (n = 279) mice. The size of each circle is relative to the number of CNV events on that chromosome. A CNV event is represented as a node, which is depicted as one dot in the heatmap. Edges are lines connecting nodes if the two CNVs present a genomic overlap greater than 40%. Red represents a higher number of overlaps, while blue represents unique CNVs. Differences can be seen between the number of CNVs per chromosome and the number of recurrent and unique CNVs in CD-1 and NMRI mice.
A CNV-LP was used to depict CNV events at their respective genomic location by their copy number state (gain, partial loss or complete loss) per individual mouse sample (Figure 2.12; Appendix 2G). Eleven unique CNVRs were identified - nine in CD-1 mice and two in NMRI mice, and six CNVRs were shared in both stocks. Unique CNVRs in CD-1 mice are found on chromosomes 2, 3, 9, 10, 14 and 18, whereas unique CNVRs in NMRI mice are found on chromosomes 3 and 5. The shared CNVRs are found on chromosomes 1, 2, 5, 11, 17 and the X chromosome.
Figure 2.12: Distribution of CNV events across the autosomes in CD-1 and (n = 99) NMRI (n = 279) mice and the X chromosome in male CD-1 (n = 43) and NMRI (n = 279) mice. Different colours represent the state of CNV events, specifically complete loss (state 0; red), partial loss (state 1; yellow), and gains (states 3 and 4; blue). Unique CNVRs in CD-1 mice are found on chromosomes 2, 3, 9, 10, 14 and 18. Unique CNVRs in NMRI mice are found on chromosomes 3 and 5. The shared CNVRs are found on chromosomes 1, 2, 5, 11, 17 and the X chromosome.
2.3.5 Enrichment of immune-related and metabolic pathways and terms in genes encompassed by or overlapping a CNV and CNVRs in CD-1 and NMRI mice

To finish the characterization of CNVs in CD-1 and NMRI mice as detected in this chapter, the genic content of CNVs was analyzed. Galaxy identified 1220 genes in CD-1 mice and 1793 genes in NMRI mice encompassed or overlapped by a CNV (Appendix 2H). The most common genic CNVs and CNVRs in CD-1 mice are summarized in Table 2.6, in NMRI mice in Table 2.7, and in both CD-1 and NMRI mice in Table 2.8.

Metascape identified 40 and 68 enriched terms for CD-1 and NMRI mice, respectively. The top enriched terms for complete losses are humoral immune response \( (2.51 \times 10^{-4}) \) in CD-1 mice, and there were no enriched terms in NMRI mice. The top enriched terms for partial losses were ‘flavonoid glucuronidation’ \( (9.77 \times 10^{-23}) \) and ‘antigen processing and presentation of peptide antigen’ \( (3.89 \times 10^{-16}) \) in CD-1 and NMRI mice. The top enriched terms for gains were ‘lymphocyte-mediated immunity’ \( (9.33 \times 10^{-26}) \) and ‘phagocytosis, recognition’ \( (1.0 \times 10^{-66}) \) in CD-1 and NMRI mice.

The Gene Ontology Resource identified 114 and 201 GO terms for CD-1 and NMRI mice, respectively (Appendix 2I). The top GO term for complete loss CNVs is ‘humoral immune response’ \( (2.38 \times 10^{-4}) \) and ‘cell surface receptor signalling pathway’ \( (1.69 \times 10^{-7}) \) for CD-1 mice and NMRI mice. The top GO term for partial loss CNVs is ‘xenobiotic glucuronidation’ \( (3.55 \times 10^{-16}) \) for CD-1 mice and ‘antigen processing and presentation of peptide antigen’ \( (4.30 \times 10^{-13}) \). The top GO term for gain CNVs is ‘positive regulation of immune response’ \( (1.82 \times 10^{-20}) \) and ‘humoral immune response mediated by circulating immunoglobulin’ \( (3.52 \times 10^{-55}) \) for CD-1 and NMRI mice.

DAVID identified 27 and 64 KEGG pathways in CD-1 and NMRI mice, respectively (Appendix 2J). The most common pathways seen in CD-1 and NMRI mice in CNV events that were partial losses were metabolism pathways. In CN gains, the most common
pathways seen in both stocks were disease pathways and immune system pathways.

PANTHER identified 10 and 12 enriched terms for CD-1 and NMRI mice, respectively (Appendix 2K). The top enriched terms in CD-1 mice were ‘angiogenesis’ ($1.35 \times 10^{-3}$), ‘nicotine degradation’ ($3.86 \times 10^{-4}$) and ‘endogenous cannabinoid signalling’ ($7.31 \times 10^{-3}$) for complete losses, partial losses and gains, respectively. No terms were enriched for complete losses or gains in NMRI mice. The top enriched pathway for partial losses in NMRI mice was ‘gonadotropin-releasing hormone receptor pathway.’
Table 2.6: Most common CNVs and CNVRs in CD-1 mice

<table>
<thead>
<tr>
<th>Genomic location</th>
<th>CN state</th>
<th>Number of mice affected</th>
<th>Gene symbol (gene type)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:88205703–88306778</td>
<td>Loss</td>
<td>58</td>
<td>Hjurb (pc), Mroh2a (pc), Trpm8 (pc), Ugt1a1 (pc), Ugt1a8 (pc), Ugt1a6a (pc), Ugt1a5 (pc), Ugt1a7c (pc), Ugt1a2 (pc), Ugt1a6b (pc), Gm20528 (ncRNA), Ugt1a9 (pc), Ugt1a10 (pc), Dnajb3 (pc)</td>
</tr>
<tr>
<td>2:24657205–24825055</td>
<td>Both</td>
<td>61</td>
<td>Gm13459 (ps), Cacna1b (pc), Ehmt1 (pc)</td>
</tr>
<tr>
<td>2:71814473–71856719</td>
<td>Both</td>
<td>60</td>
<td>Itq6 (pc)</td>
</tr>
<tr>
<td>3:33800207–33804912</td>
<td>Gain</td>
<td>52</td>
<td>Ttc14 (pc)</td>
</tr>
<tr>
<td>3:121988501–122024550</td>
<td>Gain</td>
<td>55</td>
<td>Arhgap29 (pc)</td>
</tr>
<tr>
<td>5:7024671–7048239</td>
<td>Loss</td>
<td>69</td>
<td>Zfp804b (pc)</td>
</tr>
<tr>
<td>10:22148501–22404734</td>
<td>Both</td>
<td>51</td>
<td>Gm4895 (ps), Gm26528 (ps), C920009B18Rik (lncRNA), Gm5421 (ps), Gm34656 (lncRNA), Gm10825 (lncRNA), E030030106Rik (pc), Raet1e (pc), Gm1972 (ps), Gm26581 (lncRNA), Raet1d (pc), Gm34607 (lncRNA), H60b (pc), Gm26740 (lncRNA), 4930444G20Rik (pc), Gm19791 (ps), Gm7678 (ps), Gm2539 (ps)</td>
</tr>
<tr>
<td>14:81373274–81807414</td>
<td>Gain</td>
<td>69</td>
<td>Gm19311 (ps), Gm41219 (lncRNA)</td>
</tr>
<tr>
<td>17:30456718–30924641</td>
<td>Both</td>
<td>76</td>
<td>Gm9874 (pc), Gm8373 (ps), 1700097N02Rik (lncRNA), Gm24661 (snRNA), Glo1 (pc), Gm41561 (lncRNA), Dnah8 (pc), Bbvd9 (pc), H2-Ab1 (pc), H2-Aa (pc), H2-Eb1 (pc), Gplr1r (pc), Gm20513 (lncRNA)</td>
</tr>
<tr>
<td>18:57656600–57739153</td>
<td>Gain</td>
<td>89</td>
<td>Gm26038 (snRNA), Gm44491 (miRNA), Gm30636 (lncRNA)</td>
</tr>
<tr>
<td>X:169953985–169990786</td>
<td>Gain</td>
<td>90</td>
<td>G530011006Rik (lncRNA), Gm15247 (lncRNA), Gm15726 (lncRNA), Mtd1 (pc)</td>
</tr>
</tbody>
</table>

*Gene type: Protein coding (pc), pseudogene (ps), non-coding RNA (ncRNA), long non-coding RNA (lncRNA), microRNA (miRNA)
Table 2.7: Most common CNVs and CNVRs in NMRI mice

<table>
<thead>
<tr>
<th>Genomic location</th>
<th>CN state</th>
<th>Number of mice affected</th>
<th>Gene symbol (gene type)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:88205703–88306778</td>
<td>Loss</td>
<td>214</td>
<td>*Hjurp (pc), Mroh2a (pc), Trpm8 (pc), Ugt1a1 (pc), Ugt1a8 (pc), Ugt1a6a (pc), Ugt1a5 (pc), Ugt1a7c (pc), Ugt1a2 (pc), Ugt1a6b (pc), Gm20528 (incRNA), Ugt1a9 (pc), Ugt1a10 (pc), Dnahb3 (pc)</td>
</tr>
<tr>
<td>2:77875006–78021364</td>
<td>Both</td>
<td>90</td>
<td>*Gm13727 (incRNA), Gm13726 (ps), Cwc22 (pc)</td>
</tr>
<tr>
<td>3:74289954–74407866</td>
<td>Loss</td>
<td>99</td>
<td>*Gm37050 (ps), Gm6098 (ps)</td>
</tr>
<tr>
<td>3:7511220–76164831</td>
<td>Loss</td>
<td>119</td>
<td>*Platr10 (incRNA) Gm25846 (snoRNA), Gm29133 (pc), Serpini1 (pc), Gm37685 (incRNA), Gm18428 (ps), Gm37256 (incRNA), Golim4 (pc), Fstl5 (pc)</td>
</tr>
<tr>
<td>3:106114046–106309539</td>
<td>Loss</td>
<td>57</td>
<td>*Gm43709 (ps), Chil3 (pc), Chila1 (pc), Chil4 (pc), Gm6522 (ps)</td>
</tr>
<tr>
<td>4:112071427–112626691</td>
<td>Gain</td>
<td>59</td>
<td>*Gm12820 (ps), Gm12815 (ps), Gm12821 (ps), Skint4 (pc), Skint3 (pc), Skint9 (pc), Gm12819 (ps), Skint2 (pc), Gm24099 (snRNA), Gm12814 (ps)</td>
</tr>
<tr>
<td>5:3259208–3408118</td>
<td>Loss</td>
<td>99</td>
<td>*Cdk6 (pc), Gm29868 (incRNA), Gm36470 (incRNA), Gm36548 (incRNA)</td>
</tr>
<tr>
<td>5:36505037–36545077</td>
<td>Both</td>
<td>119</td>
<td>*Tbc1d14 (pc), Gm32528 (incRNA)</td>
</tr>
<tr>
<td>6:41043394–41163544</td>
<td>Loss</td>
<td>79</td>
<td>*Trbv3 (gs), Trbv5 (gs), Trbv6 (psgs), Trbv9 (psgs), Trbv16 (gs), Trbv2 (gs), Trbv4 (gs), Trbv7 (psgs), Trbv10 (psgs), Trbv17 (gs), Trbv8 (psgs), Trbv11 (psgs), Trbv12-2 (gs), Trbv12-1 (gs), Trbv13-2 (gs), Trbv13-1 (gs), Trbv14 (gs), Trbv12-3 (psgs), Trbv13-3 (gs), Trbv15 (gs), 5830405F06Rik (incRNA), Gm37610 (incRNA), Tcrb (cc)</td>
</tr>
<tr>
<td>6:130222754–130315058</td>
<td>Gain</td>
<td>66</td>
<td>*Klra10 (pc), Gm24072 (snRNA), Klra7 (pc), Klra3 (pc), Klra13-ps (ps)</td>
</tr>
</tbody>
</table>

Continued on next page
### Table 2.7 – continued from previous page

<table>
<thead>
<tr>
<th>Genomic location</th>
<th>CN state</th>
<th>Number of mice affected</th>
<th>Gene symbol (gene type)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:71142229–71285515</td>
<td>Both</td>
<td>103</td>
<td>Gm23266 (snoRNA), Nlrp1a (pc), Gm23311 (snoRNA), Gm30219 (lncRNA), Nlp1c-ps (ps), Nlrp1b (pc)</td>
</tr>
<tr>
<td>11:93158030–93160977</td>
<td>Loss</td>
<td>85</td>
<td>Gm9874 (pc), Gm8373 (ps), 1700097N02Rik (lncRNA), Gm24661 (snRNA), Glo1 (pc), Gm41561 (lncRNA), Dnah8 (pc), Bbbd9 (pc), H2-Ab1 (pc), H2-Aa (pc), H2-Eb1 (pc), Glp1r (pc), Gm20513 (lncRNA)</td>
</tr>
<tr>
<td>17:30456718–30924641</td>
<td>Loss</td>
<td>62</td>
<td>H2-T22 (pc), H2-T23 (pc), Trim71 (pc), Gm20950 (ps), Gm8810 (ps), H2-B1 (ps), Gm8815 (ps), Gm20530 (ps), H2-T10 (ps), BC023719 (lncRNA), Gm7030 (pc), Gm9574 (ps), Gm20478 (lncRNA), H2-T24 (pc), A930015D03Rik (lncRNA), 2410017I17Rik (lncRNA), H2-T-ps (ps), Gm6659 (ps), Gm10499 (ps), Gm8909 (pc), Gm17782 (ps), Gm19684 (pc), Gm6034 (pc), H2-T3 (pc), Gm6623 (ps), Gm20495 (ps), Gm8835 (ps), Gm4246 (ps), Gm20447 (ps), Gm20392 (ps), Gm10074 (ps), Gm18604 (ps), Gm20545 (ps), Gm29402 (ps), Gm6633 (ps), Gm2452 (ps), Gm20546 (lncRNA), Trim39 (pc), Rpp21 (pc), Gm5682 (ps)</td>
</tr>
<tr>
<td>17:36017266–36325814</td>
<td>Loss</td>
<td>134</td>
<td>G530011O06Rik (lncRNA), Gm15247 (lncRNA), Gm15726 (lncRNA), Mid1 (pc)</td>
</tr>
<tr>
<td>X:169953985–169990786</td>
<td>Gain</td>
<td>255</td>
<td>Gm15247 (lncRNA), Gm15726 (lncRNA), Mid1 (pc)</td>
</tr>
</tbody>
</table>

*Gene type: Protein coding (pc), pseudogene (ps), gene segment (gs), pseudogenic gene segment (psgs), complex cluster (cc), non-coding RNA (ncRNA), small nucleolar RNA (snoRNA), long non-coding RNA (lncRNA), microRNA (miRNA)
Table 2.8: **Most common CNVs and CNVRs in both CD-1 and NMRI mice**

<table>
<thead>
<tr>
<th>Genomic location</th>
<th>CN state</th>
<th>Number of mice affected</th>
<th>Gene symbol (gene type)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:88205703–88306778</td>
<td>Loss</td>
<td>272</td>
<td><em>Hjurp (pc), Mroh2a (pc), Trpm8 (pc), Ugt1a1 (pc), Ugt1a8 (pc), Ugt1a6a (pc), Ugt1a5 (pc), Ugt1a7c (pc), Ugt1a2 (pc), Ugt1a6b (pc), Gm20528 (ncRNA), Ugt1a9 (pc), Ugt1a10 (pc), Dnajb3 (pc)</em></td>
</tr>
<tr>
<td>2:77875006–78021364</td>
<td>Both</td>
<td>131</td>
<td>Gm13727 (ncRNA), Gm13726 (ps), Cwc22 (pc)</td>
</tr>
<tr>
<td>3:74289954–74407866</td>
<td>Loss</td>
<td>108</td>
<td>Gm37050 (ps), Gm6098 (ps)</td>
</tr>
<tr>
<td>3:106114046–106309539</td>
<td>Both</td>
<td>105</td>
<td>Gm43709 (ps), Chil3 (pc), Chia1 (pc), Chil4 (pc), Gm6522 (ps)</td>
</tr>
<tr>
<td>4:112071427–112626691</td>
<td>Gain</td>
<td>89</td>
<td>Gm12820 (ps), Gm12815 (ps), Gm12821 (ps), Skint4 (pc), Skint3 (pc), Skint9 (pc), Gm12819 (ps), Skint2 (pc), Gm24099 (snRNA), Gm12814 (ps)</td>
</tr>
<tr>
<td>5:36505037–36545077</td>
<td>Both</td>
<td>125</td>
<td>Tbc1d14 (pc), Gm32528 (lncRNA)</td>
</tr>
<tr>
<td>6:130222754–130315058</td>
<td>Gain</td>
<td>75</td>
<td>Klr10 (pc), Gm24072 (snRNA), Klr7 (pc), Klr3 (pc), Klr13-ps (ps)</td>
</tr>
<tr>
<td>11:71142229–71285515</td>
<td>Both</td>
<td>140</td>
<td>Gm23266 (snoRNA), Nlrp1a (pc), Gm23311 (snoRNA), Gm30219 (lncRNA), Nlp1c-ps (ps), Nlrp1b (pc)</td>
</tr>
<tr>
<td>17:30456718–30924641</td>
<td>Both</td>
<td>138</td>
<td>Gm9874 (pc), Gm8373 (ps), 1700097N02Rik (lncRNA), Gm24661 (snRNA), Glo1 (pc), Gm41561 (lncRNA), Dnah8 (pc), Btd (pc), H2-Ab1 (pc), H2-Aa (pc), H2-Eb1 (pc), Glp1r (pc), Gm20513 (lncRNA)</td>
</tr>
</tbody>
</table>

Continued on next page
Table 2.8 – continued from previous page

<table>
<thead>
<tr>
<th>Genomic location</th>
<th>CN state</th>
<th>Number of mice affected</th>
<th>Gene symbol (gene type)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>17:36017266–36325814</td>
<td>Both</td>
<td>136</td>
<td>(H2)-T22 (pc), (H2)-T23 (pc), Trim71 (pc), Gm20950 (ps), Gm8810 (ps), H2-B1 (ps), Gm8815 (ps), Gm20530 (ps), H2-T10 (ps), BC023719 (lncRNA), Gm7030 (pc), Gm9574 (ps), Gm20478 (lncRNA), H2-T24 (pc), A930015D03Rik (lncRNA), 2410017I17Rik (lncRNA), H2-T-ps (ps), Gm6659 (ps), Gm10499 (ps), Gm8909 (pc), Gm17782 (ps), Gm19684 (pc), Gm6034 (pc), H2-T3 (pc), Gm6623 (ps), Gm20495 (ps), Gm8835 (ps), Gm4246 (ps), Gm20447 (ps), Gm20392 (ps), Gm10074 (ps), Gm18604 (ps), Gm20545 (ps), Gm29402 (ps), Gm6633 (ps), Gm4252 (ps), Gm20546 (lncRNA), Trim39 (pc), Rpp21 (pc), Gm5682 (ps)</td>
</tr>
<tr>
<td>X:169953985–169990786</td>
<td>Gain</td>
<td>345</td>
<td>G530011006Rik (lncRNA), Gm15247 (lncRNA), Gm15726 (lncRNA), Mid1 (pc)</td>
</tr>
</tbody>
</table>

*Gene type: Protein coding (pc), pseudogene (ps), non-coding RNA (ncRNA), small nucleolar RNA (snoRNA), long non-coding RNA (lncRNA),
2.3.6 Mutation cluster detection and heterozygous SNP genotypes-CNV association

To explore HI in the genomes of CD-1 and NMRI mice, the association between chromosomal heterozygosity and CNVs with proximal association to heterozygous SNP loci was analyzed. A MANOVA test determined that there is a significant difference between the association in CD-1 and NMRI mice ($p = 6.18 \times 10^{-8}$). The percent heterozygosity was graphed against the percent of proximal associated CNVs per chromosome, and a regression line was fitted for each stock individually (Figure 2.13A). When the estimated coefficients were compared using a Student’s t-test, they were found to be not significantly different ($p = 0.693$; Figure 2.13A). A Spearman rank correlation coefficient analysis on CD-1 and NMRI mice combined indicated there is a positive correlation (Figure 2.13B, $R = 0.45$, $p = 0.0049$) between increased levels of heterozygous SNP loci and proximal CNVs, identifying evidence supporting HI in the genomes of CD-1 and NMRI mice.

Figure 2.13: Statistical analysis of the occurrence of proximal associations per autosome and autosomal heterozygosity. A. Distribution of proximal CNVs per chromosome and average chromosomal heterozygosity for CD-1 and NMRI mice. B. Spearman rank correlation coefficient analysis of chromosomal heterozygosity and proximal CNVs in CD-1 and NMRI mice indicates a positive correlation. The value of the correlation coefficient ($R$) can vary between -1 and 1, with 0 indicating no association, -1 indicating a strong negative correlation, and 1 indicating a strong positive correlation.
To examine the heterozygous landscape at SNP loci in both outbred mouse stocks, rainfall plots were created. Rainfall plots show clusters of heterozygous SNP genotypes in the genomes of both CD-1 and NMRI mice (Figure 2.14), with larger clusters of heterozygous SNP loci in CD-1 mice and larger deserts in NMRI mice. Regions of clustered heterozygous SNP loci are separated by regions of clustered homozygous SNP loci, creating a discontinuous landscape of heterozygosity. Different patterns of discontinuous landscapes are found across different chromosomes (Figure 2.14) and within the same chromosome (Figure 2.15) when compared within stocks and between stocks.

**Figure 2.14:** Visualization of heterozygous SNP genotypes across the genome through chromosomal position and distance between consecutive heterozygous SNPs. Comparison of the distribution of heterozygous SNP genotypes in CD-1 mice (A, B, C) and NMRI mice (D, E, F) on chromosomes 3 (A, D), 5 (B, E) and 11 (C, F). All chromosomes display a discontinuous landscape of heterozygosity, with differences in the landscape between chromosomes and outbred mouse stock. A discontinuous landscape is characterized by regions of clustered heterozygosity separated by regions of clustered homozygosity. Rainfall plots display heterozygous SNPs at their respective genomic location (X-axis) against the distance from their neighbouring heterozygous SNP plotted on a logarithmic scale (Y-axis). SNP loci shown in these rainfall plots are only those assayed by the array and heterozygous.
Figure 2.15: The discontinuous landscapes of heterozygosity at SNP loci on chromosome 17 in six different mice show the inter-animal and inter-stock differences in landscapes, ideal for studying HI. Comparison of the discontinuous landscapes of heterozygous SNP loci in CD-1 mice (A, B, C) and NMRI mice (D, E, F) on chromosome 17. A discontinuous landscape is characterized by regions of clustered heterozygosity separated by regions of clustered homozygosity. When comparing rainfall plots it is clear to see there are differences in consecutive heterozygosity and homozygosity within CD-1 mice (A, B, C) and NMRI mice (D, E, F) and between the landscapes in each stock. Differences include number of heterozygous clusters and size of homozygous regions interspersed between heterozygous clusters. Rainfall plots display heterozygous SNPs at their respective genomic location (X-axis) against the distance from their neighbouring heterozygous SNP plotted on a logarithmic scale (Y-axis). SNP loci shown in these rainfall plots are only those assayed by the array and heterozygous.

To examine the spatial association between heterozygosity and CNVs and test for indirect evidence supporting HI, rainbow plots and J statistic plots were used. Rainbow plots show a potential association between heterozygous SNP genotypes and CNVs, which was confirmed using the J function (Figure 2.16). All rainfall, rainbow and J statistic plots for CD-1 and NMRI mice can be found in Appendix 2L. CD-1 and NMRI mice have 41.0 and 21.8% of chromosomes that contained CNVs displaying heterozygous SNP genotypes with proximally associated CNVs. Chromosome 17 contains more CNVs with
proximal spatial distribution than either both or distal spatial distributions in CD-1 and NMRI mice. Chromosome 4 in CD-1 mice and chromosome 6 in NMRI mice also contains more CNVs with the proximal distribution. The majority of CNV events on chromosomes 5, 8, 9, 10, 12, and 13 in CD-1 mice and chromosome 14 in NMRI mice have proximally distributed heterozygous SNPs. Overall, indirect evidence supporting HI in the genomes of CD-1 and NMRI mice is detected across multiple autosomes.

Figure 2.16: Visualization of the proximal association of heterozygous SNP genotypes and CNVs on the autosomal chromosomes in CD-1 (top row) and NMRI (bottom row) mice. Rainbow and J statistic plots were generated to display the heterozygosity and CNVs on a single chromosome. Rainbow plots display the CNVs in red dots along the X-axis of the graph at their respective chromosomal position, whereas black circles represent heterozygous SNPs. The closer these black dots are plotted to the x-axis, the closer they are to the CNV along the genomic sequence (proximal). J function statistically analyzes the spatial distribution of the heterozygous SNP genotypes to the closest CNV. After running 1000 simulations, J statistic produces blue confidence bands, and if the J function (black line) crosses the lines, the result is significant.
2.4 Discussion

This chapter presents indirect evidence of HI in the genomes of CD-1 and NMRI mice and identifies the need for further study to confirm HI and study potential mechanisms. Additionally, this chapter presents the most extensive characterization of structural variants and single nucleotide polymorphisms in two outbred mouse stocks, CD-1 and NMRI. Pairwise comparisons of genetic distances were calculated for all mice using heterozygous SNPs or CNVs. CNVs were identified and characterized by location, state, length, recurrence, genic impact and pathway enrichment. The levels of heterozygosity at SNP loci were analyzed in both stocks, and the discontinuous landscape of clusters of heterozygosity was detected. Association between clusters of heterozygosity and CNVs was detected in both CD-1 and NMRI mice, highlighting the need to analyze the nature and distribution of the clusters of heterozygous SNP loci to explore possible mutational mechanisms associated with deletions or duplication creating CNVs.

2.4.1 Levels of heterozygosity at SNP loci in CD-1 and NMRI mice more closely represent heterozygosity in the human genome

Both CD-1 and NMRI mice have a higher average level of heterozygosity than inbred mice, with C57BL/6 mice displaying average heterozygosity of 0.6% per mouse. Comparatively, CD-1 and NMRI mice better represent the level of heterozygosity seen in humans. In humans, the average heterozygosity at all SNP loci is between 24.6-28.5% [38], which is approximately three times greater than the average heterozygosity in CD-1 mice and approximately five times greater than the average heterozygosity in NMRI mice (Table 2.1). Additionally, pairwise comparisons calculated from SNP genotypes in 344 humans are between 0.002 and 0.109 [53], which is similar to the SNP genetic distances seen in
Chapter 2. Characterization of two stocks of outbred mice

CD-1 (0.009–0.180) and NMRI (0.008–0.110) mice, further indicating that CD-1 and NMRI mice are good models of the genetic variation seen in the human population, and much more representative models.

2.4.2 SNP genetic distance better represents relatedness between mice

CD-1 and NMRI mice both had SNP trees that well represented genetic background and genetic relatedness, but the CNV trees did not accurately represent the relatedness displayed in the SNP trees. SNP trees accurately depicting relatedness, with relatedness being lost in CNV trees, has been previously reported in inbred mice [20, 54]. This relationship is visualized by comparing the coloured subset of samples in the neighbour-joining trees for SNP and CNV genetic distances. In both SNP trees, the selected samples share a recent common ancestor and are thus on the same branch of the tree. Looking at the CNV tree, it is evident that these samples are no longer close to each other on the tree, and hence the aspect of relatedness appears to be lost. Considering the Mantel test showed no significant difference between the SNP and CNV trees, relatedness is not entirely lost but confounded by other factors. This disparity could be due to de novo CNVs, which would confound the relatedness of individuals [20]. Technical error may also act as a confounding factor. There are eight probes to call each SNP and detect either the reference/major (A) or alternative/minor (B) allele. A duplicate probe for both the sense and antisense strand creates redundancy, and redundant probes are placed at different locations on the microarray. SNP genotype calls are based on the relative intensity of one probe location, represented by multiple probes on the microarray, creating high redundancy. In contrast, CNV calls are based on the increase or decrease of intensity at three consecutive probes compared to a reference [20, 31]. This leaves more possibility for technical error, making SNP calls more reliable than CNV calls [20, 54] and this difference indicates SNPs are a much more accurate method of tracking inheritance.
between individuals.

### 2.4.3 CD-1 and NMRI mice CNV profiles reported here are comparable to reported trends in humans, and in inbred, wildtype and wild-derived mice

Overall, the trends in structural variants seen in this study are comparable to trends seen in humans and trends reported by Locke *et al* (2015), but there are notable differences. CD-1 mice have an average of 17 CNVs per mouse, and NMRI mice have an average of 14 CNVs per mouse, which puts them both within the normal range seen in humans - 12-20 CNVs per human [14, 55]. The ratio of losses to gains of 0.84, 2.25 and 2.92 are reported for classical inbred, wild-derived and wild-caught mice [20]. CD-1 mice display a ratio similar to classical inbred mice. In contrast, NMRI mice display a ratio approximately three times higher than wild-derived and wild-caught mice (Table 4). A similar frequency of losses was seen in a population affected by founder effects and subsequent inbreeding. This founder population had losses as 70% of their CNV events [55]. NMRI mice may have experienced genetic drift associated with small population size when transported to America and subsequently inbred for 51 generations (Figure 2.1).

Inbred, wild-derived and wild-caught mice contained CNVs that affected 8.15% of the genome [20], which is similar to the levels observed in CD-1 mice (9.2%) but higher than the levels in NMRI mice (4.2%). This difference could be explained because NMRI mice display more CN losses, which tend to be shorter in length due to their higher probability of being more deleterious than gains. Comparatively, one study reported that 12.0% of the human genome was affected by copy number variable regions [17]. As previously reported in inbred, wild-derived, and wild-caught mice, most CNVs ~ 80.7% in CD-1 mice and 89.6% in NMRI mice, entirely encompass a gene or partially overlap with a gene [20]. Similar to what is reported in inbred mice [17, 20, 56], CD-1 and NMRI mice displayed a
lower proportion of CNVs that overlapped or encompassed genes when the CN state was a partial or complete loss. In humans, it was found that the majority of CNVRs detected encompassed or overlapped a gene or functional sequences [17].

As previously reported, chromosome length did not correlate with the number of CNVs seen per chromosome [20, 57], which may be due to different levels of recombination events and mutation rates across the genome [2]. Recombination hotspots occur throughout the genome, and thus, some chromosomes may have more recombination events occurring, leading to more opportunities for CNVs to occur [58].

The average length of CNVs reported in this study is higher than what research has previously reported. Classical inbred, wild-caught and wild-derived mice had an average length of 54,037 base pairs [20]. Similarly, a population affected by founder effects had CNVs with an average length of 52,390 base pairs [55], measured using the same technologies used in this study. The average length of CNVs is approximately 10,000 base pairs longer in NMRI mice and approximately 30,000 base pairs longer in CD-1 mice. As observed previously in inbred mice [17, 20], the average length of deletions was shorter than the average length of gains.

Overall, the trends in CNVs seen in CD-1 and NMRI mice here are comparable to trends seen in inbred, wild-derived and wild-caught mice. NMRI is unique in that the trends seen in CNVs in their genome more closely represent trends seen in the genome of a human founder population, which differs from the CNV profiles in CD-1, inbred, wild-derived and wild-caught mice.

2.4.4 Cataloguing unique and recurrent CNVs and copy number variable regions

The majority of CNV events are recurrent, whereas only a small fraction are unique CNVs. This result was previously observed in inbred mice [20]. CNVRs were detected across the genome, similar to CNVR distribution in the human genome [17]. This study
reports 18 novel CNVRs and eight CNVRs that were previously reported in inbred [20, 54], wild-derived and wild-caught mice [20], with one of these CNVR overlapping a CNV which has been previously detected in CD-1 mice [30]. The most common CNVR was detected at X:169,953,985–169,990,786 in both CD-1 and NMRI mice, followed by the CNVR at 1:88,205,703–88,306,778. All CNVRs and recurrent CNVs encompassed or overlapped with one or more genes, except for one region on chromosome 3. The CNVR at 17:36,017,266-36,325,814 in NMRI mice overlapped or encompassed the most number of genes (Table 2.8). Multiple of the genes that are encompassed or overlapped by the CNVR on chromosome 17 are histocompatibility genes. Histocompatibility genes are highly recombinant, which creates increased opportunity for CNV formation and leads to a copy number variable region [18, 57]. Four CNVRs identified in CD-1 and NMRI mice were previously reported in inbred, wild-caught and wild-derived mice [20] at 1:88205703-88306778, 4:112071427–112626691, 11:71047723–71124806, and 17:30456718–30924641. The CNVR detected at 4:112071427–112626691 overlaps Skint4, Skint3, and Skint9, which are genes that are associated with anxiety and have been previously detected in CD-1 mice [30]. CNVs have been previously reported in inbred mice affecting the Klra locus, Raet1 genes [54], which were detected in CD-1 and NMRI mice (6:130222754–130315058) and CD-1 mice (10:22148501–22404734). Finally, the CNVRs detected at 1:88,205,703–88,306,778 and 2:71,814,473–71,856,719 were previously reported to be found uniquely in wild-caught M. m. musculus and M. m. domesticus, respectively [20].

2.4.5 Enrichment of pathways and terms that respond to a changing external environment in genes encompassed by or overlapping a CNV and CNVRs

For both CD-1 and NMRI mice, the majority of enriched terms were immune-related or metabolic terms. Metascape, The Gene Ontology Resource, and DAVID identified
disease and immune system pathways enriched in genes involved in CN gains in CD-1 and NMRI mice, CN partial losses in NMRI mice, and CN complete losses in CD-1 mice. Laboratory mice are bred to study disease phenotypes, possibly explaining why immune system pathways and GO terms are enriched among the CNVs. Classical inbred mice showed the same enrichment [20]. In another study of inbred mouse strains, the top enriched GO terms determined from genes overlapped by CNVs were pheromone receptor and olfactory receptor-related functions, followed by defence response-related terms and antigen-binding, processing and presentation related terms [54]. While an enrichment in olfactory receptor-related functions was not seen in CD-1 or NMRI mice, the enrichment of defence response-related genes was mirrored in CD-1 and NMRI mice, and antigen-binding, processing and presentation-related terms were mirrored in NMRI mice. In humans, an enrichment in immune-related genes overlapped by CNVs was reported in previous studies [14, 56].

In CD-1 mice, the top Metascape term and GO term for CN partial losses were ‘flavonoid glucuronidation’ and ‘xenobiotic glucuronidation,’ both metabolic terms. In both CD-1 and NMRI mice, the top enriched terms for CN partial losses identified using DAVID were metabolic terms. This enrichment in the partial loss of metabolic terms may represent their adaption to a laboratory diet. Comparatively, a study that found wildtype mice had enriched metabolism pathways in CN gains, whereas classical laboratory strains did not [20]. Considering wildtype mice have a more varied, adaptive diet than inbred and outbred laboratory mice, it is plausible that wildtype mice have needs for a greater variety of metabolic functions and pathways. Interestingly, the phenotypes conferred by genetic variants are selected for in response to changing external environments.
2.4.6 Discontinuous clusters of heterozygosity show proximal spatial association with CNVs

Similar to the association reported in *Arabidopsis* and peach plants, this chapter demonstrates that in both CD-1 and NMRI mice, there are examples of an elevated occurrence of CNVs near regional areas of heterozygosity. These examples present indirect evidence of HI. MANOVA analysis confirmed a difference between increased heterozygosity and proximal CNVs in CD-1 and NMRI mice. A Spearman rank correlation coefficient confirmed a positive correlation between increased heterozygosity and increased numbers of proximal CNVs. Interestingly, when analyzing the proximal associations per autosome by autosomal heterozygosity for CD-1 and NMRI mice individually, the distribution showed no correlation. However, when combining data for CD-1 and NMRI mice, a positive correlation became clear. This indicates that a large range of heterozygosity may be needed to assess HI accurately. The proximal distribution of heterozygous SNP genotypes and CNVs was seen globally across the genome. Ten chromosomes displayed more proximal distribution between the two stocks than distal or both, which shows that no chromosome is a hotspot for Heterozygote Instability. The only chromosome that displayed more proximal distribution in both stocks is chromosome 17. Chromosome 17 is where the major histocompatibility complex is located. These genes are responsible for immune responses and antigen processing and presentation [59]. Duplications and deletions are common in the major histocompatibility complex to diversify antigen presentation [59]. Considering the proximal distribution of CNVs to localized clusters of heterozygous SNP loci was detected in both CD-1 and NMRI mice, this indicates support for HI, and thus, the nature of these clusters should be closely analyzed. Considering how many mechanisms contribute to the formation of CNVs, it is intuitive that CNVs occur in other locations across the genome. The fact that evidence was detected for a proximal association of CNVs with heterozygous regions highlights the need to further
study potential mutational mechanisms of HI and attributes of heterozygous clusters.

Rainfall plots for both mouse stocks showed discontinuous clusters of heterozygous SNP loci across chromosomes, with uneven distribution in CD-1 and NMRI mice. NMRI mice have larger deserts of heterozygous SNP genotypes because of their 51 generations of inbreeding, which introduced more homozygosity into their genome. CD-1 mice were only inbred for 15 generations, and thus CD-1 mice have genomes with larger clusters of heterozygosity [26]. Rainbow plots, along with J function analysis, showed CNVs are associated with these clusters.

Different levels and patterns of heterozygosity can be seen at the genome-wide level, the chromosome-specific level, and at regional areas of heterozygosity. Overall, CD-1 mice have a higher level of genomic heterozygosity at SNP loci and have a greater average number of CNVs per mouse, 17 CNV calls per mouse in CD-1 mice compared to 14 CNV calls per mouse in NMRI mice. Chromosomes with CNVs located proximally to heterozygous SNP loci tend to have higher heterozygosity, ranging from 7.8–11.0% and 4.3–9.3% in CD-1 and NMRI mice, respectively. Future studies need to closely examine the characteristics of regional areas of heterozygosity, including density, length, cluster patterns, and the number of regions per chromosome. This characterization will allow more in-depth study into the phenomenon first reported in *Arabidopsis* [31] and the potential association between heterozygosity at SNP loci and CNVs in CD-1 and NMRI mice.

Potential mechanisms for HI-induced mutations differ during meiosis and mitosis. Potential mechanisms for HI during recombination and meiosis are active as regions of heteroduplex DNA forms. Repair mechanisms may misidentify heterozygous pairs in the regions of heteroduplex DNA as DNA mismatches and repair them through mismatch mediated repair. This repair leads to crossovers and is prone to error and slippage, introducing mutations like CNVs at heterozygous loci [60]. Additionally, heterozygous sites can lead to poor pairing during heteroduplex DNA formation, leading to *Spo11-*
mediated double-strand breaks. These double-strand breaks are repaired by non-allelic homologous recombination, which leads to crossovers with a similar strand of DNA. With crossovers, proximal repeat sequences can lead to misalignment and formation of CNVs [6]. A potential mechanism for HI during replication in mitosis and meiosis is fork stalling and template switching (FoSTeS) related to heterozygosity at SNP loci [61]. Occasionally, there is decreased coupling between heterozygous loci, which may cause the polymerase and the replication fork to stall. As the polymerase switches templates to similar region on a different strand of DNA to avoid the stalled location, insertions, deletions, or duplications are possible [7]. Additionally, as these mistakes are repaired, there are extra rounds of replication and increased replication stress which creates opportunity for other mistakes or slippage.

2.4.7 Study limitations

All evidence supporting HI in this chapter was indirect and associative rather than a direct test or confirmation. This evidence is a launch point for future researchers exploring HI in outbred mouse models and to confirm the presented indirect, associative evidence. Future studies can do this by performing controlled breeding experiments to track the landscape of heterozygosity and CNV occurrence through multiple generations and characterize the regions of heterozygosity where CNVs are occurring proximally.

Additionally, alternative mechanisms of CNV formation act as confounding factors when interpreting the potential evidence of HI proposed in this chapter. CNVs may be co-localizing near clusters of heterozygous SNP loci, but the heterozygosity may not be contributing to the development of CNVs. Other mechanisms could be causing CNV formation near heterozygosity, and future studies must characterize the locations where CNVs are occurring next to clusters of heterozygosity to elucidate if heterozygous loci contribute to CNV mechanisms.

When using the mouse diversity genotyping array or other microarray technologies,
the heterozygous landscape detected is only a sampling of the total heterozygosity within
the genome. While the SNP probes are representative across the genome, there is no
inter-probe heterozygosity information meaning there could be heterozygosity between
the probes that went undetected in this study. Additionally, calling CNVs with probes
can lead to false positive and false negative calls, which can be determined through
confirmation assays. Future studies can improve these limitations by using a method with
more extensive coverage like whole-genome sequencing, which would provide the entire
heterozygous landscape of these mice and allow for CNV detection. This method would
be much more costly than using microarray technology. Additionally, CNVs could be
detected using array comparative genome hybridization.

The greatest limitation of the data in this chapter is that by using publicly available
data, there is no tissue available for the samples used to generate the CEL files accessed
for analysis, and thus it is not possible to confirm mouse-specific CNVs and landscapes
of heterozygosity. While this chapter provided indirect confirmation of specific CNVRs
detected in CD-1 and NMRI mice through previously reported CNVRs in CD-1 mice,
inbred mice, wild-derived mice and wild-caught mice, future research can confirm strain-
specific CNVs and CNVRs of greatest interest identified using the CNV-LP through
available CD-1 and NMRI outbred mice. To reconstitute a population study of mutation
landscapes researchers must repeat the population study with CD-1 and NMRI mice.
Tissues could be extracted from this study for internal validation of CNVs detected in
these samples and applied to CNVs detected in this study, confirming strain-specific or
population-specific CNVs.

2.4.8 Future directions

Future research should perform controlled parent-progeny breeding studies with CD-1 and
NMRI mice to track Heterozygote Instability and germline mutations through meiosis in
both male and female gametes over multiple generations of mice. Over the whole genome,
there will be differences in clusters and deserts of heterozygosity, and by using both CD-1 and NMRI mice, there will be a large enough spread of heterozygosity to assess for correlation between clusters of heterozygosity and the occurrence of CNVs. Researchers will then be able to use DNA from the controlled breeding experiment to study the nature of the heterozygous clusters - size, density, occurrence, the distance between clusters, and how different clusters co-localize more frequently with CNV occurrence. Defining a profile for the heterozygous regions where CNVs occurred more frequently, combined with an examination of sequence context and proximal heterozygosity landscape will help elucidate possible mechanisms for CNV occurrence due to HI. Using next-generation sequencing in combination with array comparative genomic hybridization to study this complex phenotype at a single base resolution will allow an in-depth definition of heterozygous clusters, detection of CNVs and detection of other de novo mutations. Additionally, future studies must perform validation assays to confirm the CNVs and heterozygous regions within CD-1 and NMRI mice. These validations would be performed through digital droplet PCR and sequencing using CD-1 and NMRI mouse samples to detect reproducibility. The goal would be to reproduce what was seen in the publicly available dataset by detecting strain-specific CNVs and CNVRs. Although the exact mouse samples are not available for confirmation, other mice in the stock will share similar landscapes of discontinuous heterozygosity and CNV occurrence. Finally, this experimental plan looks to assess HI during meiosis in germ cells. Future research must look for evidence of HI in somatic cells, including stem cells and cancer cells, to evaluate associations of this phenomenon with typical development, stem cell tissue regeneration and cancer. Once evidence indicates a possible mechanistic location, repair-deficient mice could be used to elucidate possible mechanisms of HI-induced mutations.
2.5 Conclusion

This chapter provides examples of an increase of CNVs at regions of localized heterozygosity, supporting the Heterozygote Instability hypothesis through indirect evidence. This chapter provides the most extensive characterization of CNVs in CD-1 and NMRI mice to date, with novel reporting of previously unidentified CNVs and CNVRs. The abundance and diversity of CNVs within and between CD-1 and NMRI mice were characterized and compared to inbred, wild-derived and wild-caught mice. CD-1 mice had a much higher ratio of CN gains, whereas NMRI mice had a much higher level of CN partial losses. While this may be due to founder effects and population bottleneck, it is still important for researchers to consider these differences when using CD-1 and NMRI mice to study complex phenotypes. Differences in the CNV landscape of CD-1 and NMRI mice were identified, including differences in locations of recurrent CNVs, which have different genic effects. This catalogue of CNVRs will be an essential resource for researchers using these mice as a proxy for complex phenotypes seen in humans, given the genic nature of CNVs and their potential relevance to phenotypic studies. This chapter provides a characterized landscape of heterozygosity at SNP loci, and as predicted, there are clusters and deserts of heterozygosity creating discontinuous landscapes of heterozygosity in the genome. Differences were seen in the discontinuous landscapes of heterozygosity in CD-1 and NMRI genomes, with CD-1 mice having greater clusters of heterozygosity, as predicted based on breeding records. This chapter identified the tendency for CNVs to co-localize near clusters of heterozygosity. This chapter uses new visualization and statistical tools to portray CNV landscapes and portray and assess spatial distribution and association. Proximal association of CNVs to heterozygosity is demonstrated in both stocks but more strongly in CD-1 mice, consistent with the breeding history of CD-1 mice. There are larger, more prominent clusters of heterozygous SNPs in CD-1 mice due to being inbred for a shorter number of generations, which may affect HI and its effect
on CNV occurrence. Attributes of heterozygous clusters should be characterized in the future to study this possibility and potential mechanisms further.
References


REFERENCES


[33] Luo, B. [2018], Statistical tools for assessment of spatial properties of mutations observed under the microarray platform, PhD thesis, The University of Western Ontario.


Chapter 3.

Toward characterization of genetic variation in mouse models of basal cell carcinoma

3.1 Introduction

Evidence suggests that Heterozygote Instability (HI) is a new mutational mechanism occurring in meiosis in germ cells. Evidence for HI occurring in mitosis in somatic cells has not been examined. HI has been demonstrated in germ cells by analyzing heterozygosity at single nucleotide polymorphism (SNP) loci and small insertions and deletions in Arabidopsis [1] and peach plants [2] and Chapter 2 presented additional indirect evidence supporting HI in the genomes of two outbred mouse stocks. Notably, HI has not been studied in somatic cells and cases of high cell division. Two cases of somatic cells with high genetic variation and either typical or atypical cell division are stem cells and cancer. Recent research has identified genetic instability due to heterozygous mutations in Saccharomyces cerevisiae cancer models and proposed that similar mechanisms may contribute to genetic instability in human cancer cells [3]. This observation implicates cancer models as a helpful model for assessing HI with somatic, acquired mutations and a novel, unstudied potential mechanism for genomic instability and mutation accumulation in cancer cells.
3.1.1 **Genetic variation is high and dynamic in the epidermis**

The normal epidermis has high levels of genetic variation. Researchers tracked somatic variation in the DNA sequence as keratinocytes proliferated and differentiated, using a single nucleotide polymorphism (SNP). Individuals with homozygous germline alleles displayed both homozygous genotypes and the heterozygous genotype sequence in skin surface scrapings. Additionally, the epidermal samples taken from the same location showed variation in the sequences at the SNP loci at different times. Thus, the somatic mutation burden in the epidermis is highly dynamic over time, changing horizontally across the epidermis and vertically up from the dermal to epidermal layers as the cells divide (Figure 3.1A) [4], making the epidermis ideal for the study of genetic variation in a model of high dynamic genetic diversity. In humans, the skin is composed of three major layers, the epidermis, dermis and hypodermis (subcutaneous layer), with the epidermis stratified into five layers (Figure 3.1B). Keratinocyte stem cells originate in the stratum basale and proliferate into differentiated keratinocytes throughout the stratum spinosum, stratum granulosm, stratum lucidum, and stratum corneum. Terminally differentiated keratinocytes reside in the stratum corneum, where they are continually shed and are replaced by differentiated cells as they are pushed to the surface.

Oddly, recent research found that aged, UVB-exposed eyelid epidermal tissues have a \textit{de novo} mutation burden with UVB signature mutations comparable to squamous cell carcinoma, basal cell carcinoma and melanoma and positive selection of mutations in known cancer driver genes, yet no malignant transformation [5]. To study a tissue with a similar structure, yet no UVB exposure to act as a mutagen, researchers examined esophageal epithelium tissue [6]. Interestingly, more cancer-associated mutations had significant positive selection in esophageal tissue than UVB-exposed eyelid epidermal tissue. The mutation burden displayed signatures of intrinsic mutational processes and shared mutations in known esophageal cancer driver genes, yet no detectable malignancy.
Chapter 3. Mutagenesis in models of BCC

Figure 3.1: Dynamics of genetic variation and structure in the skin. A. Mutation burden in the skin is highly dynamic, and it is changing horizontally across the epidermis, throughout the depth of the epidermis, vertically through the dermal layers and over time (four dimensions). B. The epidermis, dermis and hypodermis (subcutaneous) layers make up the skin, with the epidermis having five unique layers. Created using BioRender.com.

occurred [5, 6]. The combined observations that were seen in the aged, UVB-exposed epidermal tissue and esophageal epithelium tissue highlight the high level of genetic variability that is seen in human skin without cancer.

3.1.2 High genetic variation in cancer

Hypotheses regarding genetic variants and cancer initiation focus on the accumulation of post-zygotic, acquired mutations in the same cell, including driver gene mutations, which confer a growth advantage to the cell and passenger gene mutations, which confer
less or no apparent growth advantage [7, 8]. The aggregate of accumulated mutations in a cell is termed the mutation burden. The mutation burden can provide certain traits to a cell contributing to carcinogenesis. These traits or hallmarks include: self-sufficiency in growth signals, evading growth suppressors, limitless replicative potential, evading apoptosis, angiogenesis, tissue invasion and metastasis, genome instability, evading immune destruction, altering cell metabolism, and tumor enhanced inflammation [9, 10].

As a cell acquires mutations that give a phenotypic advantage over other cells, it will produce daughter cells with the same advantageous mutations, called clonal expansion. These daughter cells will compete with other cells in the surrounding microenvironment, with a strong positive selection for the most advantageous phenotypes. Positive selection of phenotypes conferred by mutations can be inferred by detecting clonal expansion using sequencing data, specifically detecting high variant allele fraction with read-depth analysis of next-generation sequencing data [5, 6].

### 3.1.3 BCC as a case for the highest genetic variation

Basal cell carcinoma (BCC) is a skin cancer that originates from stem cells in the hair follicle bulge of the epidermis [11, 12]. BCC provides a unique opportunity to study HI and other complex phenotypes, considering BCC has one of the highest mutation burdens of human cancers yet is one of the most benign [5]. Contributing to the genetic variation in BCC is the genetic instability of carcinogenesis and the high genetic variation inherent to skin. The mutation burden in basal cell carcinomas is at least 25 to 150 somatic mutations/Mb/cell, compared to anywhere between 1 to 380/Mb in squamous cell carcinoma, and 0.5 to 200/Mb in melanoma [5]. There is a larger range of mutation burdens in squamous cell carcinoma and melanoma, but comparatively, BCC has a higher average mutation burden [5]. A high mutation burden leads to a discontinuous landscape of mutation burden with hotspots in many genes across the genome and positive selection for advantageous phenotypes conferred by mutations. The landscape of heterozygosity has
not been documented and must be characterized. In addition to studying the mutational mechanisms, like HI, the combined observations in BCC, aged UVB-exposed epidermal tissue, and esophageal epithelium tissue predict factors in addition to mutations per se are required for tumor initiation and aggression. These observations highlight the need for a characterized BCC mouse model to study somatic mutation burdens and specific mutations previously labelled as driver mutations in more depth to understand the tumorigenic potential of mutations and mutation burdens.

BCC is the most frequently occurring human cancer and the most common skin cancer, accounting for 75–85% of all skin cancers [13, 14]. BCC incidence is increasing due to the increased use of tanning beds and increased sun exposure. BCC incidence is rising five times faster than other cancers, regardless of ethnicity [15, 16]. UVB is the primary mutagen; thus, risk increases with prolonged sun exposure and BCC generally occurs on sun-exposed skin, including the face, hands and forearms [17]. BCC is locally invasive and rarely metastasizes but recurs at a rate of 50% per five years [18, 19]. Most treatments are non-specific procedures, including surgery, cryosurgery, photodynamic therapy, radiation therapy, and topical agents [14]. Due to the locally invasive nature of BCC tumors and current surgical treatment options, there is often an increase in patient morbidity [12]. Prevention methods are limited to sun avoidance and physical barriers, as there is currently little evidence to support sunscreen decreasing the incidence of BCC [20, 21].

Research has determined aberrant activation of the Hedgehog (Hh) pathway is responsible for BCC initiation [11, 22–26]. The Hh pathway plays a role in embryonic development and tissue homeostasis in adults (Figure 3.2). Hh signaling is very active during development — regulating cell differentiation, specification and proliferation, and then activity decreases with age [27, 28]. In mammals, there are three Hh-related proteins: Sonic hedgehog (Shh), Indian hedgehog and Desert hedgehog. In the skin, Shh signaling is responsible for maintaining the stem cell population and regulating hair follicle
Hedgehog (Hh) pathway: Key oncogenic pathway in BCC. Hh pathway mediated by \(PTCH1\) inhibition of \(SMO\). When \(SHH\) is absent, \(PTCH1\) binds to and inhibits \(SMO\), leading to repression of pathway activation. Active \(SHH\) binds to \(PTCH1\), preventing \(PTCH1\) binding to \(SMO\), which enables \(SMO\) to be transported to the cillum. \(SMO\) promotes the release of GLI transcription factors, which are transported to the nucleus and activate transcription. Target genes of the Hh pathway include cell cycle regulators \(Cyclin\ D/E\), and \(Myc\), and \(Gli1,\ Ptch1,\ and\ HIP\).

Mutations leading to constitutive activation of Hh signalling are crucial for BCC initiation, causing hair follicle disruption and uncontrolled proliferation [11, 22, 29]. These mutations include gain of function mutations in \(PTCH1,\ SUFU,\ and\ GLI3\), or loss of function mutations in \(SMO,\ SHH,\ GLI1,\ GLI2,\ and\ KIF7\), leading to constitutive activation [22]. The majority of BCC tumors contain somatic mutations in \(PTCH1\), the major repressor in the Hh pathway. Somatic mutations in \(SMO,\ SUFU,\ and\ TP53\ are
known to be important in BCC initiation [23, 26, 28, 30, 31]. While these genes are
the most frequently mutated, research has reported a wide variety of additional genes
mutated in BCC tumors, predicting that mutations in these driver genes are not the only
factor contributing to BCC initiation [23, 31]. A large majority of mutations in these
genes bear the mutation signature of UVB damage, a C to T transition at a dipyrimidine
sequence or a CC to TT double transition [17, 31–33].

While sporadic BCC is most common, there are inherited diseases that increase
predisposition to BCC tumorigenesis. Gorlin syndrome or nevoid basal cell carcinoma
syndrome (NBCCS) is an autosomal dominant disease characterized by the development
of multiple BCCs early in life. NBCCS is primarily caused by mutations in PTCH1,
including frameshift, missense, nonsense, insertions, deletions, and splice-site mutations
[34–36]. Xeroderma pigmentosum (XP) is an autosomal recessive disease characterized by
the development of skin cancers — basal or squamous cell carcinoma, by eight years of
age. Mutations in nucleotide excision repair genes responsible for repairing UV damage
are the primary cause of XP [37, 38].

3.1.4 Mouse models of BCC

Multiple mouse models accurately represent human BCC tumors. The first BCC mouse
models overexpressed Shh [39] or Smo [40], both of which were sufficient for BCC
tumorigenesis in mice. Ptch1+/− mice develop small skin tumors that resemble human
BCC tumors. There is a shift in histologic features with UV radiation, resulting in tumors
that more closely resemble human BCCs and grow more rapidly [41]. Mice overexpressing
Gli1 or Gli2 form multiple sporadic skin tumors that were histologically similar to human
BCCs [11, 42, 43]. Sufu+/− mice develop a skin phenotype similar to those observed
in patients with Gorlin syndrome and had aberrant activation of the Hh pathway [44].
There are now multiple Ptch1 knockout and conditional knockout models used to study
preclinical BCC treatment drugs [45]. Considering PTCH1 mutations in human tumors
are the most common driver mutations and Ptch1 mouse models most accurately match the morphology of human BCC [45]. Ptch1 mouse models are the best choice for mouse models representing human BCC tumors. The Jackson Laboratory has five strains of mice listed as ‘useful for studying cutaneous basal cell carcinoma’: heterozygous Ptch1 mice (ptc-lacz), overexpressing of human GLI1 (Tg(tetO-GLI1)10Rup), constitutive expression of Smo (R26SmoM2), overexpression of Shh (ShhcreER\textsuperscript{T2}), and a model targeting K14, one of the potential initiating cells of BCC and squamous cell carcinoma (K14CreERT). Charles River offers knockout mice for Ptch1 and Sufu, through collaboration with genOway. Taconic Biosciences has an XP mouse model susceptible to basal cell carcinoma, squamous cell carcinoma and melanoma formation after UV exposure. It is clear that there are mouse models of BCC that closely mimic the morphology of human BCC. However, it is unclear whether the genetic variation in these models represents the same extent and dynamic seen in human BCC.

3.1.5 Similarities and differences between human and mouse skin physiology

Similarities and differences between mouse and human skin physiology must be considered when evaluating mouse skin to represent the human skin. Like human skin, mouse skin is composed of the epidermis, dermis and subcutaneous layer. Unlike human skin, mouse skin has an additional muscle layer and is haired [46, 47]. Additional, shared components in mouse and human skin include keratinocytes, fibroblasts, ECM components - including collagen and hyaluronan, dermal blood vessels, adipocytes and immune cells [46]. Mouse skin is much thinner than human skin, with mouse skin having two to three layers of keratinocytes, while human skin has five to ten layers [46, 47]. A study identified 201 shared skin-associated genes between mouse and human skin, with GO terms associated with maintaining structure. Differences in enriched GO terms showed human skin has more genes involved in keratinocyte processes, whereas mouse skin has more genes involved
in muscle contraction [46]. While the differences should be taken into account while analyzing results and applying them to human skin, mouse skin is a valuable model due to the similarities mentioned above, the abundance of genetic variants with targeted disruptions in known oncogenic driver pathways, the critical use of UVB to accelerate tumorigenesis in mouse models and the more intense study of skin physiology in mouse compared to other mammals. Mouse skin has been important for studying human skin physiology, wound healing and potential topical treatments. In mice, distinct niches of stem cells are responsible for renewal in the epidermis. Stem cells in the stratum basale replenish the interfollicular epidermis, whereas stem cells in the hair follicle bulge are responsible for renewing the hair follicle and can also replace interfollicular epidermis when skin is wounded. Hh signalling is responsible for the activation of these bulge stem cells, and without activation, the stem cells will remain quiescent [48, 49].

3.1.6 Available mouse models of basal cell carcinoma provide the means to analyze HI in rapidly dividing somatic cells

Given the high levels of genetic variation detected in human skin, cancer and BCC specifically and the likeness in mouse and human epidermal layers and keratinocyte development, available mouse models of BCC provide a means to analyze HI in rapidly dividing somatic cells and explore the mystery of one of the highest mutation burdens within a benign human cancer. A study analyzing mutations in oncogenes found that 55% of all driver mutations were heterozygous [50] and intuitively, considering the loss of heterozygosity reported in human BCC tumors [23, 30, 31, 51–54], there will be a non-random heterozygous landscape in BCC applicable for HI testing. Additionally, CD-1 outbred mice have been crossed with \( Ptch1^{neo67/+} \) mice to create a mouse model of BCC that develops tumors when exposed to radiation [55, 56]. This breeding will lead to clusters of heterozygosity in the genome of this BCC mouse model, suitable for the investigation of HI in somatic cells. Before analyzing potential mutational mechanisms like HI, the
Chapter 3. Mutagenesis in models of BCC

genomic landscape of heterozygosity in mouse models of BCC must be characterized, and the level of genetic variation must be compared to the level reported in human BCC. A characterized model allows the study of HI through localized regions of heterozygosity and mutations like copy number variants (CNVs). Elucidation of HI in somatic cells would provide an additional mutational mechanism for cancer researchers to consider when studying tumorigenesis and mutation burdens in cancer.

Additionally, a characterized mouse model of BCC allows the study of the contributing factors and mechanisms that sustain BCC as a benign cancer despite a high mutation burden, applicable to what is seen in aged, UVB exposed eyelid and esophageal tissue. Characterized mouse models will allow researchers to study mutational mechanisms and the clonal architecture in mouse BCC tumors and detect and track the dynamic of subclonal architecture with far more detail than possible in humans. Finally, a characterized model allows the better evaluation of targeted therapies for treating BCC tumors and suppressing tumor recurrence.

3.1.7 Research goal and specific objectives

Research goal: To evaluate available mouse models of BCC as proxies for human somatic diversity and suitability for the study of HI. A systematic literature review was performed to document the reported genetic variation in human and mouse models of BCC to determine whether mouse models of BCC reproduce a similar burden and dynamic of de novo somatic mutations that characterize human BCC tumors. The systematic literature review revealed a significant gap in knowledge of the genetic variants in mouse BCC driving the next steps. Thus, the goal is to design an experiment to detect the genetic variation in mouse models of BCC and propose methods to test HI and the tumorigenic potential of mutation burdens.
The specific objectives are:

1. To perform a systematic literature review to compile a database of gene mutations reported in either or both human and mouse BCC tumors.

2. To compare the extent of the characterization of the reported mutation burden in both human and mouse BCC to evaluate mouse models of BCC as a proxy to study HI in the context of somatic cells.

3. To identify enriched terms and pathways using the list of mutated genes reported in human BCC tumors to establish pathways affected in human BCC tumors.

4. To design tools and pipelines to quantify and analyze the mutation burden in mouse BCC to fill the knowledge gap and permit the study of HI in a somatic cell model.

3.2 Materials and methods

3.2.1 Systematic literature review

To understand the mutation burden in BCC, a systematic literature review was performed to compile a database of gene mutations reported in human and mouse BCC tumors. The data were collected through a series of web searches using PubMed, Google Scholar and Web of Science to identify research published between January 1990 and April 2020 regarding genes mutated in either human or mouse BCC tumors.

Preliminary search terms were TOPIC:(basal cell carcinoma) AND TOPIC:(mutations), which gave 1,942 results. From these, the search terms were narrowed into a more specific set using (TOPIC:(basal cell carcinoma) AND TOPIC:(reported mutations) AND TOPIC:(human)) NOT TOPIC:(expression) to identify mutations in human tumors and (TOPIC:(basal cell carcinoma) AND TOPIC:(reported mutations) AND TOPIC:(mouse)) to identify mutations in mouse models of BCC. These searches returned 364 results for
human samples and 32 results for mouse samples. Primary research articles were selected, and studies were included if they reported one or more mutated gene(s) in a tumor sample from sporadic BCCs, patients with NBCCS, or mouse models of BCC. Study characteristics and results were tabulated and reported.

3.2.2 Comparison of reported mutations in human and mouse BCC tumors

To evaluate mouse models of BCC tumors as a proxy for the genetic variation seen in human BCC tumors and as a proxy to study HI in the context of somatic cells, a list of genes with reported mutations in human BCC tumors was created and compared to a list of genes with reported mutations in mouse BCC tumors. A table was created to compile mutation frequency, base or protein change, and mutation type when available.

3.2.3 Characterization of reported mutations in human and mouse BCC tumors

To identify pathways affected in human BCC tumors, the list of 110 genes with mutations reported in human BCC tumors was used to detect enrichment of pathways and GO terms. The list was input into Metascape for enrichment analysis [57]. Pathway-enrichment analyses were performed using Protein Analysis Through Evolutionary Relationships (PANTHER) [58] and the Database for Annotation, Visualization and Integrated Discovery v6.8 (DAVID), which identifies Kyoto Encyclopedia of Genes and Genomes (KEGG) [59, 60] pathway enrichment. Gene ontology (GO) Enrichment Analysis was performed using The Gene Ontology Resource [61, 62]. GO annotations of Biological Processes (GO_BP_DIRECT), Molecular Functions (GO_MF_DIRECT), and Cellular Component (GO_CC_DIRECT) were analyzed. Enriched terms were included using a P-value cut-off of less than 0.05 and were compared to identify similarities and differences between reported
mutations in human and mouse BCCs.

### 3.2.4 Experimental design and bioinformatics pipelines

To fill the knowledge gap regarding mutated genes in mouse models of BCC and permit the study of HI in a somatic cell model, three experimental designs and cost evaluations were developed and compared for mutation detection by a targeted exome panel, whole-exome sequencing (WES) and RNA-sequencing for 12 mouse samples. Experimental designs for mutation confirmation by digital droplet PCR (ddPCR) and Sanger sequencing were also designed. All methods use mm10 as the mouse reference genome, which is available for download from the Genome Reference Consortium [63]. A next-generation sequencing (NGS) targeted exome panel was custom-designed for DNA-extracted from mouse samples using Illumina DesignStudio Software [64]. This program allows the creation of a panel by selecting genomic regions of interest. All driver genes compiled in the systematic review were included. For each gene, the 5’ and 3’ untranslated regions (UTR), promoter, exons, and splice junctions were included for sequencing using the targeted exome panel. Targeted exome sequencing, WES and RNA-sequencing costs were determined and compared. Additionally, ddPCR and Sanger sequencing costs were determined and compared. The best bioinformatics tools for quality control, processing, alignment, analysis, and visualization steps were compiled into one bioinformatics pipeline for WES data and targeted exome sequencing data and another for RNA-sequencing data. ddPCR experimental protocols were designed to detect the copy number of a subgroup of genes. Sanger sequencing experimental protocols were designed to confirm mutations in a small subgroup of genes critical to BCC initiation.
Chapter 3. Mutagenesis in models of BCC

3.3 Results

3.3.1 Summary of study characteristics

Between 1990–2020, 62 studies reported one or more mutated gene(s), with 52 studies reporting mutations detected in human BCC tumors (Table 3.1) and 10 studies examining genes in mouse models of BCC (Table 3.2). 40 studies examined mutations in patients with sporadic BCC, 17 studies examined mutations in patients with NBCCS, two studies examined mutations in patients with advanced or metastatic BCC, and one study examined mutations in XP patients. Genes reported in mice are the current genes used to model BCC in mice, which target aberrant activation of the Hh signalling pathway, initiating tumorigenesis [65].
Table 3.1: **Studies reporting genic mutations detected in human BCC tumors.**

<table>
<thead>
<tr>
<th>Literature Data</th>
<th>Year</th>
<th>Samples</th>
<th>Genes</th>
<th>Occurrence</th>
<th>Key Finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>van der Schroeff et al [66]</td>
<td>1990</td>
<td>30</td>
<td>H-ras, N-ras, K-ras</td>
<td>Sporadic</td>
<td>Missense mutations in codon 12 of H-ras (G-C or G-D)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>and codon 61 K-ras (E-H) genes</td>
</tr>
<tr>
<td>Gailani et al [67]</td>
<td>1992</td>
<td>19</td>
<td>PTCH1</td>
<td>Sporadic, NBCCS</td>
<td>Allelic loss of chromosome 9q31 in 11 sporadic BCC cases and all hereditary cases</td>
</tr>
<tr>
<td>Rady et al [68]</td>
<td>1992</td>
<td>14</td>
<td>TP53</td>
<td>Sporadic</td>
<td>14 mutations, all with UV signatures</td>
</tr>
<tr>
<td>Ziegler et al [69]</td>
<td>1993</td>
<td>27</td>
<td>TP53</td>
<td>Sporadic</td>
<td>UV-induced hotspot mutations in TP53</td>
</tr>
<tr>
<td>van der Riet et al [70]</td>
<td>1994</td>
<td>36</td>
<td>TP53</td>
<td>Sporadic</td>
<td>Mutations in TP53 and inactivation of one but not both TP53 alleles</td>
</tr>
<tr>
<td>Hahn et al [71]</td>
<td>1996</td>
<td>252</td>
<td>PTCH1</td>
<td>NBCCS</td>
<td>Mutations in PTCH1</td>
</tr>
<tr>
<td>Johnson et al [72]</td>
<td>1996</td>
<td>72</td>
<td>PTCH1</td>
<td>Sporadic, NBCCS</td>
<td>Mutations in multiple exons of PTCH1</td>
</tr>
<tr>
<td>Gailani et al [73]</td>
<td>1996</td>
<td>37</td>
<td>PTCH1</td>
<td>Sporadic</td>
<td>Mutations in PTCH1</td>
</tr>
<tr>
<td>Chidambaram et al [74]</td>
<td>1996</td>
<td>59</td>
<td>PTCH1</td>
<td>NBCCS</td>
<td>Eight new PTCH1 mutations</td>
</tr>
<tr>
<td>Unden et al [75]</td>
<td>1996</td>
<td>22</td>
<td>PTCH1</td>
<td>Sporadic, NBCCS</td>
<td>Germline mutations in PTCH1 in three families and somatic mutations in sporadic BCCs</td>
</tr>
<tr>
<td>Lench et al [76]</td>
<td>1997</td>
<td>16</td>
<td>PTCH1</td>
<td>NBCCS</td>
<td>Five novel germline mutations in PTCH1</td>
</tr>
<tr>
<td>Wicking et al [77]</td>
<td>1997</td>
<td>71</td>
<td>PTCH1</td>
<td>NBCCS</td>
<td>28 mutations in PTCH1, with 86% of mutations causing protein truncation</td>
</tr>
<tr>
<td>Aszterbaum et al [78]</td>
<td>1998</td>
<td>86</td>
<td>PTCH1</td>
<td>Sporadic</td>
<td>Mutations in PTCH1 gene in sporadic BCC and patients with NBCCS</td>
</tr>
<tr>
<td>Ouhtit et al [79]</td>
<td>1998</td>
<td>128</td>
<td>TP53</td>
<td>Sporadic</td>
<td>Incidence increased with more TP53 mutations, CC&gt;TT mutation at codons 247 and 248 associated with increased risk of BCC</td>
</tr>
<tr>
<td>Xie et al [40]</td>
<td>1998</td>
<td>3</td>
<td>SMO</td>
<td>Sporadic</td>
<td>Activating missense mutations in SMO</td>
</tr>
<tr>
<td>Lam et al [80]</td>
<td>1999</td>
<td>20</td>
<td>SMO</td>
<td>Sporadic</td>
<td>1604G-T is a mutation hotspot in SMO</td>
</tr>
<tr>
<td>Shen et al [51]</td>
<td>1999</td>
<td>20</td>
<td>PTCH1</td>
<td>Sporadic</td>
<td>LOH at 9q22.3 in sporadic BCCs</td>
</tr>
</tbody>
</table>

*Continued on next page*
## Table 3.1 – continued from previous page

<table>
<thead>
<tr>
<th>Literature Data</th>
<th>Year</th>
<th>Samples</th>
<th>Genes</th>
<th>Occurrence</th>
<th>Key Finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smyth <em>et al</em> [81]</td>
<td>1999</td>
<td>85</td>
<td>PTCH2</td>
<td>NBCCS</td>
<td>Characterized the genomic structure of PTCH2 and identified a truncating mutation and a splice site mutation</td>
</tr>
<tr>
<td>Ratner <em>et al</em> [82]</td>
<td>2001</td>
<td>2</td>
<td>TP53, PTCH1</td>
<td>Sporadic</td>
<td>First report of UV-specific TP53 and PTCH1 mutations in the same BCC sample</td>
</tr>
<tr>
<td>Kim <em>et al</em> [32]</td>
<td>2002</td>
<td>15</td>
<td>TP53, PTCH1</td>
<td>Sporadic</td>
<td>UV-induced mutations and missense mutations most common in TP53 and PTCH1 in Korean patients</td>
</tr>
<tr>
<td>Lam <em>et al</em> [80]</td>
<td>2002</td>
<td>3</td>
<td>PTCH1</td>
<td>NBCCS</td>
<td>Three new mutations detected in PTCH1 in probands of three families</td>
</tr>
<tr>
<td>Couv-Privat <em>et al</em> [83]</td>
<td>2004</td>
<td>33</td>
<td>SHH</td>
<td>XP</td>
<td>SHH mutations in basal but not squamous cell carcinoma</td>
</tr>
<tr>
<td>Sironi <em>et al</em> [52]</td>
<td>2004</td>
<td>19</td>
<td>ING2, SAP30</td>
<td>Sporadic</td>
<td>LOH at 4q32-35 spanning ING2 and SAP30</td>
</tr>
<tr>
<td>Freier <em>et al</em> [84]</td>
<td>2006</td>
<td>273</td>
<td>NMYC</td>
<td>Sporadic</td>
<td>NMYC copy number gain</td>
</tr>
<tr>
<td>Reifenberger <em>et al</em> [30]</td>
<td>2006</td>
<td>42</td>
<td>Multiple</td>
<td>Sporadic</td>
<td>PTCH1, TP53, SMO and SUFU mutations</td>
</tr>
<tr>
<td>Heitzer <em>et al</em> [85]</td>
<td>2007</td>
<td>60</td>
<td>PTCH1</td>
<td>Sporadic</td>
<td>UV signature in most of the PTCH1 mutations and no difference between mutations in early-onset, regular and multiple BCCs.</td>
</tr>
<tr>
<td>Fan <em>et al</em> [86]</td>
<td>2008</td>
<td>25</td>
<td>PTCH2</td>
<td>NBCCS</td>
<td>Novel mutation in exon 15 of PTCH2 (215G-A)</td>
</tr>
<tr>
<td>Pastorino <em>et al</em> [87]</td>
<td>2009</td>
<td>5</td>
<td>SUFU</td>
<td>NBCCS</td>
<td>First report of a germline SUFU mutation</td>
</tr>
<tr>
<td>Fernandes <em>et al</em> [53]</td>
<td>2010</td>
<td>21</td>
<td>PTCH1</td>
<td>Sporadic</td>
<td>PTCH1 LOH seen more in aggressive BCC</td>
</tr>
<tr>
<td>Pan <em>et al</em> [88]</td>
<td>2010</td>
<td>44</td>
<td>PTCH1</td>
<td>Sporadic, NBCCS</td>
<td>Standard two-hit model accounts for 30% of the inactivated PTCH1, a one-hit model accounts for 32% and the remaining cases had no alteration to PTCH1</td>
</tr>
<tr>
<td>Fujii <em>et al</em> [89]</td>
<td>2013</td>
<td>1</td>
<td>PTCH2</td>
<td>NBCCS</td>
<td>Frameshift mutation in PTCH2 causing NBCCs in a Chinese family</td>
</tr>
<tr>
<td>Huang <em>et al</em> [90]</td>
<td>2013</td>
<td>31</td>
<td>PTCH1</td>
<td>Sporadic</td>
<td>17 of 31 BCCs had PTCH1 mutations, with majority insertions and deletions</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Literature Data</th>
<th>Year</th>
<th>Samples</th>
<th>Genes</th>
<th>Occurrence</th>
<th>Key Finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jayaraman et al [31]</td>
<td>2014</td>
<td>12</td>
<td>Multiple</td>
<td>Sporadic</td>
<td>Characterized the mutation landscape and detected a uniquely high number of mutations per tumor</td>
</tr>
<tr>
<td>Populo et al [91]</td>
<td>2014</td>
<td>196</td>
<td>TERT</td>
<td>Sporadic</td>
<td>Detected TERT promoter mutations in 78% of sporadic BCC tumors and in 68% of tumors from patients with NBCCS</td>
</tr>
<tr>
<td>Scott et al [92]</td>
<td>2014</td>
<td>42</td>
<td>TERT</td>
<td>Sporadic, NBCCS</td>
<td>TERT promoter mutations</td>
</tr>
<tr>
<td>Wang et al [93]</td>
<td>2014</td>
<td>30</td>
<td>TP53, PTCH1</td>
<td>Sporadic</td>
<td>Incidence of UV-specific mutations in TP53 and PTCH1 is much lower in Chinese than in white populations</td>
</tr>
<tr>
<td>Denisova et al [94]</td>
<td>2015</td>
<td>137</td>
<td>DPH3, OXNAD1</td>
<td>Sporadic</td>
<td>Frequent DPH3 and OXNAD1 mutations</td>
</tr>
<tr>
<td>Martincorena et al [5]</td>
<td>2015</td>
<td>234</td>
<td>Multiple</td>
<td>Sporadic</td>
<td>Mutation burden in BCC tumors is between 25–150 somatic mutations/Mb/cell with a high prevalence of C&gt;T mutations</td>
</tr>
<tr>
<td>Temel et al [95]</td>
<td>2015</td>
<td>75</td>
<td>ING2</td>
<td>Sporadic</td>
<td>Decreased levels of ING2 expression</td>
</tr>
<tr>
<td>Bonilla et al [23]</td>
<td>2016</td>
<td>293</td>
<td>Multiple</td>
<td>Sporadic</td>
<td>85% of BCCs had mutations in Hh pathway genes and 85% of the BCCs had mutations in additional driver genes</td>
</tr>
<tr>
<td>Temel et al [96]</td>
<td>2016</td>
<td>27</td>
<td>SIRT1-7</td>
<td>Sporadic</td>
<td>Significantly decreased expression levels of SIRT2 and SIRT3 mRNAs</td>
</tr>
<tr>
<td>Bal et al [97]</td>
<td>2017</td>
<td>48</td>
<td>ACTRT1</td>
<td>NBCCS</td>
<td>Germline mutations in ACTRT1 or sequences surrounding ACTRT1.</td>
</tr>
<tr>
<td>Kato et al [98]</td>
<td>2017</td>
<td>10</td>
<td>PTCH1</td>
<td>NBCCS</td>
<td>73 mutations in PTCH1 – 10 confirmed to be splicing mutations that affected the splice donor site</td>
</tr>
<tr>
<td>Matsudate et al [99]</td>
<td>2017</td>
<td>10</td>
<td>Multiple</td>
<td>NBCCS</td>
<td>TES-based methods are better for NBCCS diagnosis</td>
</tr>
<tr>
<td>Goodman et al [100]</td>
<td>2018</td>
<td>9</td>
<td>Multiple</td>
<td>Advanced, metastatic</td>
<td>Tumor mutation burden (TMB) and genomic landscape of nine samples from advanced or metastatic basal cell carcinoma patients</td>
</tr>
</tbody>
</table>

Continued on next page
Table 3.1 – continued from previous page

<table>
<thead>
<tr>
<th>Literature Data</th>
<th>Year</th>
<th>Samples</th>
<th>Genes</th>
<th>Occurrence</th>
<th>Key Finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laga <em>et al</em> [101]</td>
<td>2019</td>
<td>11</td>
<td><em>PTCH1</em></td>
<td>Advanced, metastatic</td>
<td>Identical <em>PTCH1</em> nonsense mutations detected in primary tumor and metastasis, plus increased UV-induced DNA damage</td>
</tr>
<tr>
<td>Maturo <em>et al</em> [102]</td>
<td>2020</td>
<td>191</td>
<td><em>PTCH1</em>, <em>TP53</em>, <em>TERT</em>, <em>DPH3</em></td>
<td>Sporadic</td>
<td>Detected and characterized <em>PTCH1</em>, <em>TP53</em>, <em>TERT</em> promoter, and <em>DPH3</em> promoter mutations, with a high proportion being C &gt;T mutations. Hotspot mutations detected in the <em>TERT</em> promoter region</td>
</tr>
</tbody>
</table>

**Abbreviations used:** Basal cell carcinoma (BCC); Nevoid basal cell carcinoma syndrome (NBCCS); Zeroderma pigmentosum (XP); Ultraviolet (UV); Tumor protein 53 (*TP53*); Patched (*PTCH*); smoothened (*SMO*); Sonic hedgehog (*SHH*); Inhibitor of Growth Family Member 2 (*ING2*); Sin3a Associated Protein 30 (*SAP30*); Loss of heterozygosity (LOH); Suppressor of Fused Homolog (*SUFU*); Telomerase Reverse Transcriptase (*TERT*); Diphthamide Biosynthesis 3 (*DPH3*); Oxidoreductase NAD Binding Domain Containing 1 (*OXNAD1*); Actin Related Protein T1 (*ACTRT1*)
Table 3.2: Studies reporting genic mutations that lead to BCC tumors in mouse models.

<table>
<thead>
<tr>
<th>Literature Data</th>
<th>Year</th>
<th>Transgenic Mice</th>
<th>Genes</th>
<th>Key Finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oro et al [39]</td>
<td>1997</td>
<td>Overexpressing Shh</td>
<td>Shh</td>
<td>Overexpression of Shh is sufficient in inducing BCC</td>
</tr>
<tr>
<td>Xie et al [40]</td>
<td>1998</td>
<td>Overexpressing Smo</td>
<td>Smo</td>
<td>Activating mutations in Smo lead to sporadic BCC</td>
</tr>
<tr>
<td>Aszterbaum et al [41]</td>
<td>1999</td>
<td>Lacking one Ptch1 allele</td>
<td>Ptch1</td>
<td>Ptch1&lt;sup&gt;+/−&lt;/sup&gt; mice develop follicular neoplasms resembling human tumors. Exposure to UV- or ionizing-radiation results in an increased number and size of tumors and a shift in their histologic features so that they more closely resemble human BCC.</td>
</tr>
<tr>
<td>Nilsson et al [42]</td>
<td>2000</td>
<td>Overexpressing Gli1</td>
<td>Gli1</td>
<td>Overexpression of Gli1 is sufficient for tumor initiation without mutations in Tp&lt;sup&gt;53&lt;/sup&gt; or ras</td>
</tr>
<tr>
<td>Grachtchouk et al [43]</td>
<td>2000</td>
<td>Conditionally expressing Gli2</td>
<td>Gli2</td>
<td>Overexpressed Gli2 can spontaneously cause BCC</td>
</tr>
<tr>
<td>Mancuso et al [55]</td>
<td>2004</td>
<td>One Ptch1 allele inactivated</td>
<td>Ptch1</td>
<td>Characterized a multistep model of BCC development with and without exposure to UV- or ionizing-radiation</td>
</tr>
<tr>
<td>Svård et al [44]</td>
<td>2006</td>
<td>Lacking one Sufu allele</td>
<td>Sufu</td>
<td>Sufu&lt;sup&gt;+/−&lt;/sup&gt; mice develop a skin phenotype similar to Gorlin syndrome</td>
</tr>
<tr>
<td>Mancuso et al [103]</td>
<td>2006</td>
<td>Lacking one Ptch1 allele</td>
<td>Ptch1</td>
<td>Tumorigenesis is strongly affected by the hair growth cycle</td>
</tr>
<tr>
<td>Yousef et al [12]</td>
<td>2010</td>
<td>Conditionally expressing Smo</td>
<td>Smo</td>
<td>BCCs arise from long-term resident progenitor cells of the interfollicular epidermis and upper infundibulum</td>
</tr>
<tr>
<td>Nitzki et al [45]</td>
<td>2012</td>
<td>Conditional Ptch1 knockout</td>
<td>Ptch1</td>
<td>Multiple different models of Ptch1 knockout mice develop tumors similar to human BCC tumors</td>
</tr>
</tbody>
</table>

**Abbreviations used:** Basal cell carcinoma (BCC); sonic hedgehog (Shh); smoothened (Smo); patched1 (Ptch1); ultraviolet (UV); GLI Family Zinc Finger 1 (Gli1); tumor protein 53 (Tp<sup>53</sup>); GLI Family Zinc Finger 2 (Gli2); Suppressor of Fused Homolog (SUFU)
3.3.2 Disparity in reported mutation burden in human tumors and mouse models

A list of 110 genes reported with mutations in human BCC tumors was compiled. Out of these genes, six were confirmed to be mutated in mouse BCC tumors (Table 3.3). These six genes are *Ptch1, Smo, Shh, Sufu, Gli1* and *Gli2*. Location of mutation and amino acid substitution were recorded (Appendix 3A). *PTCH1* was the most frequently reported gene in human and mouse BCC tumors. Inactivating mutations in *PTCH1* include missense mutations, nonsense mutations, insertions, deletions, splice site mutations, UVB signature mutations, loss of heterozygosity and promoter mutations, with the most common being nonsense and splice-site mutations. Overall there is no evidence for hotspot mutation regions in *PTCH1* (Figure 3.3). UVB specific mutations were reported in *PTCH1, TP53, SHH, SMO, CRNKL1, NEBL, STAT5B, DPH3* and *OXNAD1*, with the majority of mutations in *TP53* being C to T transitions. Both *PTCH1* and *TP53* mutations had a prevalence of C to T transitions and CC to TT tandem substitutions. Mutation hotspots in TP53 are at codons 177, 196, and 245. Promoter mutations were detected in *DPH3* [94, 102], *OXNAD1* [94], and *TERT* [91, 92, 100, 102].

Figure 3.3: Reported mutations in *PTCH1*. Exons are represented as blue bars. Each lollipop represents one point mutation. Mutation data was collected during the literature review.

Table 3.3: List of genes with reported mutations in human BCC tumors.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Nature of Mutation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ARID1A</em></td>
<td>Nonsense</td>
<td>[23, 100]</td>
</tr>
<tr>
<td><em>ACTRT1</em></td>
<td>Insertion, substitution</td>
<td>[97]</td>
</tr>
<tr>
<td><em>ASXL1</em></td>
<td>Nonsense</td>
<td>[100]</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Gene</th>
<th>Nature of Mutation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRAF</td>
<td>Point mutation</td>
<td>[5]</td>
</tr>
<tr>
<td>C7</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>CASP8</td>
<td>Nonsense</td>
<td>[23]</td>
</tr>
<tr>
<td>CCND3</td>
<td>Copy number gain</td>
<td>[23]</td>
</tr>
<tr>
<td>CD274*</td>
<td>Amplification</td>
<td>[100]</td>
</tr>
<tr>
<td>CDKN1A*</td>
<td>Missense</td>
<td>[100]</td>
</tr>
<tr>
<td>CDKN2A*</td>
<td>Missense, nonsense, point mutation, deletion</td>
<td>[23, 100]</td>
</tr>
<tr>
<td>CREBBP*</td>
<td>Frameshift</td>
<td>[100]</td>
</tr>
<tr>
<td>CRNKLI1*</td>
<td>Missense, C &gt; T Transition</td>
<td>[31]</td>
</tr>
<tr>
<td>CSMD1</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>CSMD2</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>CSMD3</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>CTNNA1*</td>
<td>Missense</td>
<td>[100]</td>
</tr>
<tr>
<td>DCC</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>DPCR1</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>DPH3</td>
<td>Substitution (Promoter), CC&gt;TT tandem substitutions</td>
<td>[94, 102]</td>
</tr>
<tr>
<td>DPP10</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>EPHA3*</td>
<td>Missense</td>
<td>[100]</td>
</tr>
<tr>
<td>ERBB2*</td>
<td>Recurrent, missense</td>
<td>[23]</td>
</tr>
<tr>
<td>ERBB4*</td>
<td>Missense</td>
<td>[100]</td>
</tr>
<tr>
<td>E2F3</td>
<td>Copy number gain</td>
<td>[23]</td>
</tr>
<tr>
<td>FAT1</td>
<td>Point mutation</td>
<td>[5]</td>
</tr>
<tr>
<td>FBXW7*</td>
<td>Deleterious, missense, nonsense</td>
<td>[23]</td>
</tr>
<tr>
<td>FGFR3</td>
<td>Missense</td>
<td>[23]</td>
</tr>
<tr>
<td>FLT1*</td>
<td>Missense</td>
<td>[100]</td>
</tr>
<tr>
<td>GLI1*†</td>
<td>Missense</td>
<td>[12, 26, 30, 99, 100]</td>
</tr>
<tr>
<td>GLI2†</td>
<td>Missense</td>
<td>[12, 26, 99]</td>
</tr>
<tr>
<td>GLI3</td>
<td>Missense</td>
<td>[99]</td>
</tr>
<tr>
<td>GPR139</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>GRIN2A*</td>
<td>Nonsilent</td>
<td>[31, 100]</td>
</tr>
<tr>
<td>GRM3*</td>
<td>Missense</td>
<td>[100]</td>
</tr>
<tr>
<td>HEATR7B2</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>HHIP</td>
<td>Point mutation</td>
<td>[26, 99]</td>
</tr>
<tr>
<td>HRAS*</td>
<td>Recurrent, substitution</td>
<td>[5, 23]</td>
</tr>
<tr>
<td>ING2</td>
<td>Deletions, LOH</td>
<td>[52, 95]</td>
</tr>
<tr>
<td>INPP4B*</td>
<td>Nonsense</td>
<td>[100]</td>
</tr>
<tr>
<td>ITIH2</td>
<td>Nonsilent</td>
<td>[31, 104]</td>
</tr>
<tr>
<td>JAK2*</td>
<td>Amplification</td>
<td>[100]</td>
</tr>
<tr>
<td>KCNT2</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>KDM5A*</td>
<td>Missense</td>
<td>[100]</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Gene</th>
<th>Nature of Mutation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>KDR*</td>
<td>Missense</td>
<td>[100]</td>
</tr>
<tr>
<td>KEAP1*</td>
<td>Missense</td>
<td>[100]</td>
</tr>
<tr>
<td>KEL 1*</td>
<td>Missense</td>
<td>[100]</td>
</tr>
<tr>
<td>KIF7</td>
<td>Missense</td>
<td>[99]</td>
</tr>
<tr>
<td>KNSTRN*</td>
<td>Missense</td>
<td>[23]</td>
</tr>
<tr>
<td>KRAS*</td>
<td>Recurrent, substitution</td>
<td>[5, 23]</td>
</tr>
<tr>
<td>LATS1*</td>
<td>Recurrent, missense, nonsense</td>
<td>[23]</td>
</tr>
<tr>
<td>LATS2*</td>
<td>Missense</td>
<td>[23]</td>
</tr>
<tr>
<td>LRP1B*</td>
<td>Nonsense, splice site</td>
<td>[100]</td>
</tr>
<tr>
<td>MAGI2*</td>
<td>Nonsense</td>
<td>[100]</td>
</tr>
<tr>
<td>MLL2*</td>
<td>Frameshift</td>
<td>[100]</td>
</tr>
<tr>
<td>NEBL*</td>
<td>Missense, C &gt;T Transition</td>
<td>[31]</td>
</tr>
<tr>
<td>NF1</td>
<td>Point mutation</td>
<td>[5]</td>
</tr>
<tr>
<td>NMYC*</td>
<td>Recurrent, CNV (Gain), missense</td>
<td>[23, 84, 105]</td>
</tr>
<tr>
<td>NOTCH1*</td>
<td>Missense, nonsense</td>
<td>[5, 23, 31, 100]</td>
</tr>
<tr>
<td>NOTCH2*</td>
<td>Nonsense</td>
<td>[5, 23, 31, 100]</td>
</tr>
<tr>
<td>NOTCH3</td>
<td>Point mutation</td>
<td>[5]</td>
</tr>
<tr>
<td>NPAP1</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>NRAS*</td>
<td>Recurrent, substitution</td>
<td>[5, 23]</td>
</tr>
<tr>
<td>OR5M3</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>OXNAD1</td>
<td>Substitution (Promoter), CC&gt;TT tandem substitutions</td>
<td>[94]</td>
</tr>
<tr>
<td>PCDHB2</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>PCDHB3</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>PCDHGB3</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>PDCD1LG2*</td>
<td>Amplification</td>
<td>[100]</td>
</tr>
<tr>
<td>PDGFRA*</td>
<td>Missense</td>
<td>[100]</td>
</tr>
<tr>
<td>PIK3CA*</td>
<td>Recurrent</td>
<td>[23]</td>
</tr>
<tr>
<td>PIK3R1*</td>
<td>Nonsense</td>
<td>[100]</td>
</tr>
<tr>
<td>PIK3R2*</td>
<td>Nonsense</td>
<td>[100]</td>
</tr>
<tr>
<td>POM121L12</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>PPIAL4G</td>
<td>Missense</td>
<td>[23]</td>
</tr>
<tr>
<td>PPP1R3A</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>PPP6C*</td>
<td>Recurrent, missense</td>
<td>[23]</td>
</tr>
<tr>
<td>PREX2</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>PTCH1*†</td>
<td>Missense, truncating, nonsense, InDels, splice site, C&gt;T transitions, CC&gt;TT transitions, LOH</td>
<td>[23, 26, 30–32, 51, 53, 67, 71–78, 80, 82, 88, 90, 98–102, 105]</td>
</tr>
<tr>
<td>PTCH2</td>
<td>Truncating, Frameshift, Missense</td>
<td>[81, 86, 89, 99]</td>
</tr>
<tr>
<td>PTEN*</td>
<td>Splice site</td>
<td>[100]</td>
</tr>
<tr>
<td>PTPN14*</td>
<td>Missense, Truncating</td>
<td>[23]</td>
</tr>
</tbody>
</table>
Table 3.3 – continued from previous page

<table>
<thead>
<tr>
<th>Gene</th>
<th>Nature of Mutation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAC1*</td>
<td>Missense</td>
<td>[23, 100]</td>
</tr>
<tr>
<td>RB1*</td>
<td>Deleterious, truncating, missense, LOH</td>
<td>[23]</td>
</tr>
<tr>
<td>RBM10*</td>
<td>Splice site</td>
<td>[5, 100]</td>
</tr>
<tr>
<td>RPL22</td>
<td>Missense</td>
<td>[23]</td>
</tr>
<tr>
<td>RUNX1*</td>
<td>Missense</td>
<td>[100]</td>
</tr>
<tr>
<td>SAP30</td>
<td>LOH</td>
<td>[52]</td>
</tr>
<tr>
<td>SHH†</td>
<td>Silent, missense, C&gt;T transitions, CC&gt;T tandem</td>
<td>[83, 99]</td>
</tr>
<tr>
<td>SLIT2*</td>
<td>Nonsense</td>
<td>[100]</td>
</tr>
<tr>
<td>SMARCA4*</td>
<td>Nonsense</td>
<td>[100]</td>
</tr>
<tr>
<td>SMO*†</td>
<td>Missense, silent, C&gt;T transition, CC&gt;T, G&gt;T transversion</td>
<td>[23, 30, 40, 80, 99, 100]</td>
</tr>
<tr>
<td>STAG2*</td>
<td>Nonsense</td>
<td>[100]</td>
</tr>
<tr>
<td>STAT5B*</td>
<td>Missense, C&gt;T Transition</td>
<td>[31]</td>
</tr>
<tr>
<td>STEAP4</td>
<td>Nonsilent</td>
<td>[31, 106]</td>
</tr>
<tr>
<td>STK19*</td>
<td>Recurrent, Missense</td>
<td>[23]</td>
</tr>
<tr>
<td>SPTA*</td>
<td>Missense</td>
<td>[100]</td>
</tr>
<tr>
<td>STK36</td>
<td>Deletion</td>
<td>[99]</td>
</tr>
<tr>
<td>SUFU*†</td>
<td>Substitution</td>
<td>[23, 99]</td>
</tr>
<tr>
<td>SULT1C3</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>TAF1*</td>
<td>Splice site</td>
<td>[100]</td>
</tr>
<tr>
<td>TET2*</td>
<td>Frameshift</td>
<td>[100]</td>
</tr>
<tr>
<td>TSC1*</td>
<td>Loss</td>
<td>[100]</td>
</tr>
<tr>
<td>TERT*</td>
<td>Substitution, (Promoter)</td>
<td>[91, 92, 100, 102]</td>
</tr>
<tr>
<td>TGFBR1</td>
<td>Deletion</td>
<td>[99]</td>
</tr>
<tr>
<td>TMEM132D</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>TMEM217</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>TP53*</td>
<td>Missense, truncating, nonsense, splice site, frameshift, C&gt;T transitions, CC&gt;T tandem substitutions</td>
<td>[5, 23, 31, 32, 68–102, 70, 79, 82, 100, 102, 105]</td>
</tr>
<tr>
<td>UGT2B10</td>
<td>Nonsilent</td>
<td>[31]</td>
</tr>
<tr>
<td>WT1*</td>
<td>Missense</td>
<td>[100]</td>
</tr>
<tr>
<td>YAP1</td>
<td>Point mutation</td>
<td>[23]</td>
</tr>
</tbody>
</table>

* Mutation location and amino acid substitution recorded in Appendix 3A
† Gene has been studied in mouse models of BCC
3.3.3 Analysis of pathway and gene ontology enrichment in reported genes with mutations in BCC tumors

All 110 genes identified in the systematic literature review were input into Metascape, PANTHER, DAVID and GO Enrichment Analysis. Metascape detected 77 enriched terms, with the top three enriched terms being ‘Pathways in Cancer’ (1.2x10^{-35}), ‘Melanoma’ (4.1x10^{-29}) and ‘Regulation of smoothened signalling pathway’ (3.4x10^{-18}) (Figure 3.4). PANTHER analyses reported 33 enriched pathways (Appendix 3B), with the most genes involved in the ‘Hedgehog signalling pathway’ (1.2x10^{-12}), followed by the ‘p53 Pathway feedback loops2’ (3.0x10^{-11}) and ‘EGF receptor signalling pathway’ (7.8x10^{-10}). All genes reported in mouse tumors were in the Hh pathway. KEGG analyses reported 62 enriched pathways (Appendix 3C), with ‘Pathways in cancer’ (2.8x10^{-20}) being the most enriched term followed by other cancers. ‘Hedgehog signalling pathway’ (5.7x10^{-10}) is the first non-cancer enriched term. There was significant enrichment for 1029 Biological Processes, with the top three enriched terms being ‘regulation of developmental process’ (3.4x10^{-15}), ‘regulation of cell population proliferation’ (2.5x10^{-15}), and ‘regulation of smoothened signalling pathway’ (6.6x10^{-15}). There was significant enrichment for 44 Molecular Functions, with the top three enriched terms being ‘transcription factor binding’ (5.3x10^{-5}), ‘protein kinase activity’ (7.7x10^{-4}), and ‘ion binding’ (2.3x10^{-3}). There was significant enrichment for 29 Cellular Component terms, with the top three enriched terms being ‘nucleoplasm’ (1.9x10^{-4}), ‘nucleus’ (2.4x10^{-4}), and ‘ciliary tip’ (1.9x10^{-4}). A complete list of enriched terms can be found in Appendix 3D.

3.3.4 Multiple experimental designs provided for characterization of somatic variation in BCC

It is essential to characterize the mutation burden in available mouse models of BCC to confirm they are appropriate proxies for the mutation burden and dynamic seen in human
BCC. Methods including WES, TES, RNA-sequencing can be used to detect the landscape of heterozygosity and mutations in the genomes of mouse models of BCC to evaluate their applicability to studying HI. For all experimental designs, there would be at least two groups of mice — \( P\text{tch1}^{+/−} \) mice either exposed to UVB or not. Other modifying factors or mutagens could be added depending on the research question to study the tumorigenic potential of the mutation burden. The reported mutation burden and dynamic can be tracked with UVB exposure and other treatments or preventative measures. Table 3.4 provides a summary of the mutation detection and confirmation methods analyzed to characterize mouse models of BCC, allowing elucidation of the research questions. All costs were calculated for 12 samples, two non-UVB exposed mice and four UVB exposed mice with two technical replicates per sample. This number of samples provides multiple samples per group to survey the mutation burden in tumors with statistical analysis.

Epidermis samples would be harvested, DNA would be extracted and prepared for WES. Mouse exome library prep is $337.50 per library prep using Agilent SureSelect mouse kit plus. To achieve coverage of 150–200X, four exomes can be loaded per sequencing
Table 3.4: **Summary of potential methods for mutation discovery and confirmation in mouse models of BCC.**

<table>
<thead>
<tr>
<th>Method</th>
<th>Application</th>
<th>Cost ($)</th>
<th>Scope</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Targeted exome</td>
<td>Detection</td>
<td>63,620.37</td>
<td>100 genes</td>
<td>Directly confirm reported mutation in human BCC tumors</td>
</tr>
<tr>
<td>WES</td>
<td>Detection</td>
<td>12,990.00</td>
<td>All coding genes</td>
<td>Directly confirm the reported mutation in human BCC tumors and detect novel variants</td>
</tr>
<tr>
<td>RNA-Sequencing</td>
<td>Detection</td>
<td>7,470.00</td>
<td>Transcriptome</td>
<td>Determine differentially expressed genes in mouse models</td>
</tr>
<tr>
<td>ddPCR</td>
<td>Validation</td>
<td>3,456.00</td>
<td>4 genes</td>
<td>Validate mutations detected in mouse models</td>
</tr>
<tr>
<td>Sanger</td>
<td>Validation</td>
<td>9,483.48</td>
<td>4 genes</td>
<td>Validate mutations detected in mouse models</td>
</tr>
</tbody>
</table>

lane on a high throughput flow cell. One lane costs $2,980.00; thus, lane costs for 12 samples would be $8,940.00. The total cost for WES of 12 samples is $12,990.00. Data received would be analyzed using the designed bioinformatic pipeline (Figure 3.5A). First, data would be processed, including quality control, and aligned to the mouse reference genome (mm10). Then, the data would be processed and prepared to detect variants and mutations. Copy number variants and single nucleotide variants would be detected. Data would be visualized, including a CNV-LP to display the mutation landscape. Mutation burdens would be compared between groups of mice. Rainfall plots and CNV-LPs would visualize the single nucleotide variant and copy number variant landscapes in these mice. Genes with confirmed mutations would be compared to the reported burden in humans and analyzed using Metascape, PANTHER and KEGG pathway analysis, along with Gene Ontology enrichment analysis.

Epidermis samples would be harvested, RNA would be extracted and prepared for RNA-sequencing. Libraries would be prepared using NEBNext Ultra II Directional RNA library prep kit, costing $250.00 per sample. Paired sequencing would be done on a
HiSeq 2500 instrument. To achieve 360 million paired-end reads, 12 samples can be run in one and a half sequencing lanes for a total cost of $4,470.00. The total cost for RNA-sequencing of 12 samples is $7,470.00. Data received would be analyzed using the designed bioinformatic pipelines (Figure 3.5B). First, data would be assessed for quality and then aligned to the mouse reference transcriptome. The aligned reads would be quantified, and then differential expression analysis would be performed. Quality would be assessed again, and expression data would be visualized using heatmaps and Volcano plots. Genes that are significantly differentially expressed would be analyzed using Metascape, PANTHER and KEGG pathway analysis, along with Gene Ontology enrichment analysis. In addition, genetic variants can be called from RNA-sequencing data using a variant analysis pipeline [107] or Strelka2 [108]. Data would be visualized, and transcriptomes would be compared between groups of mice. Both WES and RNA-sequencing pipelines can be run using Galaxy, R Studio and Python. Different approaches were compared at each step, and the method that suited this project best was selected [109–113].

Epidermis samples would be harvested, and DNA would be extracted and prepared for next-generation sequencing (NGS) using a custom panel. An AmpliSeq™ Custom DNA Panel for Illumina was designed containing 100 genes identified as mutated in human BCC tumors, either with a strong growth advantage or frequently mutated in human and mouse BCC (BCC110 panel). \textit{Ptch1}, \textit{Shh}, \textit{Smo}, \textit{Tp53}, \textit{Notch1}, \textit{Notch2}, \textit{Gli1}, and \textit{Gli2} were among the genes included. Data received would be analyzed using the bioinformatics pipeline applied to WES data. Data would be visualized and compared between groups of mice. CNV-LP would be used to visualize the mutation burden across the selected genes. The cost of the Custom DNA panel is $42,186.37, and with consumables and processing fees, the total comes to $63,620.37.

dPCR would be performed to confirm the copy number of important genes in BCC initiation. Epidermis samples would be harvested, DNA would be extracted and prepared for ddPCR. ddPCR would be performed using six TaqMan Copy Number Assays, selected
Chapter 3. Mutagenesis in models of BCC

Based on the compiled gene database and mutation information: *Ptch1* (Mm00400946.cn, Mm00401581.cn, Mm00400966.cn), *Tp53* (Mm00370089.cn), *Smo* (Mm00735101.cn), and *Rbm10*, which acts as a test gene and an internal control on the X chromosome to detect CNs of 1 (Mm00523229.cn). Considering *Ptch1* is the most commonly mutated gene in BCC, three assays were chosen throughout the gene: one assay overlaps Intron 1 — Exon 1, which must be present and thus is a good candidate to determine if the remaining *Ptch1* allele is present (Mm00400946.cn); one assay within exon 9 and the Sterol-sensing domain, which is reported to be responsible for the inhibition of *Smo* (Mm00401577.cn); one assay within exon 16 and the fourth extracellular loop, which is hypothesized to contain the *Shh* ligand-binding motif (Mm00400966.cn). One assay was chosen to confirm the copy number of *Smo*, another Hh pathway member (Mm00735101.cn). This assay has been pre-tested and would allow the determination of the CN. One assay was chosen to confirm the copy number of *Tp53*, which has been implicated in BCC tumorigenesis [23, 31]. The assay overlaps Exon 1 — Intron 1 and has been previously cited in the literature (Mm00370089.cn). Finally, one assay was chosen for RNA Binding Motif Protein 10 (*Rbm10*), which is reported to have mutations and can be used as an internal control (Mm00523229.cn). The total cost of ddPCR would be $3,456.00, including six assays and consumables.

Sanger sequencing would be performed to confirm mutations in specific genes crucial to BCC tumorigenesis. Sanger sequencing would be performed to detect mutations in *Ptch1*, *Tp53*, *Smo* and *Rbm10* in 12 mice. The cost per sample is $790.29. The total cost for 12 samples is $9,483.48. Results would be compared to reference sequences to identify any variants.
### 3.4 Discussion

#### 3.4.1 Reported mutations in BCC and a disparity between human and mouse models

This systematic literature review confirmed the diversity of genes with acquired mutations, the high mutation burden, and the prevalence of UVB-specific mutations in human BCC tumors, as previously reported [5, 17, 31]. This chapter presents the first compilation of genes with mutations in human BCC tumors and the associated research papers. Additionally, this chapter is the first study to identify and report the striking knowledge gap between the reported mutations in BCC tumors in humans and mouse BCC models.

It is clear that mouse models of BCC mirror human BCC tumorigenesis and well represent human BCC tumors in terms of morphology, histology and pathogenesis (Table 3.2) [11, 12, 41, 103]. However, throughout the review, it became apparent that 94% of mutations have not yet been assayed in mouse models of BCC, with 104 mutations...
Chapter 3. Mutagenesis in models of BCC

reported in human BCC tumors but not mouse tumors. Further, there has been neither germline nor somatic mutation landscape assessment in mouse BCC tumors. While there is no argument that the characterization of human BCC tumors is crucial to understanding BCC, mouse models provide the opportunity to explore mechanisms of mutagenesis, like HI, carcinogenesis and the tumorigenic and malignant potential of mutation burdens. Once these models are characterized, researchers can look for evidence of HI in somatic cells, followed by an exploration of possible HI-associated mutational mechanisms. Secondary to mechanistic explorations, characterized mouse models of BCC would be ideal for determining how to modify tumorigenic potential and understand malignancy to explore interventions toward controlling cancer initiation. Studies using mouse BCC can potentially provide insights into the complex phenotype seen in the human eyelid and allow researchers to study this phenotype with varying and controlled environmental and microenvironmental factors. Confirmation of mutation burden in mice with and without UVB exposure will allow identification of hotspots and characterization of the dynamic mutation burden when exposed to UVB.

In addition to the genes known to be important in BCC tumorigenesis being very well studied [23, 26, 28, 30, 31], researchers have performed whole-exome sequencing on large groups of human BCC tumor samples to define the mutation landscape [31]. Additional researchers have performed differential gene expression analysis to create expression profiles in human BCC tumors [26, 105, 114–117]. A well-characterized mutation landscape and gene expression profile are crucial for a baseline mutation profile in mouse models of BCC for any studies going forward, yet no similar research has been carried out in mouse models. The baseline characterization will allow direct investigation into the phenomenon of HI in rapidly dividing cells, comparison and further definition of how UVB affects the mutation burden and gene expression in mice, and provide a model to analyze how other carcinogens affect the mutation burden in this model. It will also allow studies to determine the tumorigenic or metastatic potential of a mutation burden and if preventative methods or
treatments affect either the BCC mutation burden or expression profile.

Mouse models used as proxies for other cancer types have been characterized, and the benefits emphasize the importance of defining mouse models for more informed research going forward and highlights the need for similar characterization in mouse models of BCC. Whole exome sequencing has been used to characterize mouse models of non-small-cell lung cancer, liver cancer, and breast cancer. Non-small-cell lung cancer was induced by activation of Kras or carcinogenic exposure, and mutational landscapes were compared, identifying differences in mutation burdens, single nucleotide variants, and copy number profiles with or without carcinogenic exposure [118]. A recent study characterized the mutation landscape of carcinogen-induced mouse models of liver cancer, comparing the mutation burden in carcinogen-exposed and unexposed mouse liver tumors to each other and the mutation burden in human hepatocellular carcinomas to identify similarities and differences with and without exposure and between mouse models and human cancer [119]. Another study compared two mouse models of breast cancer, MMTV-Neu and MMTV-PyMT, to determine similarities and differences between the mutation landscape in both. This characterization and comparison allow researchers to determine which model will better represent the sub-type of breast cancer they are studying [120].

Overall, the systematic review has shown the knowledge gap that must be filled in mouse models of BCC. It highlights the need to characterize the landscape of genetic variation in available mouse models of BCC, specifically Ptch+/– mice, in greater depth, which will allow more informed research to study HI and possible mechanisms and elucidate the phenotypes and mechanisms behind the highly mutated yet benign nature of BCC.
3.4.2 Enriched pathways and terms reflect known pathways involved in human BCC tumorigenesis and should be confirmed in mouse models of BCC

Analysis of the gene list showed enrichment in genes involved in ‘Pathways in Cancer,’ ‘Hedgehog signalling pathway,’ and ‘Regulation of smoothened pathway.’ Metascape and the Gene Ontology Resource identified enrichment in genes involved in ‘Negative regulation of cell proliferation’ and ‘Negative regulation of growth,’ which could play a role in most BCC tumors’ benign nature. This enrichment seen in human BCC must be investigated and confirmed in mouse models of BCC to use them as a valid proxy to study HI and the complex phenotype of mutation burdens and their tumorigenic, malignant or metastatic potential.

Considering the large number and variety of pathways and GO terms that show enrichment from the list of genes mutated in human BCC tumors, it is important to confirm the mutation burden in available mouse models of BCC and determine whether or not there are a similar dynamic and number of enriched terms in mouse models. Not only would these results be important to understand complex phenotypes seen in these mouse models while using them as proxies for human BCC, but these results could be used to study and better understand why clones with such a high mutation burden and mutations in known cancer driver genes are benign or undergo cancer initiation but rarely metastasize. Previous studies have confirmed that mouse models of BCC share similar morphologies with human BCC tumors [41, 44, 45, 55]. However, it is essential to confirm in vivo that the mutation burden and dynamics in available mouse models of BCC represent human BCC tumors.
3.4.3 Multiple experimental designs provide options to detect genetic variation in mouse models of BCC depending on the scope of the research question, budget and samples available

The ideal method to interrogate the mutation burden in mouse models of BCC depends on the research question, number of samples available for analysis and budget. Given a smaller number of samples or a lower budget, WES or RNA-sequencing would be the ideal method to detect the mutation burden and further characterize mouse models of BCC. These methods would allow direct comparison of genes mutated in mouse models to human tumor samples. In addition, the discovery of mutations in genes not reported in human tumors is possible, along with differentially expressed genes, which are both essential factors when fully characterizing phenotypes seen in a model. Comparatively, given a greater number of samples or a higher budget, targeted-exome sequencing provides a highly sensitive approach for mutation detection to survey driver mutations and characterize the dynamic subclonal diversity in the epidermal cells. Detecting mutations in different clones in the epidermal cells will help elucidate how the mutation burden accumulates and clonal dynamics. This information will provide an extensively characterized mutation landscape, which is ideal for studying HI when identifying the landscape of heterozygosity and potentially associated mutations. Knowing the baseline mutation burden with UVB exposure will allow researchers to account for UVB mutagenesis when assaying the genome for evidence of HI and HI mechanisms.

Targeted-exome sequencing, WES, and RNA-sequencing each have their advantages and disadvantages. Targeted-exome sequencing using the BCC110 targeted gene panel and WES would allow for direct confirmation of genes that had reported mutations in human BCC tumors in available mouse models of BCC. This confirmation of mutated genes would fill the missing mutation gap between human and mouse models of BCC.
Using a targeted gene panel would not allow for discovering new genes that may be mutated in mouse models of BCC but would be an optimal method to assess a large number of samples. In contrast, WES allows for the discovery of genes with previously unreported mutations in BCC tumors.

Currently, no panels are available to detect the somatic mutation burden in human or mouse BCC tumors, which means this would be the first targeted NGS panel to test the mutation burden in skin cancers, particularly BCC. Illumina provides eight targeted NGS panels focusing on genes with known cancer associations (Appendix 3E). These panels screen between 6 to 500 known driver genes of various cancers or known modifiers of tumor invasion or aggression. These panels are very effective and helpful in detecting mutations in a sample, which can better predict the need for preventative measures or targeted treatments. One of the eight targeted NGS panels is specifically targeted to detect mutations in six genes known to increase an individual’s risk for hereditary skin diseases, including *PTCH1* and *SUFU*. The BCC 110 panel would give researchers a tool to study the mutation burden in BCC tumors to understand factors that modify tumorigenesis and the aggressive potential of tumors, and give doctors a panel to assess BCC patients’ mutation burden.

Both targeted-exome sequencing and WES would allow for comparison and characterization of UVB-specific mutations in mouse models. RNA-sequencing would determine what is differentially expressed with and without UVB exposure or an environmental modifier, treatment or preventative method. Somatic mutations can be detected from RNA-sequencing data using a previously reported pipeline [109], which would allow for another mutation detection method. Targeted-exome sequencing allows for confirmation of what is reported in humans. It is the most expensive method, whereas WES is less expensive and allows for confirmation and discovery, past the mutations that have been reported in human BCC tumors. RNA-sequencing provides a good alternative for detecting gene expression differences and somatic mutations once the known mutation burden
in mouse models is established. RNA-sequencing is the least expensive method proposed.

3.4.4 With characterization, mouse models of BCC are promising proxies of high dynamic genetic variation to study HI in somatic cells

Currently, available mouse models of BCC are not yet a valid proxy to study HI in somatic cells as levels of genomic heterozygosity and mutations have not been detected. After the genetic variation and \textit{de novo}, somatic mutation burden are detected and characterized in mouse models of BCC, mouse models can be used as proxies to study multiple complex phenotypes seen in humans. Combined with array comparative genome hybridization to detect CNVs across the genome of mouse models of BCC, the HI hypothesis could be assessed within these characterized models. These experiments would indicate whether HI occurs within the highly variable skin tissue and, more specifically, within somatic cells or if HI occurs in germ cells and is meiosis related. The results provided by the methods and pipelines proposed for genetic variation detection can be used to analyze HI.

Not only could mouse models of BCC with characterized landscapes of genetic variation be used to study HI, but cancer researchers can utilize these models to study other complex phenotypes. Another complex phenotype is the tumorigenic potential of mutation burdens in different environments, as seen in UVB-exposed epidermal tissue and esophageal epidermal tissue, both containing high levels of mutations and mutations in cancer-associated genes yet no transformation [5, 6]. Similarly, characterized mouse models of BCC can be used to study the perplexing phenomenon seen in human BCC. A well-characterized model will facilitate the better evaluation of preventative measures and targeted therapies for BCC and suppressing tumor recurrence. Finally, elucidating whether or not HI occurs in rapidly dividing cancer cells will be invaluable to cancer researchers to understand HI and if heterozygosity breeds more variation.
3.4.5 Study limitations

The greatest limitation of this study is that there is no in vivo work performed to confirm or explore the results presented from the systematic literature review. Additionally, there is currently no direct evidence of a discontinuous landscape of heterozygosity in mouse models of BCC. Indirect evidence is implied through loss of heterozygosity in many genes in human BCC tumors [23, 30, 31, 51–54], oncogenes with 55% heterozygous driver mutations [50], and the fact that specific Ptch1+/− mouse models were crossed with CD-1 outbred mice when created [55, 56]. These limitations can quickly be assessed and ameliorated in future research using the methods provided in this chapter for mutation detection and confirmation.

3.5 Conclusion

This chapter presents the first systematic review to compile BCC tumors’ mutation burden in humans and mouse models, identifying a gap in knowledge between reported mutations in humans and mouse models. This chapter highlights the importance of characterizing these mouse models to begin to fill this knowledge gap and provide a landscape optimal for studying HI and other complex phenotypes. This chapter contributes two novel targeted panels to detect the somatic mutation burden in human BCC tumors and detect common human BCC mutations in BCC mouse models. Additionally, this chapter provides pipelines and tools for detection, analysis, and visualization of the mutation burden and transcriptome in BCC mouse models. Multiple methods and pipelines are proposed for future researchers to confirm the somatic mutation burden in available mouse models of BCC, ideally in Ptch1+/− mice. The targeted panel provides a sensitive approach for mutation detection to survey driver mutations and characterize the epidermal cells’ dynamic subclonal diversity.
Mouse models of BCC with a characterized landscape of genetic variation will be optimal to study HI in rapidly dividing cells. The results from *Arabidopsis* and peach plants indicate that HI may be active throughout meiosis [1, 2], but whether HI occurs during recombination or replication is unknown. Thus, HI may be occurring during replication in mitosis as well, and it is imperative to study this in a model of high somatic variation. Once the landscape of heterozygosity is detected in mouse models of BCC and genetic variants are characterized, researchers can analyze for evidence of co-localizing heterozygosity and mutations. If co-localization is detected, possible mechanisms can be studied. Elucidation of mutation mechanisms may also provide insight into any contribution HI has to genetic heterogeneity and mutation burden in BCC tumors. Secondarily, available mouse models of BCC, particularly *Ptch1*+/− mice, provide a means to analyze and explore the mystery of one of the highest mutation burdens within a benign human cancer. Future work must characterize the somatic mutation burden in these mouse models before analysis of these two phenomena is possible.

Future research can utilize the provided experimental workflow and analysis tool to fully characterize the mutation burden in BCC mouse models with and without UVB exposure, filling a critical knowledge gap. These pipelines can additionally be extended to other mouse models and biological questions.
References


REFERENCES


Mutations of the human homolog of drosophila *patched* in the nevoid basal cell carcinoma syndrome, *Cell* 85, 841–851.


REFERENCES


[118] Westcott, P., Halliwill, K., To, M., Rashid, M., Rust, A. et al. [2015], The mutational


Chapter 4.

Conclusions and future directions

This thesis presents preliminary indirect evidence consistent with predictions based on a phenomenon of Heterozygote Instability (HI) in the analysis of two stocks of outbred mice and identifies a mouse cancer model suited to the analysis of HI in somatic cells. Both models, individually and in combination, are powerful tools to test the HI hypothesis of de novo mutagenesis directly and elucidate the associated mutational mechanisms. Complete characterization of genetic variation in outbred mice is also valuable to discover genetic contributions and the missing heritability contributing to complex disease. Complete characterization of genetic variation in mouse models of BCC is also valuable to elucidate the nature and burden of somatic mutations and factors other than the genetic variation that contribute to carcinogenesis, malignancy and metastasis. This thesis highlights the power of the genetic model and the value in characterizing models fully.

This thesis confirms that outbred mice are better representatives of genetic variation in the human genome, as previously known, and should be utilized more frequently when using mouse models as proxies for human genomic diversity and study of contributors to complex phenotypes. The CNV landscape was detected and visualized in CD-1 and NMRI mice. In 99 CD-1 and 279 NRMI mice, 4313 and 2629 CNVs were characterized across the genome, respectively. Size, location, genes affected by CNVs, and copy number variable regions were catalogued and compared within and between CD-1 and NMRI mice. Similarities and differences were identified between CNVs reported in inbred mice, wild-caught mice, wild-derived mice, and humans. Clusters and deserts of heterozygous SNP loci were identified across the genomes of both outbred mouse stocks. Characterizing the heterozygous SNP and CNV landscapes in CD-1 and NMRI mice allowed analysis of
spatial associations possibly implicating associated mutation mechanisms. A potential association between localized regions of heterozygosity and CNV occurrence was detected in CD-1 and NMRI mice. Statistical analysis identified a positive correlation between increasing levels of heterozygosity and proximally associated CNVs. These findings are consistent with the HI previously shown in Arabidopsis [1] and peach plants [2]. This thesis confirms the discontinuous heterozygous landscape at SNP loci in CD-1 and NMRI mice, highlighting these mice as ideal models to study Heterozygote Instability further by direct testing using a breeding experiment and assay of heterozygosity and CNVs.

Additionally, towards understanding HI and its possible function in somatic mutagenesis, this thesis presented the first systematic literature review to compile all reported mutations in human BCC tumors and mouse BCC tumors. The disparity between mutation landscapes documented for human BCC tumors compared to mouse models of BCC was identified. Characterizing available mouse models of BCC will provide ideal models to study Heterozygote Instability in somatic cells given high genetic diversity, high cell replication and high de novo mutation burden. Mouse models of BCC allow for elucidation of the factors and mechanisms behind one of the highest mutation burdens within one of the most benign human cancers and provide a model of study of mutagenesis in the context of very high genetic variation. Furthermore, characterization of the mutation landscape in mouse models of BCC will provide a baseline of mutations in these mice without any modifiers or mutagens. This baseline will provide a tool for researchers while interpreting results obtained using these mouse models in terms of potential genetic variants that may have contributed to phenotypic variation or comparison if testing carcinogens, preventative measures or possible treatments. Different methods for mutation detection and subsequent mutation confirmation were compared, and three experimental workflows were provided to characterize the mutation burden in BCC mouse models. Two methods were proposed for the confirmation of select mutations. In BCC models, the mutation burden should be characterized with and without UVB exposure to fully
explore the dynamics of the mutation burden with UVB exposure, comparable to defining the mutation burden in non-small-cell lung cancer induced by activating mutations or carcinogen exposure [3]. Overall, this thesis highlights a need for characterization of mouse models of BCC to optimize them as models of high dynamic genetic variation to study HI in somatic cells, along with other complex phenotypes, like the phenomenon seen in BCC tumors, in UVB-exposed epidermal tissue [4] and esophageal epidermal tissue [5]. Studying HI in mouse models of BCC will allow elucidation of whether HI occurs in mitosis in somatic cells or meiosis in germ cells, or both.

The ideal model to test HI in germ and somatic cells would be an outbred mouse model of BCC, specifically a \(Ptch1^{+/−}\) CD-1 outbred mouse model. This level of genomic variation would better represent the human genome than current inbred mouse models of BCC. Additionally, outbred BCC mouse models would have discontinuous landscapes of heterozygosity throughout their genome, optimizing their genome to study HI. While this may seem counterintuitive considering heterosis and hybrid vigor, which is the increased fitness of the progeny from two diverse parents [6], CD-1 mice specifically have been used to study multiple cancer types, including BCC, indicating they are susceptible to tumorigenesis [7, 8]. This outbred background indicates the worth of testing the occurrence of BCC tumors in a UVB-exposed \(Ptch1^{+/−}\) CD-1 outbred mouse model. A controlled parent-progeny multi-generational study would be performed using \(Ptch1^{+/−}\) CD-1 outbred mice, with samples taken from germline and somatic cells so both inherited and acquired mutations could be examined in the context of HI. Not only would this combination provide an ideal heterozygous genomic landscape to explore HI, but it would also allow examination of potential mechanisms in meiosis, mitosis, recombination, replication, normal cell division, cancerous cell division, stem cells, oogenesis, spermatogenesis and embryogenesis. Combining whole-exome sequencing or whole-genome sequencing, proposed to characterize the genetic variation in mouse models of BCC with genome-wide array comparative hybridization, would allow for optimal testing of HI. Once evidence indicates
potential contexts where HI is occurring, repair deficient mice could be used to elucidate possible mechanisms.

Future studies can use the characterized genome of outbred mice to enrich research into the variation in diseases, complex phenotypes, and toxicology responses. Future research must continue to define a profile of heterozygous SNP clusters that are associated with proximal CNVs. The attributes included in a profile of heterozygosity in the genome should include the density of clusters, length of clusters, the number of heterozygous regions per chromosome, and the sequence context of the heterozygosity which will lead to a better understanding of Heterozygote Instability in germline cells. Controlled breeding experiments must be performed to track HI through meiosis in male and female gametes over multiple generations of outbred mice. After characterizing BCC mouse models, research can explore the mechanisms behind their benign nature yet very high mutation burden. A well-characterized model with a defined mutation landscape is not only crucial in exploring new preventative measures and targeted treatments but will be highly valuable as models of high genetic variation to study complex phenotypes and mutational mechanisms. Characterized mouse models of BCC are ideal for tracking HI through mitosis and replication in somatic cells and track mutations in rapidly dividing cells. Finally, researchers can apply the experimental design and bioinformatics pipeline provided here to define the mutation landscape of other mouse models of cancer and diseases. Mouse models offer a controlled environment to study mutational mechanisms. This thesis suggests an exquisite combination of two mouse models, outbred mice and mouse models of BCC, plus experimental design and bioinformatics pipelines to investigate Heterozygote Instability in germ and somatic cells during meiosis, mitosis, and typical and atypical cell division.
References


Appendices

Appendix 2A: 378 CD-1 and NMRI outbred mouse CEL files and sample IDs from The Center for Genome Dynamics at The Jackson Laboratory. (Online)

Appendix 2B: CNV and SNP genetic distance matrices for all pairwise comparisons of 378 sample genotypes. (Online)

Appendix 2C: Copy number variant landscape plot (CNV-LP) script. (Online)

Appendix 2D: Autosomal CNVs detected for 378 samples from both outbred mouse stocks. (Online)

Appendix 2E: X chromosome CNVs detected for male CD-1 and NMRI mice. (Online)

Appendix 2F: Summary of X chromosome CNV calls in male CD-1 and NMRI mice by copy number state.

<table>
<thead>
<tr>
<th>Stock</th>
<th>Number of Samples</th>
<th>CNV Calls</th>
<th>Copy Number State&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Loss/Gain&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD-1</td>
<td>43</td>
<td>96</td>
<td>2.23</td>
<td>95 (2.21)</td>
</tr>
<tr>
<td>NMRI</td>
<td>279</td>
<td>378</td>
<td>1.35</td>
<td>360 (1.29)</td>
</tr>
</tbody>
</table>

Values in brackets are normalized by sample count.

<sup>a</sup> Copy number states: full deletion (CN 0) and gain (CN 2+, 2 or more)

<sup>b</sup> Total number of losses (CN 0) divided by number of gains (CN 2+)

Appendix 2G: Copy number variant landscape plot (CNV-LP) data for CD-1 and NMRI mice. (Online)

Appendix 2H: Genes affected by CNVs in CD-1 and NMRI mice. (Online)

Appendix 2I: GO terms enriched in genes affected by CNVs in CD-1 and NMRI outbred mouse stocks. (Online)

Appendix 2J: KEGG pathways enriched in genes affected by CNVs in CD-1 and NMRI outbred mouse stocks. (Online)

Appendix 2K: PANTHER pathways enriched in genes affected by CNVs in CD-1 and NMRI outbred mouse stocks. (Online)
Appendix 2L: Rainfall, rainbow and J statistic plots depicting heterozygous SNP loci, spatial association between heterozygous SNP loci and CNVs and the statistical significance of the spatial association for each chromosome in every CD-1 and NMRI mouse sample. (Online)

Appendix 3A: Mutations in genes mutated in human BCC tumors identified by a systematic literature review. (Online)

Appendix 3B: PANTHER pathways enriched in genes with reported mutations in human BCC identified by the systematic literature review. (Online)

Appendix 3C: KEGG pathways enriched in genes with reported mutations in human BCC identified by the systematic literature review. (Online)

Appendix 3D: GO terms enriched in genes with reported mutations in human BCC identified by the systematic literature review. (Online)
Appendix 3E: Targeted next-generation sequencing summary for comparison to a BCC-specific targeted exome panel

<table>
<thead>
<tr>
<th>NGS Panel</th>
<th>Genes on Panel (#)</th>
<th>Genes in Common (#)</th>
<th>HH Pathway Genes</th>
<th>Genes Mutated in BCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>TruSight Tumor 15</td>
<td>15</td>
<td>6</td>
<td>None</td>
<td>BRAF, ERBB2, KRAS, NRAS, PIK3CA, TP53</td>
</tr>
<tr>
<td>AmpliSeq for Illumina Comprehensive Panel v3</td>
<td>161</td>
<td>28</td>
<td>PTCH1, SMO</td>
<td>ARID1A, BRAF, CDKN2A, CREBBP, ERBB2, ERBB4, FGFR3, FBXW7, JAK2, KDR, KNSTRN, Kras, MYCN, NOTCH1, NOTCH2, NOTCH3, NRAS, PIK3CA, PTCH1, PTEN, RAC1, RB1, SMO, TERT, TP53</td>
</tr>
<tr>
<td>TruSight Cancer Target Genes</td>
<td>94</td>
<td>9</td>
<td>PTCH1, SUFU</td>
<td>CDKN2A, HRAS, PTCH1, PTEN, RB1, SUFU, TP53, RUNX1, WT1</td>
</tr>
<tr>
<td>AmpliSeq for Illumina Cancer Hotspot Panel</td>
<td>52</td>
<td>11</td>
<td>SMO</td>
<td>BRAF, CDKN2A, ERBB2, ERBB4, FBXW7, FGFR3, HRAS, JAK2, KDR, Kras, NOTCH1, NRAS, PIK3CA, PTEN, RB1, SMO, TP53</td>
</tr>
<tr>
<td>AmpliSeq for Illumina Focus Panel</td>
<td>50</td>
<td>17</td>
<td>SMO</td>
<td>BRAF, ERBB2, ERBB4, FGFR3, HRAS, JAK2, KRAS, MYCN, NRAS, PIK3CA, SMO</td>
</tr>
<tr>
<td>TruSight Tumor 170</td>
<td>170</td>
<td>25</td>
<td>PTCH1, SMO</td>
<td>ARID1A, BRAF, CCND3, CDKN2A, CREBBP, ERBB2, ERBB4, FBXW7, FGFR3, HRAS, JAK2, KDR, Kras, MYCN, NOTCH1, NOTCH2, NOTCH3, NRAS, PIK3CA, PTCH1, PTEN, RB1, SMO, TERT, TP53</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>NGS Panel</th>
<th>Genes on Panel (#)</th>
<th>Genes in Common (#)</th>
<th>HH Pathway</th>
<th>Genes Mutated in BCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AmpliSeq for Illumina Comprehensive Panel</td>
<td>409</td>
<td>37</td>
<td><strong>PTCH1, SMO, SUFU</strong></td>
<td><strong>ARID1A, ASXL1, BRAF, CDKN2A, CREBBP, CSMD3, CTNNA1, DCC, EPHA3, ERBB2, ERBB4, FBXW7, FGFR3, HRAS, JAK2, KDR, KEAP1, KRAS, LRP1B, MLL2, MYCN, NOTCH1, NOTCH2, NRAS, PIK3CA, PTCH1, PTEN RB1, RUNX1, SMO, SMARCA4, STK36, SUFU, TET2, TSC1, TP53, WT1</strong></td>
</tr>
<tr>
<td>TruSight Oncology 500</td>
<td>500</td>
<td>58</td>
<td><strong>PTCH1, SMO, SUFU, GLI1</strong></td>
<td><strong>ARID1A, BRAF, CASP8, CD274, CCND3, CDKN1A, CDKN2A, CREBBP, CTNNA1, E2F3, ERBB2, ERBB4, EPHA3, FAT1, FBXW7, FGFR3, GLI1, GRIN2A, GRM3, HRAS, JAK2, KDM5A, KDR, KEAP1, KEL, KRAS, LATS1, LATS2, LRP1B, MAGI2, MLL, MYCN, NOTCH1, NOTCH2, NOTCH3, NRAS, PIK3CA, PIK3RI, PPP6C, PREX2, PTCH1, PTEN, RAC1, RB1, RBM10, RUNX1, SLIT2, SMARCA4, SMO, STAT5B, SUFU, TET2, TERT, TSC1, TGFB1, TP53, WT1, YAP1</strong></td>
</tr>
<tr>
<td>Skin Cancer NGS Panel</td>
<td>6</td>
<td>3</td>
<td><strong>PTCH1, SUFU</strong></td>
<td><strong>CDKN2A, PTCH1, SUFU</strong></td>
</tr>
</tbody>
</table>
# Curriculum Vitae

## Hailie A. Pavanel

### Education

<table>
<thead>
<tr>
<th>Institution</th>
<th>Years</th>
<th>Program</th>
</tr>
</thead>
<tbody>
<tr>
<td>The University of Western Ontario</td>
<td>2018 – 2021</td>
<td>M.Sc. Biology, Cell and Molecular Biology</td>
</tr>
<tr>
<td>The University of Western Ontario</td>
<td>2013 – 2018</td>
<td>B.Sc. Honours Specialization in Genetics and Biochemistry</td>
</tr>
</tbody>
</table>

### Academic Awards

<table>
<thead>
<tr>
<th>Award</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetics Society of America Travel Award</td>
<td>2020</td>
</tr>
<tr>
<td>Ruth Horner Arnold Fellowship</td>
<td>2019</td>
</tr>
<tr>
<td>Department of Biology Graduate Travel Award</td>
<td>2019</td>
</tr>
</tbody>
</table>

### Published Work


### Presentations

#### (Oral)


#### (Poster)


“Available mouse models of basal cell carcinoma provide the means to analyze and explore the mystery of one of the highest mutation burdens within a benign human cancer,” the Terry Fox Research Foundation Ontario Node Research annual symposium, Toronto, Canada, 2018.


Teaching Experience

The University of Western Ontario 2018 – 2020
Teaching Assistant, Department of Biology
3594B: Genome Organization, Mutagenesis and Repair;
3592A: Principles of Human Genetics;
3326F: Cell Biology Laboratory;
1001A: Biology for Science I