8-24-2020 11:00 AM

Heterozygosity: An Inconspicuous Meiosis-Linked Intrinsic Mutagen in Mice

Nicholas A. Boehler, *The University of Western Ontario*

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

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

Abstract

The impact of heterozygosity as an intrinsic mutagen in mammals is unknown. In plant models, existent heterozygosity increases the local *de novo* meiotic mutation rate. Mice offer study of this phenomenon given well-established genomic technologies and strains with known, diverse genomic landscapes of heterozygosity. High resolution genotyping arrays assay heterozygous single-nucleotide polymorphic (SNP) loci and copy number variants (CNVs). Using a J statistic for spatial analysis, 60.9% of autosomes from 707 publicly available array samples have nonrandom spatial associations between heterozygous SNP loci and CNVs. By crossing C57BL/6J inbred mice to DBA/2J inbred mice, heterozygous SNP loci and *de novo* CNVs were analyzed. Of 43 *de novo* CNVs in F2 mice compared to both F1 and F2 heterozygous SNP landscapes, 33 and 7 were found to co-localize with heterozygous SNP loci, respectively. Heterozygosity may be an overlooked meiosis-linked contributor to CNV mutagenesis, affecting models of disease risk prediction and evolution.

Keywords: Mutagenesis, heterozygosity, copy number variant, single-nucleotide polymorphism, genomic spatial statistical tool, genotyping microarray

Summary for Lay Audience

Understanding the factors that contribute to the rate, type, and distribution of DNA mutations across the genome is paramount to fully comprehending diversity in the form and function of an organism, development of genetic disorders, and the process of evolution. Heterozygosity – the condition of having two different nucleotide sequences at the same location between parental chromosomes – has been shown to exist nearby new and elevated levels of mutations arising during development of germ cells in plants. However, this relationship is yet to be demonstrated in mammals. Mice are ideal to study this due to the availability of strains with known diverse landscapes of heterozygosity. Single-nucleotide polymorphic (SNP) loci are sites within a genome that have variable nucleotide content between individuals in at least 1% of the population and therefore may be sampled for heterozygosity. Copy number variants (CNVs) are a type of mutation characterized as large DNA duplications or deletions. Microarrays designed to target hundreds of thousands of sites across the genome are used to detect SNP heterozygosity and CNVs. I contributed to development of a spatial statistical analysis pipeline to determine whether heterozygous SNP loci and CNVs are nearby, far apart, or randomly distributed for 707 publicly available samples. I found that in 3,533 of 5,799 chromosomes with CNVs, heterozygous SNP loci and CNVs are nonrandomly distributed with respect to one another. I also generated six three-generation lineages of mice, crossing two different low heterozygosity inbred strains to produce F1 mice with an average of 20% SNP heterozygosity. Brother-sister mating of F1 mice produced F2 mice with an average of 10% SNP heterozygosity. Microarray analysis followed by application of the analytical pipeline showed that 1,024 of 1,338 chromosomes with CNVs had a nonrandom spatial relationship between heterozygous SNP loci and CNVs. I identified 43 CNVs in F2 mice that were not inherited from their parents.

Interestingly, 33 of the non-inherited CNVs were nearby the heterozygous SNP loci found in their F1 parents. These findings indicate that heterozygosity may contribute to the formation of CNVs, therefore demanding reassessment of predictions of disease risk and evolutionary change.

Co-authorship Statement

I completed this work under the supervision and financial support of Dr. Kathleen Allen Hill who is a co-author on all works stemming from this thesis.

I performed all of the experimental work presented in this thesis with assistance in applying spatial statistical analysis by Dr. Bin Luo, Dr. Charmaine Dean, and Dr. Reg Kulperger. Further assistance was given by Dr. Camila de Souza, Hailie Pavanel, and Steven Villani in creating R scripts for statistical analysis and pipeline automation.

The work presented in this thesis contains a primary research manuscript in preparation for publication with Dr. Kathleen Allen Hill as co-author.

Sections of this thesis pertaining to the application of the J statistic are contributions to a manuscript in preparation with co-authors Dr. Bin Luo, Hailie Pavanel, myself, Freda Qi, Dr. Charmaine Dean, Dr. Kathleen Allen Hill, and Dr. Reg Kulperger.

Sections of this thesis pertaining to the breeding schemes, genetic backgrounds, and genotypes of the mouse cohorts are contributions to a manuscript in preparation testing the classification of mouse SNP genotypes using Machine Learning with Digital Signal Processing (MLDSP). Co-authors in this work include Hailie Pavanel, Dr. Gurjit Singh Randhawa, myself, Ali Coyle, Pok Wan, Dr. Lila Kari, and Dr. Kathleen Allen Hill.

Acknowledgements

I have been lucky to have an abundance of support from my supervisor, family, and friends in the pursuit of my Master's degree. Firstly, I would like to thank Dr. Kathleen Hill for her unwavering encouragement, guidance, and feedback throughout the course of this thesis. Her patience, teaching, and careful critiques were imperative to helping me achieve my academic aspirations. I would also like to express my gratitude to my advisors Dr. Marc-André Lachance, Dr. Jamie Kramer, and Dr. David Smith for their invaluable input during this project. I am particularly grateful to Dr. Lachance for serving as my reader and providing exceptional constructive criticism which dramatically improved my scientific writing ability.

I would like to thank my wife, Bridget Murphy, for her unfaltering emotional support and immense patience as I navigated the writing of this work. I would not have been able to complete this thesis without her encouragement for which I will always be grateful. I would also like to thank my son, Jamie, for always brightening my day and lifting my spirits, no matter how stressful life became.

I would also like to acknowledge a number of people for their expert assistance throughout my research project. Thank you to Dr. Bin Luo, Dr. Reg Kulperger, and Dr. Camila de Souza for their technical and statistical expertise. Thank you to Steven Villani for spending uncounted hours helping tackle my R coding problems. As well, I would like to thank Rachel Kelly and Hailie Pavanel for their innumerable research suggestions and making the lab environment a welcoming and enjoyable place to do research.

This research was funded by a Natural Sciences and Engineering Research Council of Canada Discovery Grant award to Dr. Kathleen Hill. This research was also supported by additional

funding awarded to me by the Queen Elizabeth II Graduate Scholarship in Science and Technology. Financial support for conference attendance was provided by the Department of Biology Graduate Travel Award, and Environmental Mutagenesis and Genomics Society Student and New Investigator Travel Awards.

Contents

List of Figures

List of Tables

List of Appendices

List of Abbreviations

bp	Base pair
BRLMM-P	Bayesian Robust Linear Model with Mahalanobis distance classifier - Perfect match
CD1	Caesarian derived-1
CEL	Cell intensity file [file extension]
CI	Classical inbred
CNV	Copy number variant
CSV	Comma-separated values
D-loop	Displacement loop
Distal association	A distant relationship between two genomic events
DNA	Deoxyribonucleic acid
DNA-PKcs	DNA dependent protein kinases
DSB	Double-stranded break
FTP	File transfer protocol
HD-CNV	Hotspot detector for copy number variants
HDR	Homology-directed repair
HMM	Hidden Markov Model
IGP	Invariant genomic probe

Indel	Insertion or deletion of nucleotides (< 30bp)
Kb	Kilobase (1,000 bp)
LCR	Low copy repeat
Mb	Megabase (1,000,000 bp)
MDGA	Mouse diversity genotyping array
NAHR	Non-allelic homologous recombination
NGS	Next generation sequencing
NHEJ	Non-homologous end joining
NMRI	Naval Medical Research Institute
Proximal association	A nearby relationship between two genomic events
QC	Quality control
RI	Recombinant inbred
SPO11	SPO11 initiator of meiotic double stranded breaks
SNP	Single nucleotide polymorphism
WD	Wild-derived

Introduction

1.1 Copy number variants: An underappreciated source of genetic variation and disease

DNA mutation is the fundamental fuel for the process of evolution, responsible for creating an immensely rich diversity of life. While the limelight of DNA mutation research has historically emphasized understanding how changes to single-nucleotides arise, the prevalence and impact of larger-scale structural mutation is only recently becoming apparent [2]. Copy number variants (CNVs) are one such large structural mutation. CNVs are deletions or duplications of regions of DNA historically considered 1 kilobase (Kb) or greater in size [3]. In response to improved CNV detection capacity by next generation sequencing (NGS) technologies, the minimum length is now often considered to be as little as 50 base pairs (bp) [4–6].

With the advent of single-nucleotide polymorphic (SNP) genotyping microarrays and DNA sequencing technologies, the startling impact of CNVs is becoming apparent. While the estimated rate of large *de novo* CNVs arising per generation in humans is a modest 1–2% [7–9], the larger size of these mutations results in a substantial contribution to evolutionary change. On average, more than twice as many total nucleotides that differ between human and chimpanzee genomes consist of CNV duplication compared to single base pair substitutions [10].

The departure from wild type ploidy state caused by CNVs can have significant phenotypic consequences. For example, in humans, gain or loss of gene dosage contributes to the etiology of complex traits such as neuropsychiatric disorders such as autism spectrum disorder and schizophrenia [11–13]. Deletions involving the *CHRNA7* gene have been shown to cause

epilepsy [14]. Duplications involving the *MECP2* gene in males are associated with developmental delays and intellectual disability [15]. Other complex human traits and diseases with CNV associations include Alzheimer disease, Parkinson's disease, and HIV-1/AIDS susceptibility [16–18].

Phenotypic impacts of changes to gene dosage caused by CNVs have been studied in other organisms as well. For example, duplications have been detected in bacteria of dosage-sensitive genes that confer antibiotic resistance [19]. Tandem triplication of *AMTE1* genes in maize is associated with aluminum resistance [20]. Increased copy numbers of the *EPSPS* gene in a variety of weed species has been linked to glyphosate herbicide resistance [21].

CNV analysis has also been applied to a variety of agriculturally relevant animals. CNVs have been studied in chickens in relation to body weight, egg laying, and muscle and body organ growth [22]. In pigs, CNVs have been associated with changes to backfat and intramuscular fatty acid composition and growth [23]. In cattle, CNVs contribute to phenotypes related to meat tenderness and milk composition traits [24, 25].

1.2 Copy number variants can be inherited or arise *de novo*

The vast majority of an organism's genetic information is inherited from its parents and is considered constitutive. However, rarely, DNA sequences arise that cannot be traced back through either the paternal or maternal lineage and are considered *de novo* and acquired. Pre-zygotic *de novo* mutations are referred to as germline mutations that become incorporated into the DNA of every cell of the body of the offspring. Post-zygotic *de novo* mutations are referred to as somatic mutations that are only incorporated into the DNA of the daughter cells of the affected progenitor cell.

The genome-wide rate of formation of CNVs that are larger than 500 bp in size per genome per generation has been conservatively estimated to be 3 x 10^{-2} from human HapMap data [26]. The rate of very large (>100 Kb) CNV events has also been conservatively estimated at 1.2 x 10^{-2} CNVs per genome per generation [27]. The distribution of CNV events across genomes is not random. Centromeric and telomeric regions harbour more and repeated CNV events and

are considered hotspots while gene-rich and gene dosage-sensitive regions harbour fewer CNV events and are considered cold spots [28, 29].

One substantial feature contributing to the formation of CNVs is the presence of multiple low copy repeats (LCRs) in genomic regions spanning up to 10 Mb. LCRs are often pseudogenes, nonfunctional superfluous duplications of genes, or repetitive DNA elements >1 Kb in size with >95% sequence identity that have been found to be hotspots for duplications or deletions [30]. The presence of two or more LCRs near DNA damage events along a chromosome can promote and mediate formation of CNVs via errors during DNA repair [31]. Recurrent CNVs, CNVs found to arise independently within unrelated members of a population, are usually flanked by LCRs [32]. Non-recurrent CNVs are not necessarily flanked by repetitive elements and generally have unique breakpoints [33].

1.3 Heterozygosity may be an unknown mutagenic contributor to copy number variant formation

The differences in the DNA sequence between distinct individuals of a given species represent the genetic diversity of that species. The extent of genetic diversity contributes to the ability of a species to respond favourably to environmental changes. Decreased levels of genetic diversity within a population have been associated with declines in population fitness and increased risk of extinction [34, 35].

The vast majority of eukaryotes are diploid organisms, possessing genomes consisting of two homologous sets of chromosomes. Heterozygosity is defined as the proportion of sites on a chromosome at which the maternal and paternal DNA sequence differ and is a valuable parameter in estimating the genetic diversity of a population. On average, heterozygosity has been found to be 35% lower in threatened taxa in comparison to nonthreatened taxa and indicates a decreased reproductive fitness and elevated extinction risk in the wild [36].

Single-nucleotide polymorphic (SNP) loci are genomic sites that contain sequence differences at a single base pair that appear in at least 1% of the population. SNP loci therefore have at

least two possible nucleotides that may be found in an individual. The possible nucleotides for a given SNP locus may be treated as alleles. Given that most SNP loci are biallelic, the alleles are usually categorized as the major (A) allele or the minor (B) allele. The genotype of a biallelic SNP locus is typically reported as homozygous AA, homozygous BB, or heterozygous AB. SNPs are one of the largest contributors to genomic variability between individuals [37]. In mice, over 70 million SNPs have been identified [38]. SNP loci are relatively easy and inexpensive to assay for allele composition, making them a convenient way of sampling heterozygosity in a genome.

Heterozygosity of certain loci can play an direct role in increasing organismal fitness. Heterozygote advantage refers to a scenario wherein a heterozygous genotype confers a fitness advantage when compared to either of the alternate homozygous genotypes. For example, in humans, individuals who are heterozygous for a mutation in the *HBB* gene for sickle cell anemia do not manifest the disease phenotype but also possess resistance to malarial infection [39]. In this case, heterozygous individuals have a fitness advantage in comparison to individuals who are homozygous for either the dominant allele or recessive, disease-causing allele.

While heterozygosity may be a beneficial genetic feature in some contexts, recent evidence suggests that elevated numbers of heterozygous loci may be mutagenic. A study by Yang *et al*. (2015) demonstrated a significant increase of mutation rates in relation to increased heterozygosity in *Arabidopsis*, a small flowering plant and widely used model organism [1]. By interbreeding purebred parental plants of differing ecotype, *Col* and *Ler*, they produced high heterozygosity F1 plants. The nucleotide diversity, a measure of the average number of nucleotide difference per site between two DNA sequences, between *Col* and *Ler* is approximately 0.39%. After selfing of both the low heterozygosity parentals (P0 $\rightarrow$ P1) and the high heterozygosity F1 (F1 $\rightarrow$ F2) plants, parent-progeny groups were subjected to high read-depth next generation sequencing (NGS). Single-nucleotide point mutations and small (<30 bp) insertion-deletion (indel) mutations were detected (Fig 1.1). Single-nucleotide point mutations were 3.6-fold higher in F2 plants relative to their P1 counterparts. Further, a 2.8-fold increase in the rate of indels in intergenic regions was also found in the F2s compared to the P1s. Heterozygos-

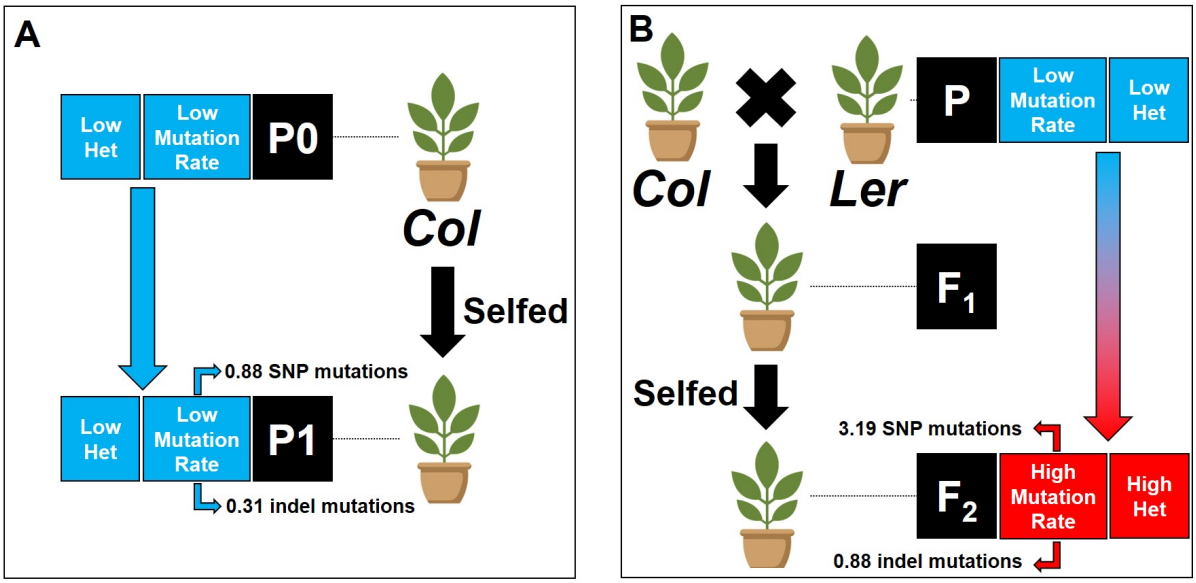

Figure 1.1: Engineered low and high heterozygosity *Arabidopsis* breeding schemes by Yang *et al*. [1]. (A) Selfing of purebred *Col Arabidopsis* plants produced plants of similar heterozygosity and a detected SNP point mutation rate of 0.88 per generation and an indel mutation rate of 0.31 per generation. (B) Crossing purebred parental plants of differing ecotypes, *Col* and *Ler*, produced plants of high heterozygosity. Selfing high heterozygosity F1 plants produced F2 plants with high heterozygosity and a detected SNP point mutation rate of 3.19 per generation and an indel mutation rate of 0.88 indel mutation per generation.

ity was implicated as the culprit for mutagenesis after F2 plants were selfed for two more generations, depleting genomic heterozygosity by approximately one-half for each generation. Sequencing of F3 and F4 plants demonstrated a positive correlation between heterozygous loci and single-nucleotide mutation rate. Mounting evidence supporting heterozygosity as an endogenous mutagen was further strengthened as the median distance (167 bp) of 273 *de novo* mutations to heterozygous sites in the F2 plants was significantly smaller than expected. Further, significantly more *de novo* point mutations were found in heterozygous regions compared to homozygous regions.

The finding of elevated mutagenesis in a high heterozygosity genome has since been corroborated in peaches [40]. Using a similar approach to the *Arabidopsis* study, Xie *et al*. generated a low heterozygosity (0.27% nucleotide diversity) F1 peach tree from an intraspecific cross and a high heterozygosity (1.24% nucleotide diversity) F1 peach tree from an interspecific cross. Both F1 trees were selfed to generate F2s and the level of SNP point mutation and indel rates were assessed by parent-progeny NGS. The high heterozygosity interspecific F2 group had a 1.8 fold increase in the relative SNP point mutation rate and a 1.7 fold increase in the indel (<30 bp) mutation rate. Additionally, the average local heterozygosity for regions surrounding *de novo* mutations was up to 1.5-fold higher than the average genomic heterozygosity in the interspecific F2.

Both the *Arabidopsis* and peach studies revealed compelling evidence that heterozygosity is associated with an elevated number of point and indel mutations that occur nearby heterozygous loci. This nearby spatial association between heterozygous loci and mutations can be termed a proximal association, which will be used throughout this thesis. However, there are gaps in knowledge yet to be addressed. Neither study addressed whether heterozygosity is associated with larger structural mutations such as CNVs. As well, the pervasiveness of this phenomenon in species outside of plants remains limited. Insight into the precise mechanism by which heterozygosity could contribute to mutagenesis is not fully understood. One potential hypothesis for heterozygosity directly promoting and mediating mutagenesis is that an elevated level of heterozygosity could cause chromosomal misalignment during meiotic recombination and lead to improper DNA repair [41]. Alternatively, heterozygosity may co-localize with mutational

events but not itself contribute to the development of new mutations.

1.4 Imperfect DNA repair contributes to mutagenesis resulting in copy number variants

There are a number of a mechanisms by which CNVs are known to arise. Two predominant mechanisms of mutagenesis leading to CNV formation occur during repair of double strand DNA breaks (DSBs), a highly cytotoxic DNA lesion [4]. The frequency of DSBs in human and mouse fibroblasts has been estimated to be ten per day per cell, thus requiring cells constantly to repair their DNA [42–44]. In response to DSBs, cells will attempt to repair the damage either through non-homologous end joining (NHEJ) or homology-directed repair (HDR) pathways. The NHEJ and HDR repair mechanisms differ in their capacity for accurate repair, efficiency, and template requirements [45].

Errors associated with NHEJ and HDR mechanisms have been implicated as the two major pathways by which large CNVs arise. One study found that of 227 CNVs larger than 7 kb from eight human genomes, 39% likely arose due to improper NHEJ and 38% arose due to improper HDR [46]. The remaining large CNVs are thought to have formed primarily due to retrotransposition events and variable nucleotide tandem repeats.

The repair mechanism NHEJ is kinetically favourable and simply ligates detected DSBs back together but does not depend on sequence homology to proceed (Fig 1.2) [47]. When a DSB occurs, the DNA ends are bound by the Ku70/Ku80 heterodimer, which recruits DNA-dependent protein kinases (DNA-PKcs) and, if necessary, nucleases (such as Artemis) or polymerases to generate compatible ends [48]. Finally, ligation of the DSB is completed by a ligation complex made up of DNA ligase IV, XRCC4 and XLF [49, 50]. The NHEJ mechanism is known to be error-prone and often introduces short insertions and deletions (usually <10 bp) during the strand resection or elongation phase [51]. While short indels are a common outcome of NHEJ, larger duplications and deletions do occur.

HDR is a high fidelity repair mechanism that depends on sequence complementarity between

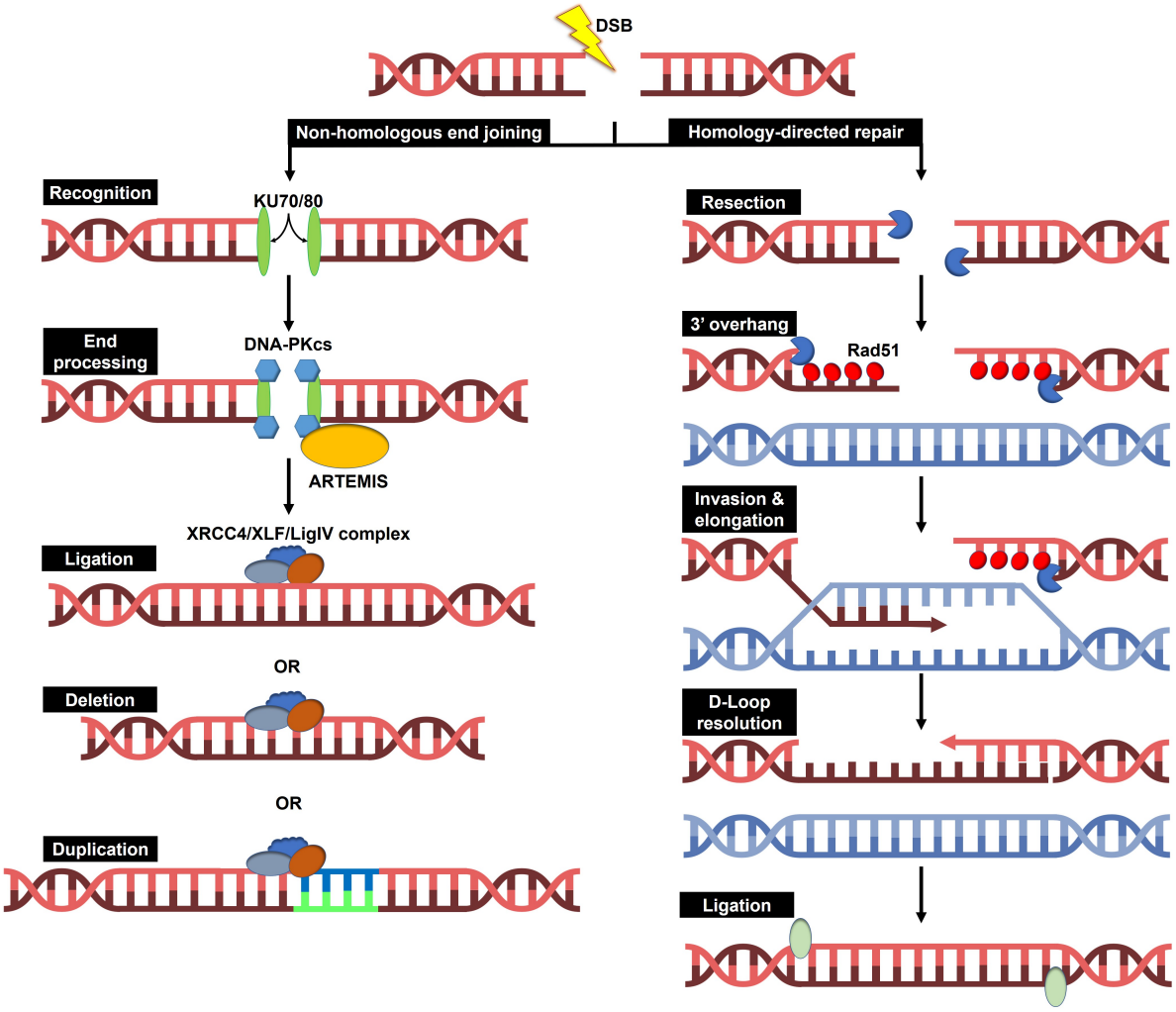

Figure 1.2: DNA double-stranded breaks can be repaired through non-homologous end joining or homology-directed repair.
Red DNA molecules represent broken strands to be repaired. Blue DNA molecules represent homologous sequences used as a template for repair. Non-homologous end joining begins with strand breakage recognition by KU70/KU80 heterodimers that recruit DNA dependant protein kinases in addition to nucleases (such as ARTEMIS) or polymerases to generate compatible ends. A ligation complex made up of XRCC4/XLF/LigIV repairs the sugar-phosphate backbone. Homology-directed repair begins with strand resection by endonucleases that leave 3' overhangs that are coated by Rad51 proteins that catalyze the recognition of sequence complementarity and strand invasion. The 3' overhang is elongated across the breakage site using the homologous DNA as a template before the D-loop is resolved. The remaining strand gap is elongated before the sugar-phosphate backbone is ligated and repair is complete.

the DNA strand to be repaired and a template strand, typically derived from the undamaged parental chromosome (Fig 1.2). During this repair process, the DSB lesion first undergoes single stranded resection of the 5' ends to generate 3' single stranded DNA overhangs [52]. The 3' overhangs are coated with recombinase proteins (such as Rad51) which promote strand invasion of the damaged DNA to the template [53]. Importantly, strand invasion relies upon sequence complementarity of the damaged strand to the template. Strand invasion results in the formation of a displacement loop (D-loop) bubble, where the 3' end of the damaged DNA is now elongated until there is sufficient overlap between it and the other damaged DNA strand [54]. The D-loop is then resolved, with the 3' overhangs of the damaged strands annealing and polymerase elongating the single stranded segments. Finally, ligase repairs the nicks in the backbone and the lesion is repaired [55].

While HDR is typically a highly faithful method of DNA repair, it is not infallible. HDR can suffer from non-allelic homologous recombination (NAHR), a mishap in repair wherein an incorrect template with a near-identical sequence is used during strand invasion instead of the proper template [30]. Large genomic rearrangements can arise from NAHR, including duplications, deletions, and inversions. Significantly, NAHR has been found to contribute directly to CNV deletions responsible for autism spectrum disorder and Williams-Beuren syndrome phenotypes [56].

The HDR mechanism is primarily active during the G2 and S phases, but NAHR can occur outside of a mitotic setting [57]. In eukaryotes, homologous recombination (HR) occurs during meiosis, playing a critical role in increasing genetic diversity by shuffling genetic material during chromosomal crossover [58]. In similar fashion to HDR, the process of HR involves strand invasion of highly homologous sequences and is susceptible to using an incorrect template for repair [59]. The prevalence and impact of NAHR during mitosis and meiosis remains unclear, although headway has been made in determining factors affecting NAHR rates such as: distance, alignment length, sequence complementarity, and chromosomal position [60]. Still, interpretation of these factors alone still leaves the exact origins of many CNVs unknown and heterozygosity has not been specifically investigated as a contributor. However, during repair of a DSB, strand invasion depends upon complementarity of similar and not necessarily

identical sites. It is reasonable to consider heterozygosity as a factor that may increase the rate at which a non-allelic template is used for repair, increasing the likelihood of CNV formation in regions of heterozygosity.

1.5 Interbreeding different classical inbred mice produces a good model for investigating the contribution of heterozygosity to copy number variant formation in mammals

The common house mouse, *Mus musculus*, has long been the staple mammalian model organism for genetic research due to phylogenetic relatedness and physiological similarities to humans [61]. Additionally, the ease of laboratory breeding has permitted the curation of many inbred strains over the last century. Classical inbred mouse strains are generated through brother-sister mating each generation, causing heterozygosity levels to be reduced by one-half with each generation. To be considered classically inbred, a mouse must be the product of at least 20 consecutive generations of brother-sister mating [62]. At 20 generations, at least 98.6% of the loci will be homozygous [63]. Some of the oldest inbred strains such as C57BL/6J (B6) and DBA/2J (DBA) have been inbred for over a century and are homozygous at virtually all of their loci.

Classical inbred mice serve as effective genetic control populations for a number of research applications. As individuals within a strain are designed to be as close to isogenic (genetically identical) as possible, experimental reproducibility is not confounded by genetic variation. With the widespread use of inbred mice, a vast arsenal of molecular tools have been developed to manipulate their genome and, in parallel, high-throughput microarrays and NGS have permitted the development of information-rich databases. For example, as of August 2020, 32 laboratory mouse strains have had their genomes sequenced. [38, 64–66].

Interbreeding of two different parental inbred mouse strains results in F1 hybrid mice that are heterozygous at all loci at which their parents differ, but remain homozygous at all loci at which

their parents are the same. F1 hybrids are useful in that their genomes contain a predictable amount of nonrandomly distributed heterozygosity in all individuals that may be controlled based on the selection of parental strains. Derived from low heterozygosity parents, F1 mice are a good model for investigating the effect of increased heterozygosity on mutagenesis leading to CNV formation in a reproducible genetic context.

Importantly, the cells undergoing gametogenesis that produce these F1 mice will have lower heterozygosity in nature. In order to evaluate the effects of heterozygosity on mutagenesis in a meiotic context, F1 mice will need to be brother-sister mated to produce F2 mice. If heterozygosity is associated with CNV formation in a meiotic context, it can be expected that only the F2 mice would demonstrate an elevated number of heterozygosity-linked *de novo* CNVs. If, however, heterozygosity is associated with CNV formation in a mitotic context, it can be expected that both F1 and F2 mice would demonstrate an elevated number of heterozygosity-linked *de novo* CNVs.

1.6 Microarray technology is a high-throughput method for sampling heterozygosity and *de novo* copy number variants in mice

SNP genotyping microarrays are a well-established, high-throughput technology used for determining the genotype at hundreds of thousands of genomic loci at a time. These arrays have been developed for a wide range of species, including humans, mice, cattle, chickens, and many others [67–70]. In addition to providing genotypes of SNP loci, genotyping microarrays permit the detection of large CNVs across the genome (Fig 1.3).

Manufacturing of SNP genotyping microarrays can involve spotting single stranded DNA probes to a glass slide that are complementary to SNP loci to be interrogated [71]. Genotyping array calls are made by extracting and amplifying whole genomic DNA, followed by fragmentation and fluorescent labelling. The labelled DNA is then hybridized to the array, where fragments of high sequence homology bind the affixed probes. After washing away the

unbound and imperfectly hybridized DNA, the microarray is scanned and a fluorescence intensity image is captured [72]. The fluorescence intensity emitted by a specific probe spot is proportional to the amount of DNA binding.

Separate probes interrogating the more common A allele and less common B allele. Relative fluorescence can be interpreted by genotype clustering algorithms to indicate whether a SNP locus is homozygous AA, homozygous BB, or heterozygous AB. Samples with fluorescence intensities outside of the genotype call clusters are designated 'no calls' and may indicate the presence of a genotype not interrogated by the array, a CNV gain or loss, inefficient hybridization of sample DNA to the probe, or insufficient algorithm training to make accurate genotyping calls (Fig 1.4).

Development of high-density chipsets has permitted the detection of CNVs alongside SNP genotyping [73]. With a greater number of probes being assayed, the resolution at which the genome can be interrogated increases while the average inter-probe distance shrinks. CNVs can be called when the fluorescence intensity for at least three consecutive probes deviate from the expected diploid fluorescence intensity. In this way, copy number states of zero (deletion of both alleles), one (deletion of one allele), or three or more (allele duplication) can be distinguished (Fig 1.4). For example, 20 consecutive probes with a relative fluorescence one half of the expected diploid intensity is indicative of a copy state one deletion.

1.7 Publicly available Mouse Diversity Genotyping Array data are ideal for investigating copy number variant mutagenesis

The Mouse Diversity Genotyping Array (MDGA) is the most probe-dense SNP genotyping array to date, targeting 624,124 SNP loci [68]. Each SNP locus is targeted by a total of eight probes to ensure redundancy and improve calling accuracy. Two pairs of duplicate probes target the A allele on the forward and reverse strands while the other two pairs of duplicate probes target the B allele on the forward and reverse strands. The total fluorescence intensity

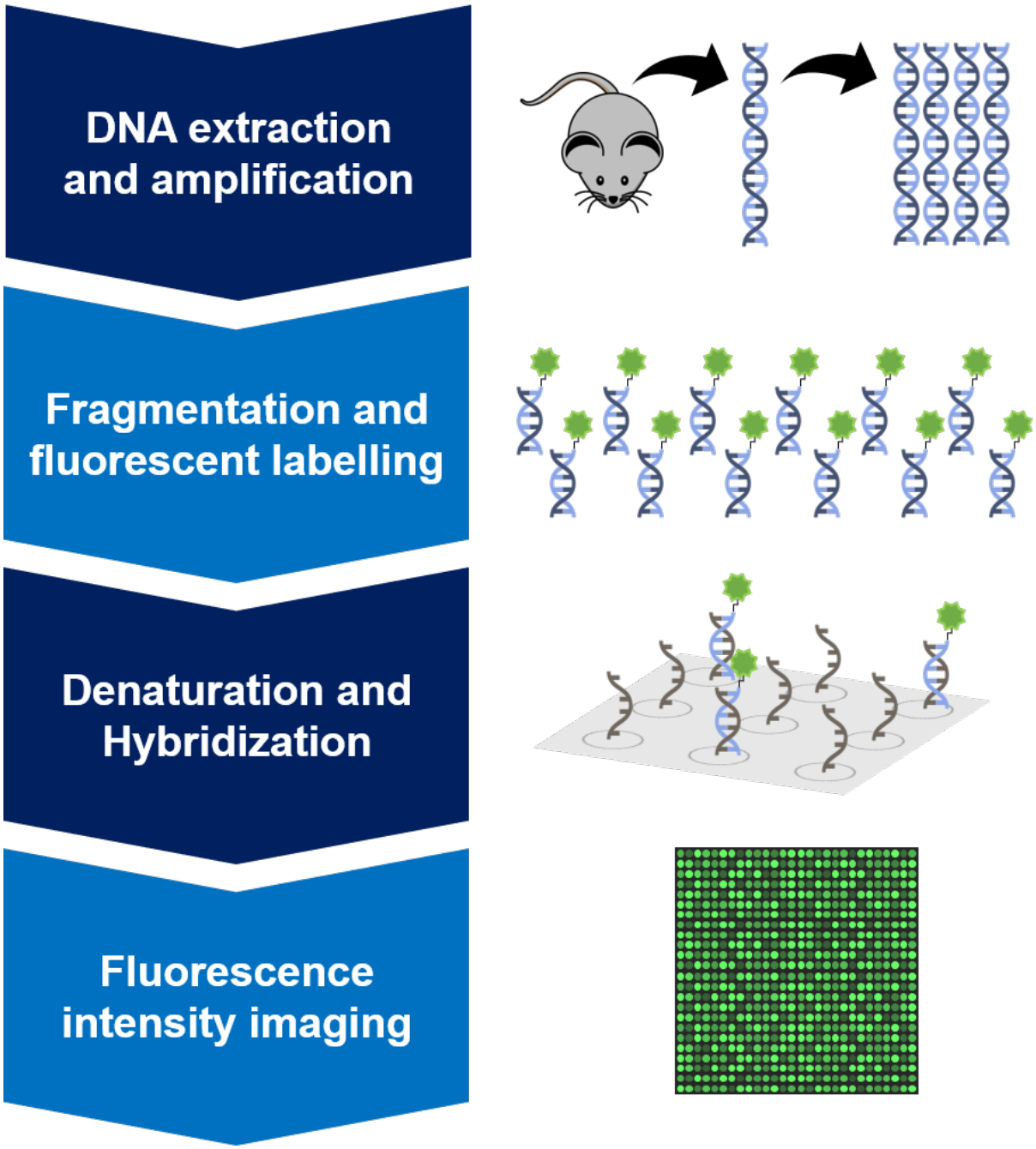

Figure 1.3: Microarray experimental design methodology. DNA is first extracted and purified from the sample of interest. Following PCR amplification, the DNA is fragmented and fluorescently labelled. DNA is then denatured and washed across the microarray, permitting hybridization to complementary single stranded probes spotted on the array. Unbound DNA is washed away and a fluorescent image is captured. Relative fluorescence intensity is used to ascertain single-nucleotide polymorphic loci genotypes and detect copy number variants through the use of various software.

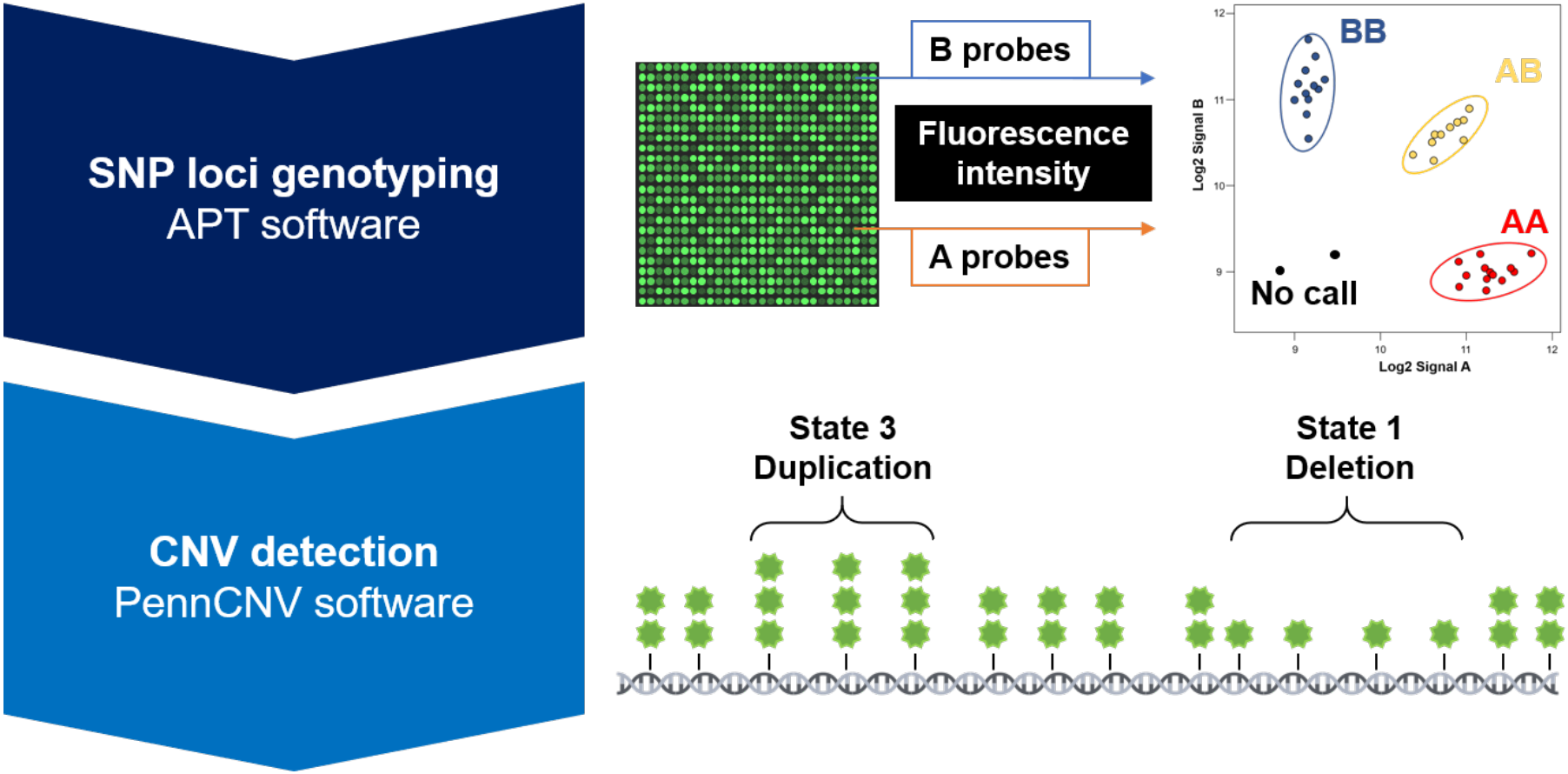

Figure 1.4: SNP loci genotyping and CNV detection methodology. Using fluorescence intensity image (.CEL) files, SNP loci genotyping and CNV detection can be conducted using Affymetrix Power Tools (APT) and PennCNV software, respectively. SNP loci genotyping software evaluates the relative fluorescence of the A and B probes for each training sample at each locus and uses genotype clustering algorithms to form call clusters AA, AB, and BB. Test samples are then called AA, AB, BB or no call depending on their signal. CNV detection is performed using PennCNV software. Relative fluorescence of consecutive probe markers is assessed using a Hidden Markov Model. CNV duplications and deletions are called by stretches of at least 3 probe markers deviating in their relative fluorescence from the diploid state.

values for all like probes targeting a locus are taken into account during genotyping to reduce the number of false-positive calls.

The MDGA further targets 916,269 loci queried by invariant genomic probes (IGPs) [68]. IGPs assay conserved regions of the genome, where sequence information is expected to be identical amongst all individuals. IGPs are particularly useful for identifying putative CNVs because the relative fluorescence of these probes is expected to remain the same for all diploid individuals within the genus *Mus*. Therefore, fluorescence deviations in IGPs are more straightforwardly and reliably interpreted by various CNV calling algorithms [74], although redundancy for these probes is reduced in comparison to SNP loci because there are only two IGPs for each queried locus. The MDGA has been further optimized for CNV analysis by filtering the SNP probeset to reduce possible sources of noise and false positive calling. There are 493,290 well performing SNP probes on the MDGA [75–77].

One of the benefits of the MDGA is the abundance and diversity of publicly available sample data. The Jackson Laboratory has made 1901 MDGA samples of varied genetic background publicly available, ranging from classical inbred B6 mice all the way to cross-species applications to the tapir and rhinoceros [78]. A subset of 334 of classical inbred, wild-derived inbred, and wild-caught mice has undergone thorough CNV characterization, identifying 9,634 putative CNVs affecting 6.87% of the mouse genome [75]. While the nature of CNV recurrence, distribution and gene-overlap was rigorously profiled for these samples, the relationship between SNP heterozygosity and CNVs has yet to be analyzed and therefore remains unknown.

1.8 Chromosomal landscape of genetic variation can be visualized using rainfall and rainbow plots to gauge the spatial relationship of genomic events

Rainfall plots are scatterplots used to visualize the distribution of genomic events [79]. First used to detect somatic point mutation hotspots within cancer genomes [80], rainfall plots can be adapted instead to visualize the distribution of SNP heterozygosity along a chromosome and

anticipate regions susceptible to mutagenesis leading to CNV formation. The x-axis shows the position of heterozygous SNP loci while the y-axis is the distance to the preceding heterozygous SNP locus in base pairs (Fig 1.5). Rainfall plots are useful in visualizing the nature, number, and distribution of heterozygous SNP loci clusters along a chromosome.

The spatial relationship between two genomic events, such as heterozygous SNP loci and CNVs, can be visually assessed using rainbow plots [Luo, unpublished, in preparation]. Rainbow plots portray the density and distribution of heterozygous SNP loci with respect to their proximity to CNV events along a chromosome (Fig 1.6). Rainbow plots are useful for discerning the heterozygous landscape around each CNV along a chromosome. CNVs in heterozygous SNP dense regions will show many more proximal heterozygous SNPs falling lower on the graph than CNVs in heterozygous SNP deserts.

In combination, rainfall and rainbow plots are powerful visualization tools that can provide compelling evidence for association between heterozygous SNP loci and CNVs. However, this analysis is subjective and may lead to errors in interpretation. For example, A rainbow plot could be misleading in showing a CNV appearing in close proximity to many heterozygous SNPs while not accounting for a far greater number of homozygous SNP probes in the region. It is necessary to have an objective, quantifiable measure by which to understand the spatial association between heterozygous SNPs and CNVs.

1.9 Novel spatial statistical tools can be used effectively to interrogate the relationship between heterozygosity and copy number variants

Recently, novel statistical tools have been proposed for characterizing the spatial relationship between heterozygous SNP loci and CNVs using microarray data [81, 82]. One such tool is a nonparametric test for spatial independence between genomic events is called the J statistic, and may be used to determine whether heterozygous SNP loci are nearby CNVs on an autosome. Adapted from the J function, a ratio statistic used for profiling the distribution of geological

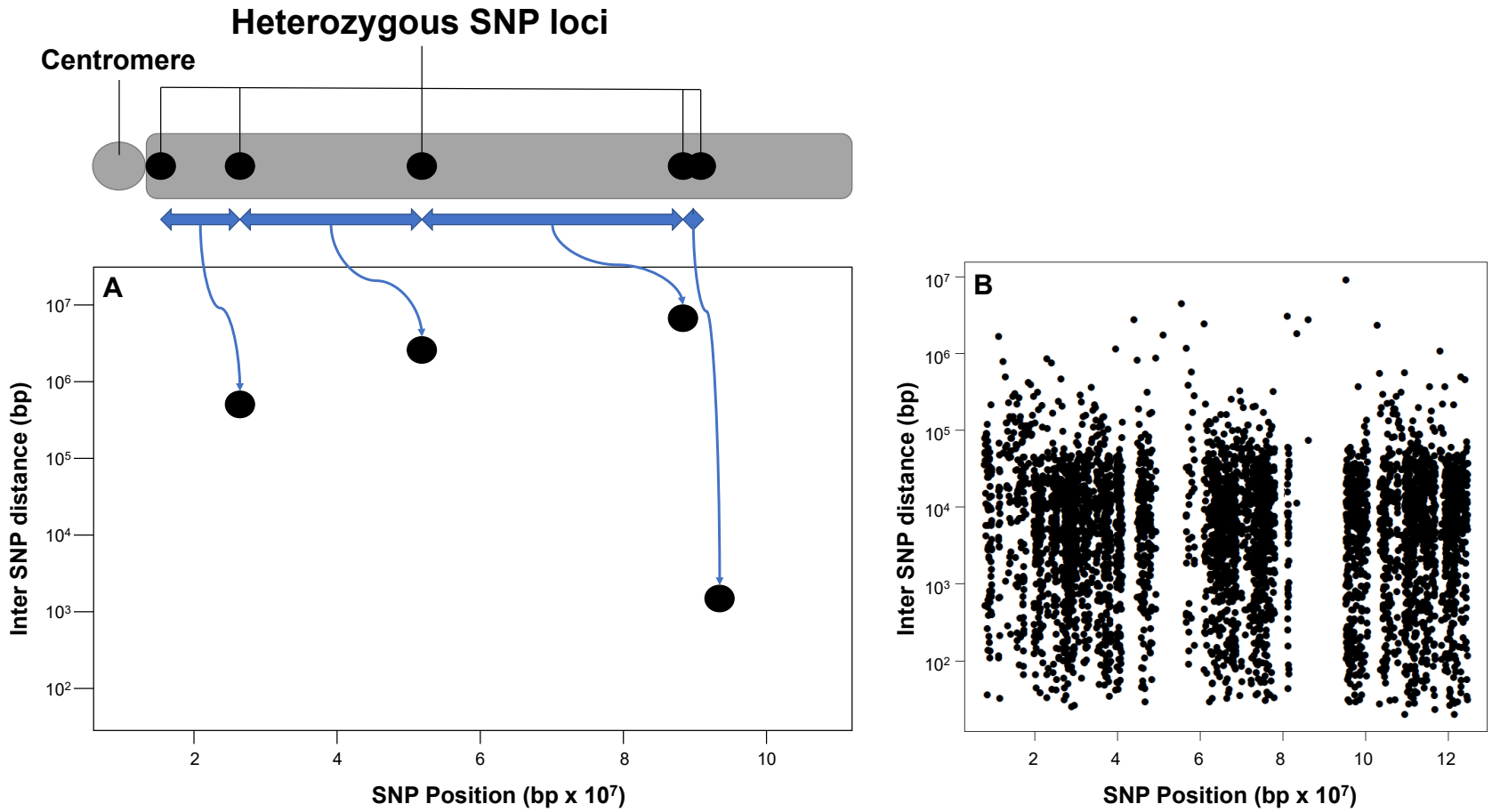

Figure 1.5: Rainfall plots enable visualization of the distribution of heterozygous single-nucleotide polymorphic loci along a chromosome. A) Black circles represent heterozygous SNP loci. The x-axis of a rainfall plot is the chromosomal position in base pairs. The y-axis represents the distance between a heterozygous SNP locus and the previous heterozygous SNP locus plotted on a logarithmic scale. **B)** Example chromosome 14 from a F1 mouse with approximately 29% SNP heterozygosity. Rainfall plots are useful for visualizing regions of heterozygous SNP loci clustering along a chromosome. For example, several instances where the distance between heterozygous SNP loci is less than 1000 bp are apparent.

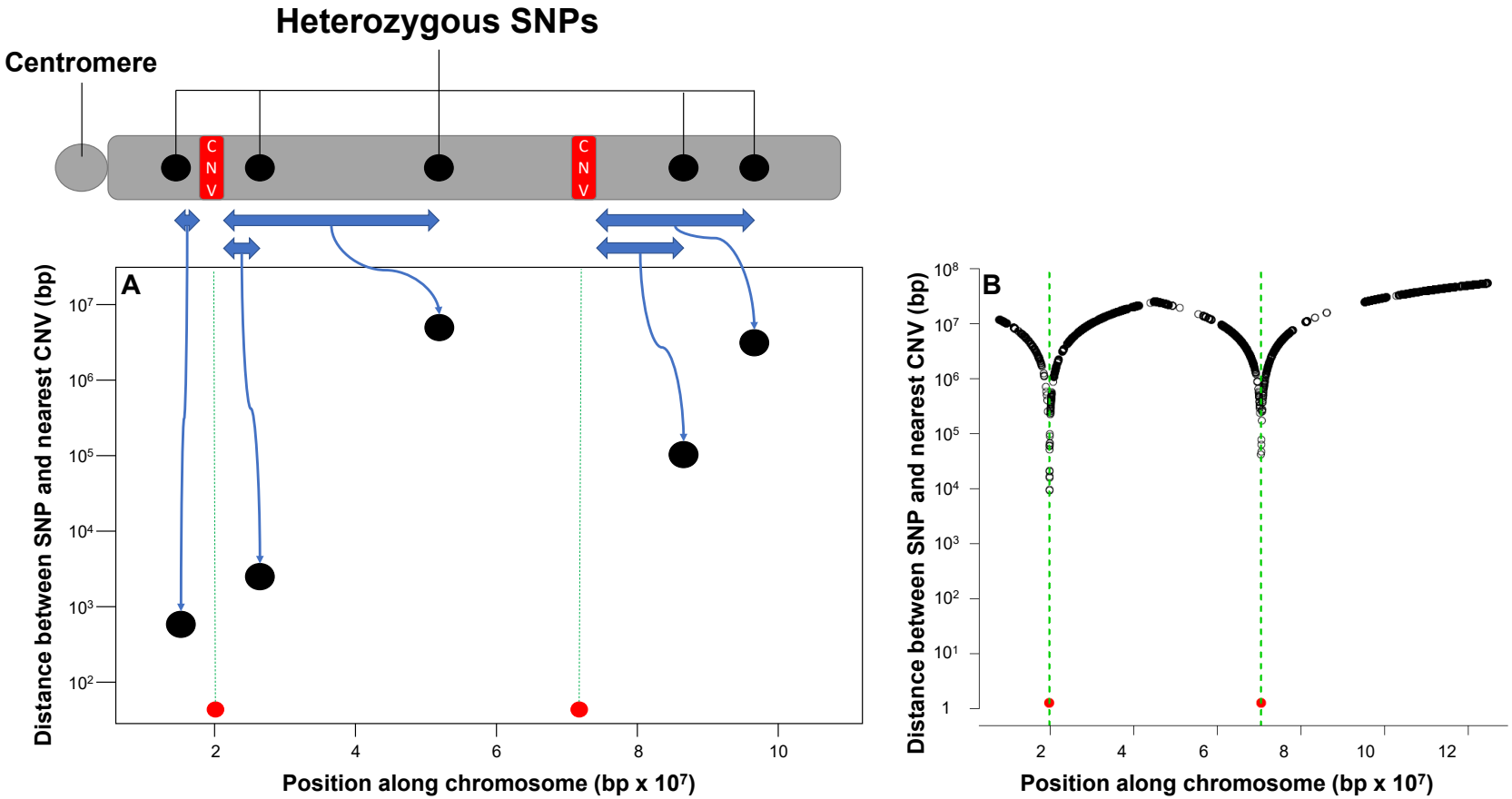

Figure 1.6: Rainbow plots enable visualization of the spatial relationship between heterozygous single-nucleotide polymorphic loci and copy number variants on a chromosome. A) Black circles represent the position of heterozygous SNPs in base pairs while red circles represent the position of a CNV in base pairs. The x-axis represents the position along the chromosome in base pairs. The y-axis is the distance of a heterozygous SNP to its nearest CNV in base pairs, causing a 'rainbow' shape to form. **B)** Example chromosome 14 from a F1 mouse with approximately 29% SNP heterozygosity and two CNV events proximally associated with heterozygous SNP loci.

fault lines to ore [83], the J statistic accommodates the problem of intermittent probe sampling across the genome by microarrays [84].

The J statistic evaluates the neighbourhood of queried SNP probes outside of CNV regions on a chromosome and determines whether the distribution of heterozyous SNP loci is following a random distribution. Using a Poisson Null process and Monte Carlo simulations, global confidence bands are constructed to test the null hypothesis that heterozygous SNPs are randomly distributed outside of CNV regions. The length of the chromosomal region interrogated surrounding CNVs can be varied and is referred to as the 'r value distance'. More specifically, the r value is the number of base pairs from start and end of a CNV to be evaluated by the J statistic. If the observed J function extends outside the global confidence bands, the null hypothesis is rejected.

If the observed J function remains within the global confidence bands, the heterozygous SNP landscape is not significantly different from a random distribution (Fig 1.7). If the observed J function crosses below the bottom global confidence band, heterozygous SNPs are significantly proximal to (nearby) CNVs for the interrogated chromosome (Fig 1.8). If the observed J function crosses above the top global confidence band, heterozygous SNPs are significantly distal to (far away from) CNVs for the interrogated chromosome (Fig 1.9).

1.10 The Axiom MouseHD array is an untapped, cost-effective candidate technology to survey the mouse genome

Although the MDGA remains the most probe dense, a newer, more cost efficient array has recently been designed for mice. The Axiom MouseHD genotyping array is a custom-made high-density array that costs approximately $75 per sample as opposed to the $600 per sample of the MDGA. The Axiom array targets 616,136 SNP loci, 488,945 of which are also targeted by the MDGA. The Axiom array and MDGA are comparable in the number of SNP loci targeted, both with an approximate resolution of 1 SNP locus per 4.4 kb. However, the Axiom array does not contain any IGPs and therefore does not interrogate conserved loci.

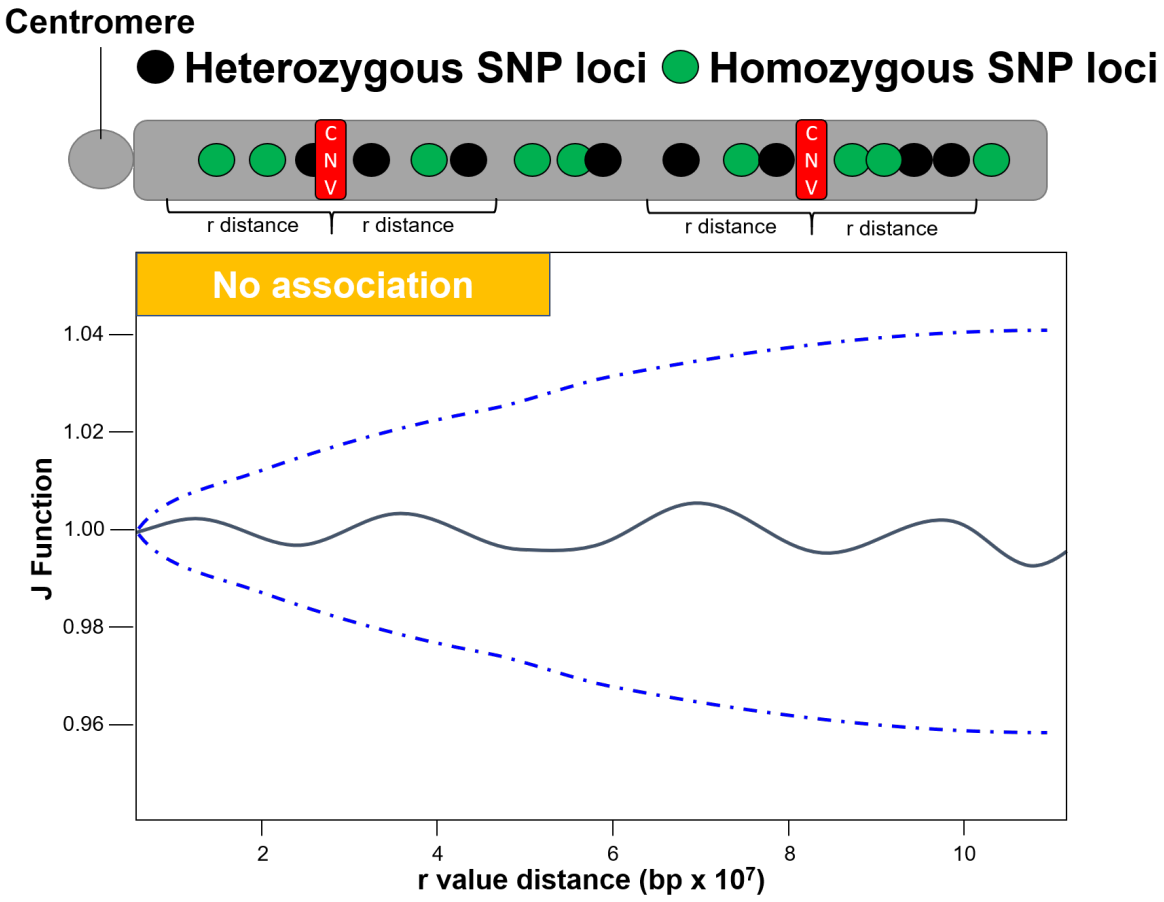

Figure 1.7: Example of a J statistic result for heterozygous single-nucleotide polymorphic loci and copy number variants that are not spatially associated to one another on a chromosome. An example mouse chromosome with sampled SNP loci is depicted above the plot. Black points represent heterozygous SNP loci while green points represent homozygous SNP loci. CNVs are represented by red vertical bars. The r distance is the length of the region surrounding CNVs evaluated by the J statistic extending outward from the start and end of each CNV. The heterozygous and homozygous SNP loci are randomly distributed in the evaluated sections of the chromosome. The x-axis of the J statistic plot is the r value distance while the y-axis is the observed J function. The dotted blue lines are global confidence bands constructed from the Poisson null process with a critical significance level of $\alpha = 0.05$. The solid black line represents the J function, a nonparametric measure of association between the observed heterozygous SNP loci distribution and random distributions generated through Monte Carlo simulations. When the J function remains within the global confidence bands, the test statistic fails to reject the null hypothesis and no evidence is found for a spatial association between heterozygous SNP loci and CNVs.

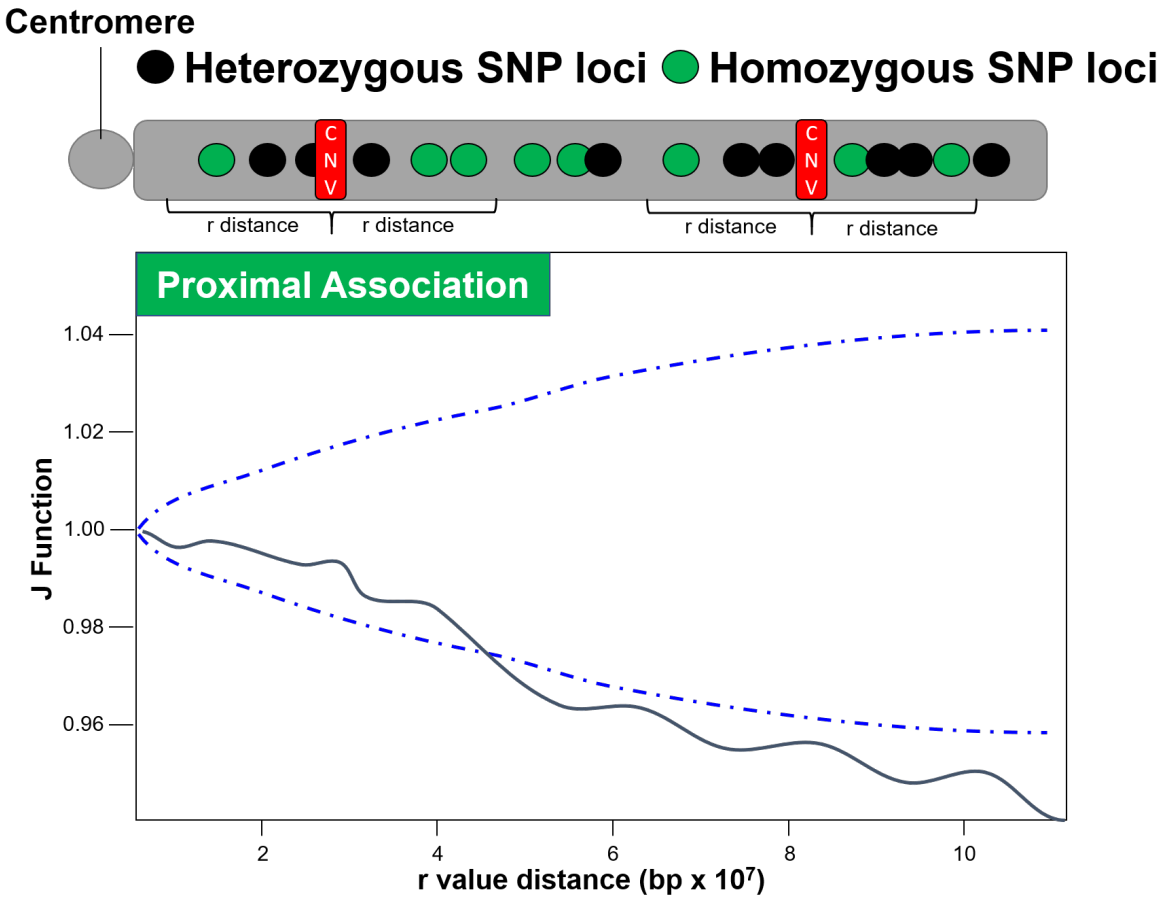

Figure 1.8: Example of a J statistic result for heterozygous single-nucleotide polymorphic loci and copy number variants that are proximally associated to one another on a chromosome. An example mouse chromosome with sampled SNP loci is depicted above the plot. Black points represent heterozygous SNP loci while green points represent homozygous SNP loci. CNVs are represented by red vertical bars. The r distance is the length of the region surrounding CNVs evaluated by the J statistic extending outward from the start and end of each CNV. The heterozygous SNP loci are distributed more closely to CNV events in the evaluated sections of the chromosome than homozygous SNP loci. The x-axis of the J statistic plot is the r value distance while the y-axis is the observed J function. The dotted blue lines are global confidence bands constructed from the Poisson null process with a critical significance level of $\alpha = 0.05$. The solid black line represents the J function, a nonparametric measure of association between the observed heterozygous SNP loci distribution and random distributions generated through Monte Carlo simulations. When the J function crosses the bottom global confidence band, the test statistic rejects the null hypothesis and indicates a proximal spatial association between heterozygous SNP loci and CNVs.

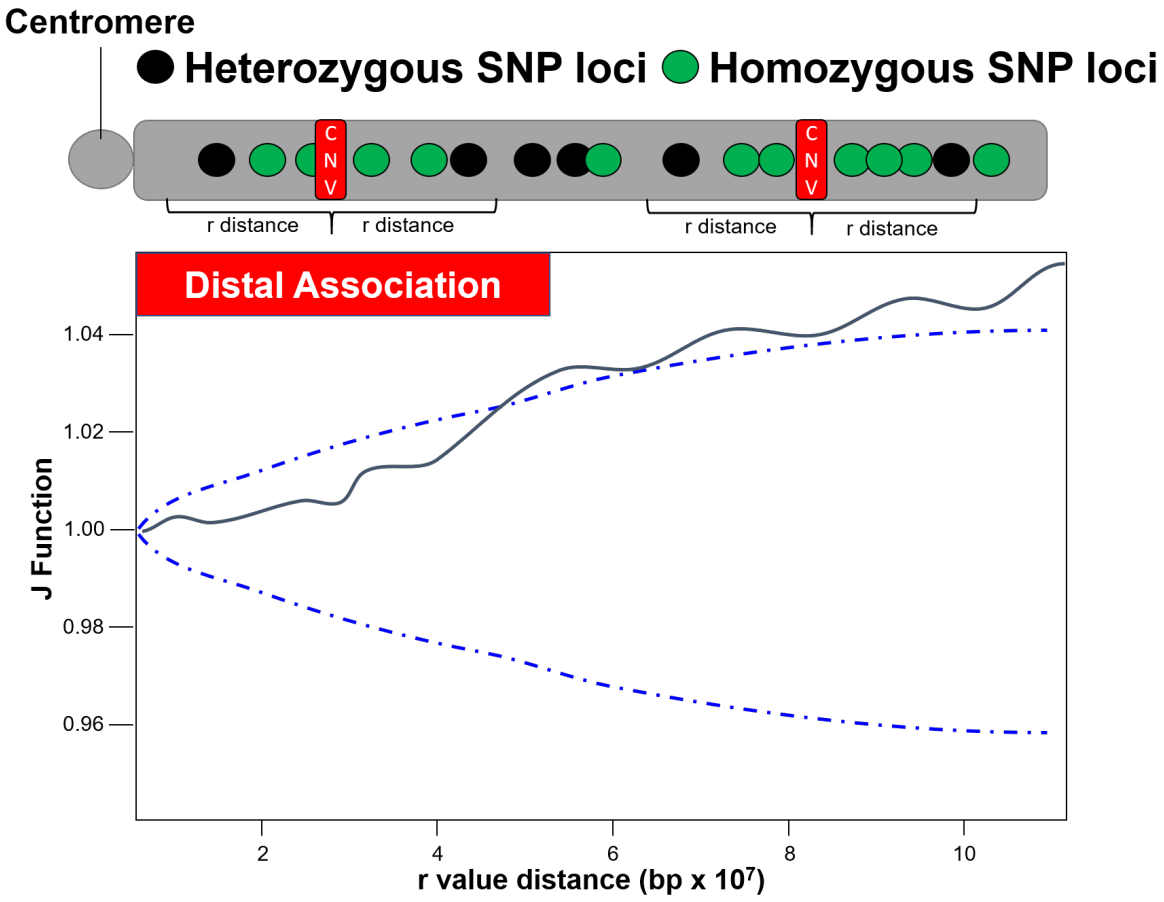

Figure 1.9: Example of a J statistic result for heterozygous single-nucleotide polymorphic loci and copy number variants that are distally associated to one another on a chromosome. An example mouse chromosome with sampled SNP loci is depicted above the plot. Black points represent heterozygous SNP loci while green points represent homozygous SNP loci. CNVs are represented by red vertical bars. The r distance is the length of the region surrounding CNVs evaluated by the J statistic extending outward from the start and end of each CNV. The heterozygous SNP loci are distributed more closely to CNV events in the evaluated sections of the chromosome than homozygous SNP loci. The x-axis of the J statistic plot is the r value distance while the y-axis is the observed J function. The dotted blue lines are global confidence bands constructed from the Poisson null process with a critical significance level of $\alpha = 0.05$. The solid black line represents the J function, a nonparametric measure of association between the observed heterozygous SNP loci distribution and random distributions generated through Monte Carlo simulations. When the J function crosses the top global confidence band, the test statistic rejects the null hypothesis and indicates a distal spatial association between heterozygous SNP loci and CNVs.

The Axiom MouseHD array has been successfully used to identify quantitative trait loci implicated in contributing to cardiac hypertrophy in Balb/Cj mice [85]. However, the Axiom Mouse HD array has never been used to detect CNVs in mice and was not designed with CNV calling in mind. Therefore, new, streamlined software packages such as Axiom Analysis Suite and Axiom CNV Tool do not come with support for CNV calling for the Axiom MouseHD array. Fortunately, PennCNV, a free software tool provided by Wang *et al*. for detecting CNVs, uses a Hidden Markov Model (HMM) CNV calling algorithm that can be used for any high-density SNP genotyping array [86].

The Axiom MouseHD array is a potentially useful and practical technology for studying the relationship between heterozygous SNPs and *de novo* CNVs because its low cost permits an increase sample size and therefore increased statistical power. Given the expected low rate of *de novo* CNV mutagenesis, it is desirable to investigate the spatial landscape of as many chromosomes as possible.

1.11 Central hypothesis

The startling findings of high heterozygosity locally affecting the rate of SNP and indel mutations per generation in both *Arabidopsis* and peach plants provided a strong foundation for extending investigation of this phenomenon to nonplant organisms. Mice are a staple mammalian model organism well-suited for replicating an experimental setup with high and low heterozygosity to assess the relative level of mutation. Given the nature of mechanisms of mutation resulting in CNVs, I hypothesized that clusters of SNP heterozygosity in mice would promote and mediate the incidence of NAHR events following DSB repair, resulting in the formation of new CNVs. I therefore predicted that clusters of SNP heterozygosity in mice would co-localize with CNVs more frequently than expected at random and mice of higher heterozygosity would harbour a greater number of CNV events. Further, if heterozygosity is mutagenic during meiotic recombination, I expect *de novo* CNVs to arise in proximity to clusters of SNP heterozygosity in the parental mice of the affected individual.

1.12 Experimental aims

My first experimental aim was to determine the spatial relationship between heterozygous SNP loci and CNVs from publicly available MDGA data. Careful scrutiny was given to classical inbred mice, wild-derived inbred mice, and classical inbred F1 hybrids, which had low, moderate, and high SNP heterozygosities, respectively. These groups are of particular interest because the classical inbred mice and their F1 hybrids had different levels and distributions of heterozygous SNP loci, yet both arose from the same low heterozygosity gametogenesis environments. In contrast, the wild-derived mice had moderate heterozygous SNP loci levels both pre- and post-zygotically.

The first step in this aim was to download, filter and categorize the publicly available MDGA data from the Jackson Laboratory. Genotyping and CNV detection were then performed, determining the location and distribution of both heterozygous SNP loci and CNVs. Rainfall, rainbow and J statistic plots were then produced for all autosomes. If mitotic mechanisms are responsible for contributing to CNV mutagenesis, I expected to see an elevation in both CNV frequency and proximity to heterozygous SNP loci in the F1 hybrids. If meiotic mechanisms are responsible, I expected the wild-derived mice to demonstrate this pattern.

My second experimental aim was to generate a novel automated pipeline by which the spatial relationship between any heterozygous SNP loci and CNVs may be investigated. This aim was tackled in tandem with aim one. To address this aim, various R coding scripts were made to automatically process raw exported data from Affymetrix genotyping software and PennCNV. Using these scripts, the spatial relationship of heterozygous SNP loci and CNVs was determined for thousands of chromosomes at a time. Additionally, this automated pipeline may be adapted to investigate the spatial relationship between other genomic features.

My final experimental aim was to uncover the spatial relationship between heterozygous SNP loci and *de novo* CNVs in a meiotic context. To address this aim, a breeding experiment using B6 and DBA classical inbred parental mice was conducted. Using publicly available data from the MDGA, I determined the expected level of heterozygous SNP loci of B6 x DBA F1 hybrids to be approximately 20% and distributed in clusters throughout the genome. These were ideal

mice for testing whether heterozygosity affects *de novo* CNV mutagenesis in a mitotic setting because this elevated heteterozygosity is post-zygotic.

The F1 mice were then brother-sister mated to produce F2 mice of unpredictable and discontinuous heterozygous SNP loci distributions. The average level of heterozygous SNP loci of these mice was approximately 10%, although inter-animal and inter-chromosomal variation were expected and observed. These F2 mice arose from cells undergoing gametogenesis with high heterozygosity and were an ideal vector to test whether heterozygosity affects *de novo* CNV mutagenesis in mice.

For six three-generation mouse lines, the landscape of heterozygous SNP loci and CNVs for the parentals, F1s, and F2s using the novel spatial statistical pipeline was elucidated. CNVs that were not inherited in the F1s and F2s were identified. If heterozygous SNP loci affect mutagenesis leading to *de novo* CNV formation in a meiotic setting, I expected to see *de novo* CNVs located close to heterozygous SNP loci in the F2 cohort. More accurately, I expected these *de novo* CNVs to be located close to the heterozygous SNP loci of the F1 cells undergoing gametogenesis they arose in.

Materials and methods

2.1 Genotyping and copy number variant calling for 800 publicly available mouse diversity genotyping array samples

Of 1901 MDGA raw data (.CEL) files publicly available from the Jackson Laboratory [78], 800 were selected based on their known diversity in SNP heterozygosity due to a spectrum of breeding strategies and genealogical histories. Samples were included if they were derived from one of the following categories: Classical inbred (CI), recombinant inbred (RI), wild-derived (WD), Caesarian derived-1 (CD1) outbred, Naval Medical Research Institute (NMRI) outbred, or F1 hybrids. A comprehensive list of sample information including sex, strain and category is provided in the supplementary online materials (Appendix B).

Genotyping was performed using the BRLMM-P algorithm implemented by Affymetrix Power Tools (APT; Thermo Fisher Scientific) software [87] with default parameters. A call rate cut-off of >97% was employed. Of the 800 samples genotyped, 66 samples had call rates below 97% and were discarded from further analysis. The total fluorescence intensity signals from each SNP probeset were summarized as log R ratio (LRR) values. The fluorescence signal ratio between the B and A allele probes at each SNP locus was represented by B allele frequency (BAF) values. Both LRR and BAF values were calculated by generating a canonical genotype clustering file using the PennAffy package [88]. An in-house population frequency of the B allele reference file based on 351 representative mice [75] was used by PennCNV software [86] to detect CNVs. The default Affymetrix affygw6.hmm file was used to supply the hidden Markov model (HMM) to PennCNV. Autosomal CNVs were detected using default

parameters.

Using suggested quality control parameters for Affymetrix arrays [89] and in accordance with previously used MDGA CNV calling methodology [75], 22 samples with LRR standard deviation values greater than 0.35 or BAF drift values greater than 0.01 were removed from further analysis. In addition, 2 files were removed for not calling any CNVs, leaving a total of 710 quality control (QC) passing samples.

2.2 Profiling the density and distribution of heterozygous single-nucleotide polymorphic loci and copy number variants for 707 quality control passing mouse diversity genotyping array samples

Following APT genotyping and PennCNV calling, files were formatted with an in-house R script for input into a novel spatial statistic pipeline [82, 84]. Briefly, this script sorted the SNP and CNV call data for each sample and merged the data from these separate files such that a singular comma-separated values (CSV) output file for each sample contained all genotype calls and detected CNVs, sorted by chromosome number and position. The spatial statistic pipeline R script was developed in-house by Bin Luo, Steven Villani, and myself. This pipeline automatically processed all properly formatted samples within a folder and, using the J statistic, determined whether two genomic features are nonrandomly spatially associated with one another.

Spatial associations between heterozygous SNP loci and CNVs were profiled for 13,490 autosomes from the 710 quality control passing samples. Window size, also referred to as the 'r value' for the range of the CNV neighbourhood, was set to 10,000,000 bp outside of CNV events and 1,000 simulations were performed per autosome. Three samples were unable to be processed by this script given that the CNVs were called by IGPs outside of the range by which SNP loci are interrogated along a chromosome. These three samples were omitted from further analysis. Four autosomes across two classically inbred samples were discarded from

further analysis for having fewer than 10 heterozygous SNP calls, an arbitrarily selected low heterozygosity cutoff to prevent erroneous J statistic false positives.

For the remaining 13,429 autosomes from 707 mouse samples, heterozygous inter-SNP locus distances were visualized as rainfall plots. Rainbow and J statistic plots were generated for 5,799 autosomes that harboured at least one CNV (Appendix B).

2.3 Engineering discontinuous landscapes of single-nucleotide polymorphic heterozygosity by breeding two genetically distinct inbred mouse lines

Six mouse families of three generations were bred using C57BL/6J and DBA/2J parental strains at the Jackson Laboratory. Three families were derived from a male C57BL/6J and female DBA/2J cross while the other three were derived from a female C57BL/6J and male DBA/2J cross (Fig 2.1). F1s were brother-sister mated to produce at least 2 male F2 mice for each of the six lines. Parental and F1 mice were euthanized and frozen on dry ice following successful breeding. F2 mice were euthanized and exsanguinated at four weeks of age. Blood, cerebrum, cerebellum, lungs, heart, liver, testes, spleen, kidney, bladder, pancreas, tail clips, and ear clips were extracted and frozen on dry ice. All tissues and carcasses were shipped from the Jackson Laboratory on dry ice.

DNA was extracted from ear and tail clips for all parental and F1 mice and from liver, lung, pancreas and tail clips for all F2 mice using the Thermo Fisher Purelink™ Genomic Extraction Kit (Thermo Fisher Scientific Inc, Waltham, MA). DNA integrity was verified using electrophoresis on a 1.0% agarose gel in Tris-borate-EDTA buffer. DNA concentration was measured using the Thermo Fisher Nanodrop™ 2000 (Thermo Fisher Scientific Inc, Waltham, MA) and the DNA diluted to 50 ng/uL. The Centre for Applied Genomics (TCAG) received 10 uL of sample DNA in a 96 well plate for Axiom MouseHD microarray (Thermo Fisher Scientific Inc, Waltham, MA) analysis.

SNP genotyping was performed according to the standard protocol outlined in the Axiom 2.0

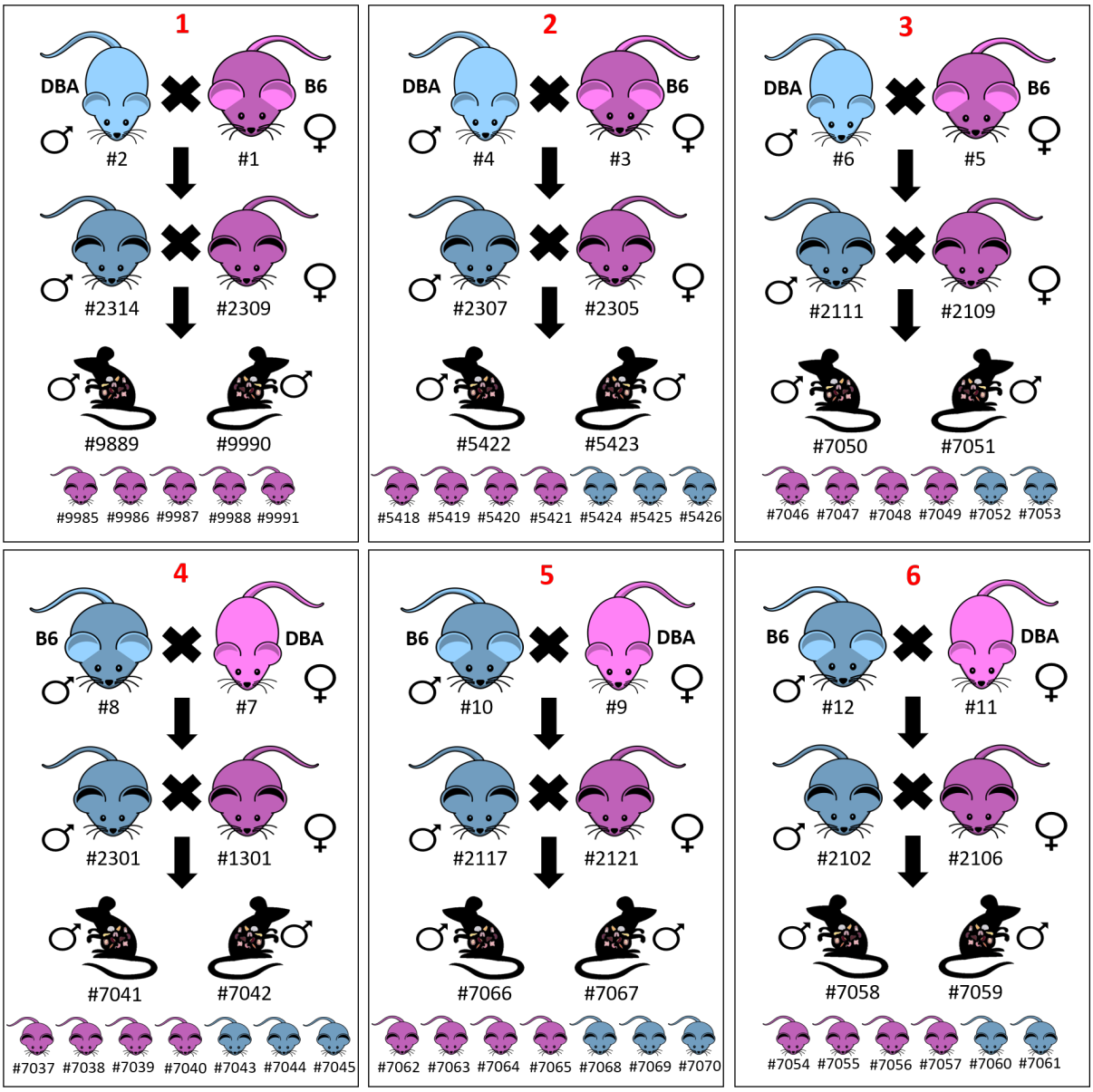

Figure 2.1: Mouse breeding and sample identification schematic for engineering single-nucleotide polymorphic heterozygosity in six mouse families. Classical inbred parental mouse strains C57BL/6J (B6) and DBA/2J (DBA) were bred to produce isogenic F1 hybrid mice with approximately 20% SNP loci heterozygosity. Brother-sister inbreeding of F1 hybrids generated F2 mice of variable SNP loci heterozygosity. Two F2 male mice were selected for tissue harvesting. All other mice, including additional F2 offspring were shipped from the Jackson Laboratory as whole carcasses. Blue coloured mice are male and pink coloured mice are female. Families 1-3 began with a male B6 mouse bred with a female DBA mouse whereas families 4-6 began with a female B6 mouse bred with a male DBA mouse. Numbers indicate unique mouse identifiers.

Assay Manual User Guide (Thermo Fisher Scientific Inc; Life Technologies, Carlsbad, CA). Briefly, 200 ng of genomic DNA was denatured and then amplified. Amplified DNA was fragmented, precipitated and then centrifuged. The pellets were dried and resuspended in buffer and added to a hybridization master mix prior to hybridization to the Axiom MouseHD microarray plate in the GeneTitan Multi-Channel Instrument for 23.5 h. Finally, the array plate was stained, washed, and scanned in the GeneTitan Multi-Channel Instrument (Thermo Fisher Scientific Inc, Waltham, MA).

2.4 Predicting heterozygous single-nucleotide polymorphism landscapes of DBA/2J x C57BL/6J F1 hybrid mice

DBA/2J and C57BL/6J inbred mice are two of the most commonly used classical inbred mouse strains and are frequently used to produce F1 hybrids. Using genotyping calls for 11 C57BL/6J and 4 DBA/2J MDGA files from the publicly available data, predictive heterozygous SNP rainfall plot landscapes for hypothetical F1 hybrids were generated for all 19 autosomes (Appendix A).

Predicted heterozygous SNP loci rainfall plots were produced using a conservative approach to determine the expected minimal heterozygosity of all F1 hybrids. SNP genotypes at 473,547 autosomal loci interrogated by the MDGA were investigated for consistent homozygous calls (AA or BB) for all 11 C57BL/6J mice and repeated separately for all 4 DBA/2J mice. At 84,799 loci (or approximately 17.9% of the total loci), the C57BL/6J mice and DBA/2J were homozygous for opposite alleles. Rare mutations notwithstanding, all F1 hybrids of these inbred mice are expected to be heterozygous at all of these loci. All 19 predicted heterozygous SNP loci rainfall plots were visually compared to rainfall plots produced by a C57BL/6J x DBA/2J F1 hybrid (18.8% SNP heterozygosity).

2.5 Genotyping and copy number variant calling of 96 Axiom MouseHD array samples

A sample set of 88 Axiom MouseHD array .CEL files was downloaded to be used as a training set for the BRLMM-P and PennCNV calling algorithms (Appendix B). All 88 mice are F2 mice derived from C57BL/6J x Balb/CJ F1 backcrosses [85]. These 88 mice will henceforth be referred to as the training set.

SNP genotyping at 616,136 loci from the 96 Axiom MouseHD array fluorescence intensity files (.CEL) was performed using the BRLMM-P algorithm implemented by Axiom Analysis Suite software (Thermo Fisher Scientific Inc; Affymetrix Inc, Santa Clara, CA) using the best practices workflow and including the training set. LRR and BAF values were calculated using the Axiom CNV Summary Tools software and each file was converted and exported into PennCNV format.

A population frequency of the B allele (PFB) file was generated using the training set as previously described [75]. The default Affymetrix affygw6.hmm files was used to supply the Hidden Markov Model to PennCNV. Autosomal CNVs were detected using default parameters.

Using suggested quality control parameters for Affymetrix arrays, all 96 files were retained for downstream analysis for having LRR standard deviation values lower than 0.35 and BAF drift values lower than 0.01. A total of 4,444 autosomal CNVs were detected across 96 samples. To reduce false positive calling and in line with previous work [75], CNVs were excluded from further analysis if they had a probe density less than 1/7000 bp. Further, state 0 (full deletion) CNVs were stringently excluded if they were called with 15 or fewer probes. After filtering, 3,479 CNVs remained for downstream analysis.

2.6 Profiling the density and distribution of heterozygous single-nucleotide polymorphic loci and copy number variants for 96 quality control passing mouse family samples

The SNP call file and filtered PennCNV output file generated by Axiom Analysis Suite software were processed by the in-house R script for input into the spatial statistic pipeline. Spatial association between heterozygous SNP loci and CNVs was profiled for 1,824 autosomes from the 96 QC passing Axiom samples. Window size (r value) was set to 10,000,000 bp outside of CNV events and 1,000 simulations were performed per autosome. Heterozygous SNP loci distances were visualized as rainfall plots. Rainbow and J statistic plots were generated for 1,338 autosomes that harboured at least one CNV (Appendix B).

2.7 Determination of recurrent copy number variants within mouse families using HD-CNV software

Recurrent CNVs within each of the six mouse families were identified using HD-CNV software. Consistent with previous methodology [75], a recurrent CNV was called if there was 40% reciprocal overlap with at least one other CNV detected within the mouse family. Gephi software [90] graph files produced for each chromosome were formatted using the Fruchterman-Reingold layout and images were scaled to represent total number of CNV calls per chromosome. A CNV was considered a singleton if it shared less than 40% reciprocal overlap with all other CNVs within the family.

2.8 Identifying inherited and *de novo* copy number variants in F1 and F2 mice

Inherited and *de novo* CNVs were identified in the F1 and F2 mice. A CNV was considered the same across samples if it shared a 40% reciprocal overlap. For a CNV to be classified as

inherited in the F1 mice, it had to be called in both the tail and ear sample for the same mouse, as well as in at least one of the four parental tissues. To be considered *de novo*, the CNV had to be called in both the tail and ear sample for the same mouse, but not found in any of the four parental tissues.

For the F2 mice, a CNV was classified as inherited if it was called in at least two of the four sampled F2 tissues (tail, lung, liver, pancreas) and was also called in at least one of the four parental tissues or the four F1 tissues. A CNV was considered *de novo* if it was called in at least two of the four sampled F2 tissues and was not found in any parental or F1 tissue.

Presumptive *de novo* CNVs were stringently filtered to have been called by at least 15 SNP probes. All *de novo* CNV BAF and log2 ratios were manually inspected using Axiom CNV Viewer software and 3 CNVs were found to be adjacent and improperly called as separate CNVs. These CNVs were merged and included in downstream analysis.

2.9 Spatial analysis of heterozygous single-nucleotide polymorphic loci and *de novo* copy number variants in F1 and F2 mice

De novo CNVs were extracted from the PennCNV output files and reformatted for the spatial analysis pipeline. The spatial relationship between heterozygous SNPs and *de novo* CNVs were profiled using rainfall, rainbow and J statistic plots. All other CNVs were omitted from this analysis.

With a meiotic mechanism of CNV mutagenesis in mind, an additional round of spatial analysis was conducted between F2 *de novo* CNVs and the heterozygous SNP landscapes of each of the F1 parents they were birthed from. Rainfall, rainbow, and J statistic plots were generated for chromosome harbouring an F2 *de novo* CNV for both maternal and paternal heterozygous SNP landscapes.

Results

3.1 Single-nucleotide polymorphic heterozygosity is not correlated with an elevated number of copy number variants per autosome for 707 mouse diversity genotyping array samples

Evaluation of genomic SNP heterozygosity across all autosomes for six distinct cohorts of mice (classical inbred, recombinant inbred, wild-derived inbred, NMRI outbred, CD-1 outbred and F1 hybrids) confirmed a wide SNP heterozygosity range from 0.09% to 45.75% (Table 3.1). On average, classical inbred mice had low SNP heterozygosity (0.31%), wild-derived mice had moderate SNP heterozygosity (3.96%), and F1 hybrids had high SNP heterozygosity (36.69%).

A total of 11,270 CNVs were detected by PennCNV from 707 publicly available MDGA samples. Autosomal SNP heterozygosity levels were similar to genomic SNP heterozygosity. Classical inbred, wild-derived, and F1 mice had 0.31%, 3.72%, and 37.47% average autosomal SNP heterozygosity, respectively (Table 3.2). The average number of CNVs per autosome was higher for the wild-derived mice (3.13 per autosome) than the classical inbred and F1 mice (1.22 and 1.15 per autosome, respectively).

There was no correlation observed between autosomal SNP heterozygosity and number of CNVs for classical inbred ($r = -0.04$, p-value > 0.001), recombinant inbred ($r = 0.01$, p-value > 0.001), wild-derived ($r = 0.00$, p-value > 0.001), CD-1 outbred ($r = 0.07$, p-value > 0.001), or F1 ($r = 0.08$, p-value > 0.001) mice (Fig. 3.1). A very weak correlation was observed between

Table 3.1 Determined whole genome SNP heterozygosity values for 707 publicly available Mouse Diversity Genotyping Array samples

Mouse cohort	Number of samples	Genomic SNP Heterozygosity (%)		
		Average[1]	Minimum	Maximum
Classical Inbred	126	0.31	0.09	1.51
Recombinant Inbred	108	0.40	0.12	3.96
Wild-Derived	40	3.72	0.51	15.52
NMRI Outbred	279	6.30	5.23	7.34
CD-1 Outbred	99	10.68	10.03	11.51
F1	55	36.69	17.99	45.75

[1] Ordered by genome average SNP heterozygosity

Table 3.2 Determined autosomal SNP heterozygosity and CNVs detected for 707 Mouse Diversity Genotyping Array samples

Mouse cohort	Number of autosomes	Average number of CNVs per autosome	Autosomal SNP Heterozygosity (%)		
			Average[1]	Minimum	Maximum
Classical	2394	1.22	0.26	0.03	4.76
Recombinant Inbred	2052	0.66	0.37	0.05	11.20
Wild Derived	760	3.13	3.55	0.14	22.54
NMRI	5301	0.33	6.49	0.40	16.71
CD1	1881	0.88	11.04	2.66	19.51
F1	1045	1.15	37.47	9.59	50.23

[1] Ordered by autosomal average SNP heterozygosity

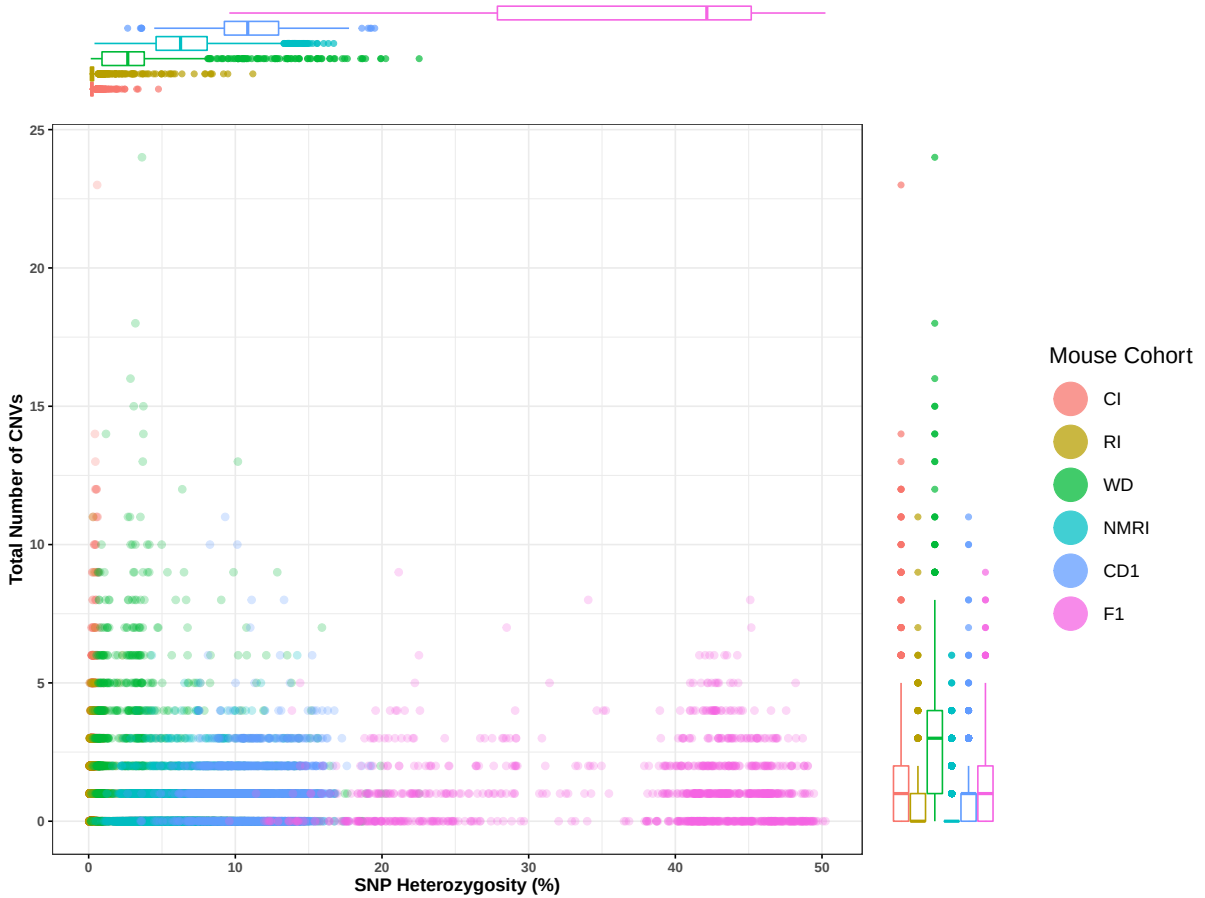

Figure 3.1: SNP heterozygosity is not correlated with a higher autosomal CNV burden in 707 MDGA samples. The x-axis represents the SNP heterozygosity (%) while the y-axis represents the number of CNVs detected per autosome. All 19 autosomes from each individual are plotted. Marginal boxplots indicate the distribution of data for each mouse cohort ordered from lowest to highest average SNP heterozygosity.

autosomal SNP heterozygosity and number of CNVs for NMRI outbred (r = 0.18, p-value < 0.001) mice.

There was no correlation observed between autosomal SNP heterozygosity and number of CNVs for classical inbred (r = -0.04, p-value > 0.001), recombinant inbred (r = 0.01, p-value > 0.001), wild-derived (r = 0.00, p-value > 0.001), CD-1 outbred (r = 0.07, p-value > 0.001), or F1 (r = 0.08, p-value > 0.001) mice (Fig. 3.1). A very weak correlation was observed between autosomal SNP heterozygosity and number of CNVs for NMRI outbred (r = 0.18, p-value < 0.001) mice.

3.2 Heterozygous single-nucleotide polymorphisms and copy number variants are frequently nonrandom in their spatial association for 707 MDGA samples

Of the 13,433 autosomes assayed, 7,630 did not harbour any CNV events, 4 had fewer than 10 heterozygous SNPs, and 5,799 had at least one detected CNV event. The observed J statistic shows a significant ($\alpha < 0.05$) spatial association between heterozygous SNPs and CNVs for 60.9% of autosomes assayed. Of the spatial associations between heterozygous SNPs and CNVs, 1,393 were proximal, 1,198 were distal, and 942 were both proximal and distal. No spatial association between heterozygous SNPs and CNVs were found for the remaining 2,266 autosomes harbouring at least one CNV.

3.3 Heterozygous single-nucleotide polymorphic loci and copy number variants are more frequently proximal than distal to one another in most assayed MDGA mouse genomes

Classical inbred, recombinant inbred, wild-derived and CD-1 mice had a higher proportion of autosomes with proximally associated heterozygous SNPs and CNVs than autosomes with distal associations (Fig. 3.2). In contrast, NMRI and F1 mice had fewer autosomes with

proximally associated heterozygous SNPs and CNVs than autosomes with distal associations.

While autosomes in the classical inbred group showed no spatial association between heterozygous SNPs and CNVs, 17.3% were found to be proximal compared to only 0.5% distal. A similar, although less pronounced, trend was observed in the wild-derived cohort, with 49.9% of autosomes having no association between heterozygous SNPs and CNVs, 27.1% were proximal and 19.4% were distal. The opposite is observed with the high heterozygosity F1 mice. These mice had only 25.9% of autosomes with no association between heterozygous SNPs and CNVs, 25.3% with proximal associations, and 39.5% distal associations.

All but one of the 2,377 autosomes assayed from the CD-1 and NMRI outbred cohorts had a significant spatial association between heterozygous SNPs and CNVs. CD-1 and NRMI outbred mice also had a high proportion (33.2% and 38.3% of autosomes, respectively) of both proximal and distal associations being found at the same time between heterozygous SNPs and CNVs. Recombinant inbred mice had a similar heterozygous SNP and CNV profile to that of the classical inbred mice, with autosomes primarily characterized as having no spatial associations between these two events (84.3% of autosomes) and a moderate number of proximal associations (11.5% of autosomes).

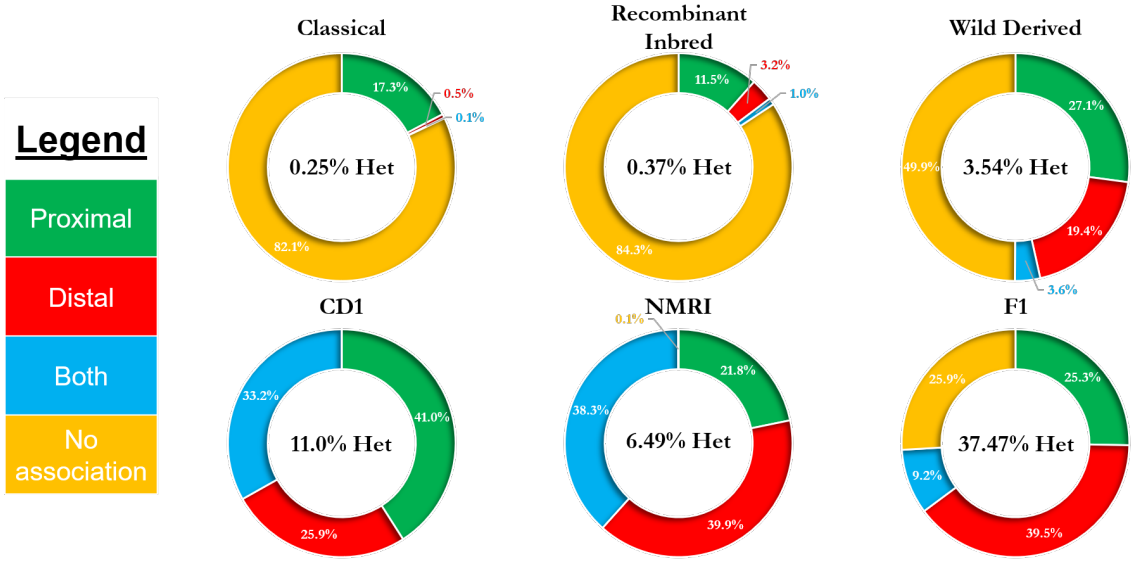

Figure 3.2: Heterozygous SNP loci and CNVs are frequently spatially associated with one another on autosomes from 707 MDGA samples. Circle plots are ordered left to right in two rows from lowest to highest SNP heterozygosity mouse cohort with average autosomal SNP heterozygosity displayed in the center. As per the legend, bar colour represents the type of spatial association between heterozygous SNP loci and CNVs called by the J statistic analytical pipeline ($\alpha = 0.05$, J statistic, Monte Carlo simulations = 1000) while bar size represents the percentage of autosomes assayed with at least one CNV. Differences in the relative abundance of proximal associations between cohorts is of particular interest to evaluate the effect of clustered SNP heterozygosity on CNV formation

3.4 Verification of expected single-nucleotide heterozygosity for 96 Axiom MouseHD array samples

The average call rate of the 96 Axiom MouseHD array three-generation mouse line samples was 99.71%, with all samples passing the recommended 98.5% default SNP call rate QC measure. The sex of all 96 samples was also correctly determined by the Axiom Analysis Suite genotyping algorithm. The average assayed genomic SNP heterozygosity across the 24 parental samples was 0.61%, with a minimum of 0.25% and a maximum of 1.26% (Table 3.3).

Comparing the empirically determined homozygous genotypes called by all six DBA/2J tail samples (average SNP heterozygosity of 0.51%) to all six C57BL/6J tail samples (average SNP heterozygosity of 0.68%), there are 110,580 autosomal loci of 567,856 total assayed autosomal loci were expected to generate heterozygous calls in all F1 hybrids. The expected heterozygous SNP loci in all F1 individuals thus equated to a minimum of 19.47%. The average assayed genomic SNP heterozygosity across the 24 F1 samples was 20.47%, with a minimum of 20.32% and a maximum of 21.57% (Table 3.4). The average assayed genomic SNP heterozygosity across the 48 F2 samples was 10.09%, with a minimum of 6.87% and a maximum of 12.75%.

Table 3.3 Determined whole genome SNP heterozygosity values for 96 Axiom MouseHD samples from six three-generation mouse lines

Mouse cohort[1]	Number of samples	Genomic SNP Heterozygosity (%)		
		Average	Minimum	Maximum
Parental	24	0.61	0.25	1.26
F1	24	20.47	20.32	21.57
F2	48	10.10	6.87	12.75

[1] Ordered by generation

Table 3.4 Determined autosomal SNP heterozygosity and CNVs detected for 96 Axiom MouseHD array samples from six three-generation mouse lines

Mouse cohort[1]	Number of autosomes	Average number of CNVs per autosome	Autosomal SNP Heterozygosity (%)		
			Average	Minimum	Maximum
Parental	456	3.05	0.60	0.12	1.45
F1	456	0.84	20.47	13.31	25.63
F2	912	1.87	10.09	0.21	25.26

[1] Ordered by generation

3.5 Single-nucleotide heterozygosity is not correlated with an elevated number of copy number variants per autosome from three-generation mouse lines

A total of 3,479 filtered CNVs were detected across 707 Axiom array samples (Fig. 3.3). Average autosomal SNP heterozygosity levels were similar to genomic SNP heterozygosity for parentals, F1s and F2s, with 0.60%, 20.32%, and 10.19%, respectively. The average number of CNVs per autosome for parentals, F1s and F2s was 3.05, 0.84, and 1.87, respectively. No correlation was observed between autosomal SNP heterozygosity and number of CNVs for parental mice ($r = 0.07$, $p > 0.001$) or F1 mice ($r = 0.01$, $p > 0.001$). A weak negative correlation was observed for F2 mice ($r = -0.27$, $p < 0.001$). There was no significant difference between the parental, F1, and F2 cohorts for total number of CNVs per autosome.

3.6 Heterozygous single-nucleotide polymorphisms and copy number variants are frequently spatially associated with one another in three-generation mouse lines

Of the 1,824 autosomes assayed from the three-generation mouse lines, 1,338 experienced at least one CNV event. The observed J statistic revealed a significant ($\alpha < 0.05$) spatial association between heterozygous SNPs and CNVs for 76.53% of autosomes assayed. In total, 359 autosomes had proximal associations, 516 had distal associations, 149 had both proximal and distal associations, and 314 had no association.

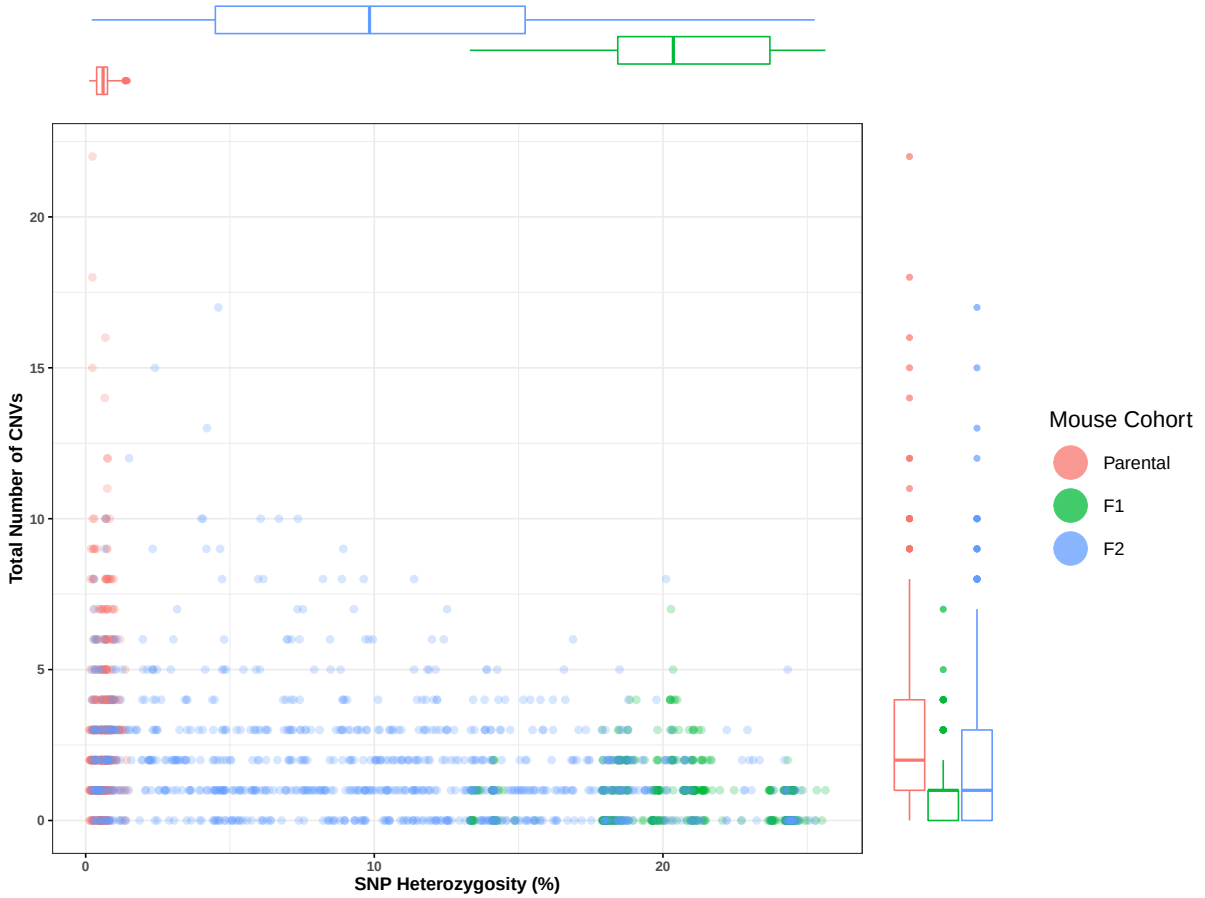

Figure 3.3: SNP heterozygosity is not correlated with a higher autosomal CNV burden in three-generation mouse lines The x-axis represents the SNP heterozygosity (%) while the y-axis represents the number of CNVs detected per autosome. All 19 autosomes from each individual are plotted. Marginal boxplots indicate the distribution of data for parental, F1, and F2 mice.

3.7 Heterozygous single-nucleotide polymorphisms and copy number variants are more frequently distal to one another in F1 and F2 mice from the three-generation mouse lines

The parental mice from the three-generation lines demonstrated a similar spatial relationship distribution to the classical inbred mice from the MDGA dataset, with 60.6% of autosomes having no association between heterozygous SNPs and total CNVs, 37.9% proximal associations, and only 1.5% distal associations (Fig. 3.4).

All autosomes from the F1 samples from the three-generation lines showed a significant spatial association between heterozygous SNPs and total CNVs, with 30.4% proximal associations, 41.5% distal associations, and 28.1% both proximal and distal associations. The difference between proximal and distal associations is greater in the F2 samples from the three-generation lines, with 18.9% proximal associations, 59.5% distal associations, 11.5% both proximal and distal associations, and 10.1% no associations.

3.8 Detection of inherited copy number variants in F1 mice from three-generation mouse lines is prone to false-negatives

The CNV landscape of 96 three-generation mouse line samples shows that CNV calling using the Axiom MouseHD array is capable of detecting inherited CNVs (Fig. 3.5). For example, a total of 36 recurrent CNVs were detected in both tissues of a parent, both tissues of a F1, and at least two tissues of a F2 from the same line across all six lines (Table 3.5). The average length of the robustly called inherited CNVs was 275,567 bp and approximately 81% of them were state 1 deletions. The remaining inherited CNVs detected were of 'mixed' state, indicating that different CNV states were called throughout the line. The average number of markers used to call the robustly called inherited CNVs across all lines was 76.7 while the minimum number of markers was 24. Three illustrative examples of robustly called inherited CNVs are highlighted

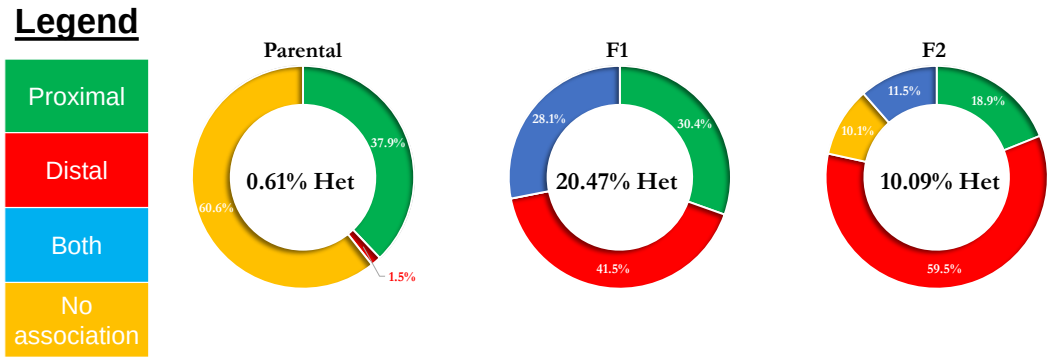

Figure 3.4: Heterozygous SNP loci and CNVs are frequently spatially associated with one another on autosomes from 96 three-generation mouse line samples. Circle plots are ordered by cohort generation with average autosomal SNP heterozygosity displayed in the center. As per the legend, bar colour represents the type of spatial association between heterozygous SNP loci and CNVs called by the J statistic analytical pipeline ($\alpha = 0.05$, J statistic, Monte Carlo simulations = 1000) while bar size represents the percentage of autosomes assayed with at least one CNV.

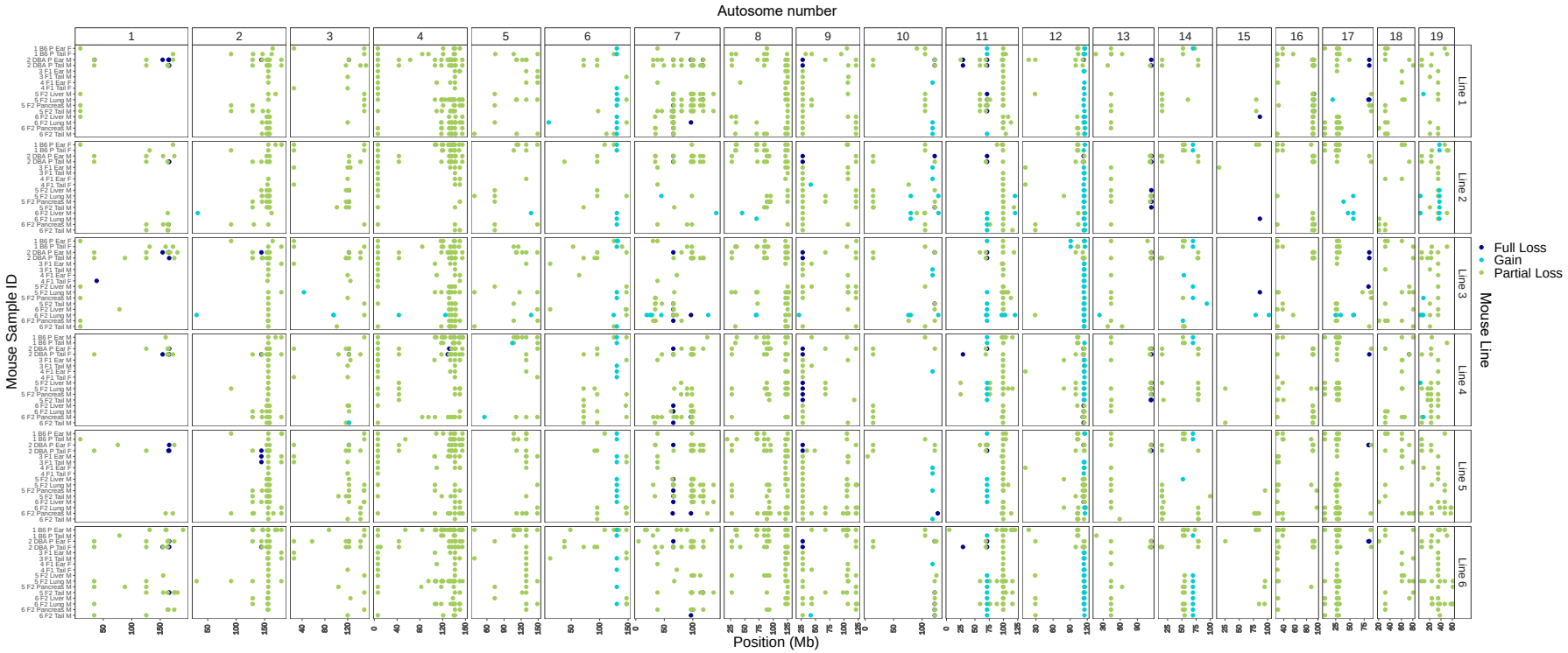

Figure 3.5: Genomic landscape of detected CNVs in 96 Axiom MouseHD array samples from six three-generation mouse lines
The x-axis represents the position along each autosome in Megabase pairs (1,000,000 bp) and autosome number is denoted at the top. Mouse sample ID nomenclature is individual mouse number within a line (1-6), strain type (B6 or DBA for parentals), cohort (P, F1 or F2), tissue type (Ear, tail, liver, lung, or pancreas), and sex (M or F) and is denoted on the left y-axis. Line number is denoted on the right y-axis. As per the legend, dark blue points are full loss (CN state 0) deletions, green points are partial loss (CN state 1) deletions, and teal points are duplications (CN state 3+).

Table 3.5 Instances of detected inherited recurrent CNVs found in three-generation mouse lines

Line	Total Number of CNVs	Number of Markers		Average CNV Length (bp)	CN State			
		Average	Minimum		0	1	3	Mixed
1	7	68	24	246 295	0	6	0	1
2	5	63	24	211 044	0	4	0	1
3	3	67	24	230 804	0	2	0	1
4	4	79	24	287 178	0	3	0	1
5	8	87	24	352 258	0	6	0	2
6	9	84	24	275 770	0	8	0	1

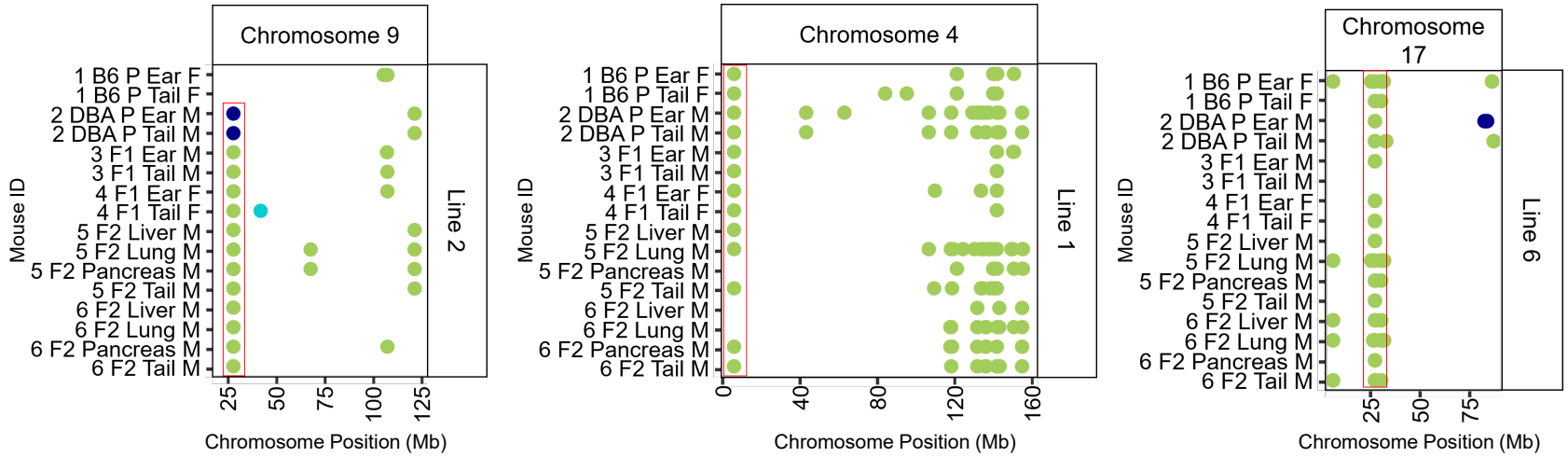

Figure 3.6: Three landscape examples of inherited CNVs in three-generation mouse lines. Examples from three different mouse lines shown in Fig 3.5 exemplifying instances of a robustly called inherited CNV detected in all three generations. The bottom x-axis represents autosomal position in Mb. The top x-axis denotes autosome number. Mouse sample ID nomenclature is individual mouse number within a line (1-6), strain type (B6 or DBA for parentals), cohort (P, F1 or F2), tissue type (Ear, tail, liver, lung, or pancreas), and sex (M or F) and is denoted on the left y-axis. The right y-axis denotes the mouse line. Red boxes surround inherited CNVs, considered as such if called in both tissues from at least one parent, both tissues from at least one F1, and at least two out four tissues from at least one F2.

in Fig. 3.6.

CNV landscape profiling also shows possible false negatives in CNV detection for the F1 cohort. For example, there are a total of 130 CNVs detected that are recurrent in at least two tissues of a parent and at least two tissues of a F2 mouse but not detected in any F1 mouse tissue (Table 3.6). The average length of these potential false negatives in the F1 cohort was 245,906 bp and approximately 83.1% were state 1 deletions, 3.1% were state 0 deletions, 1.5% were state 3 duplications, and 12.3% were of mixed state. The average number of markers used to call these potential false negative CNVs across all lines was 74.5, while the minimum number of markers was 10. Three illustrative examples of potential false negatives in CNV detection for F1 mice are highlighted in (Fig. 3.7).

3.9 HD-CNV analysis detects similar numbers of recurrent CNVs between mice from six three-generation lines

Singleton and recurrent CNVs for each autosome of each mouse line from the three-generation mouse lines are shown in Fig. 3.8. A CNV was considered recurrent if detected in at least two samples from within any given mouse line. From a total of 3479 detected CNVs, 685 identified recurrent CNVs were found across the six three-generation mouse lines (Table 3.7). A similar number of recurrent CNVs were found for each mouse line, with a minimum of 101 and a maximum of 127 across all lines.

3.10 *De novo* copy number variants are detected in all six three-generation mouse lines

Putative *de novo* CNVs in the F1 mice are those CNVs found in both tissues of an individual F1 mouse and not in any tissue from either parental (Table 3.8). There were a total of 13 *de novo* CNVs in the F1 mouse cohort across all six three-generation mouse lines. The average length of F1 *de novo* CNVs was 268,924 and called by at least 18 probe markers.

Table 3.6 Recurrent CNVs detected in parental and F2 mice but not found in F1 mice within a line indicate possible false-negative CNV calling in the F1 cohort

Line	Number of Total CNVs	Number of Markers		Average CNV Length (bp)	CN State			
		Average	Minimum		0	1	3	Mixed
1	34	90	19	295 830	0	31	0	3
2	17	61	11	201 062	1	13	1	2
3	10	91	35	317 072	0	7	1	2
4	22	54	17	167 358	3	16	0	3
5	18	62	14	197 869	0	16	0	2
6	15	89	10	322 451	0	12	0	3

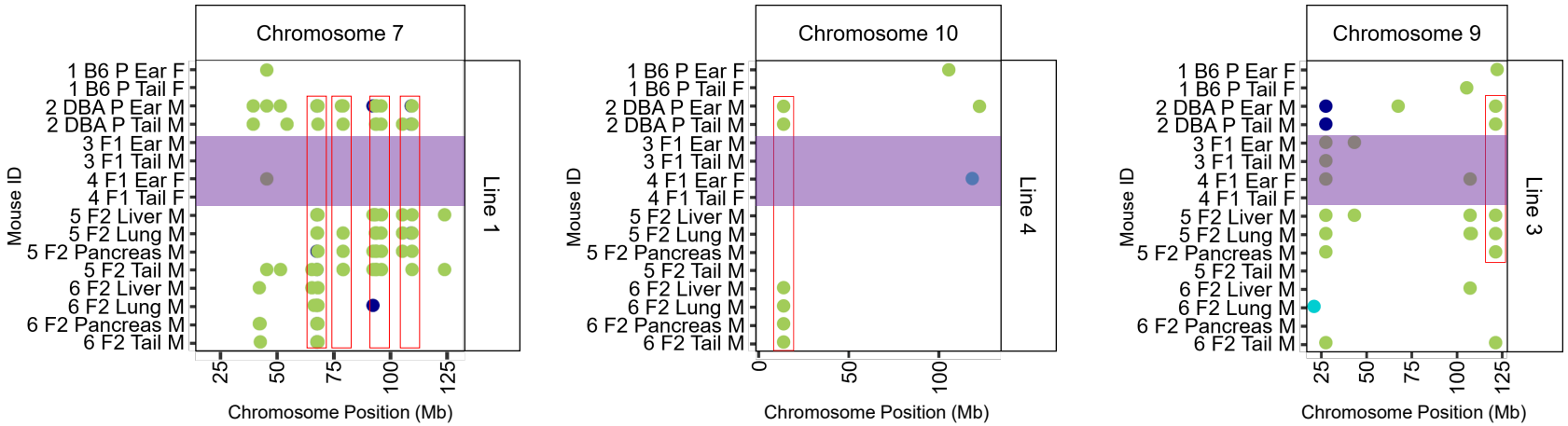

Figure 3.7: CNV landscape examples of potential false negatives in the F1 cohort of three-generation mouse lines. Examples from three different mouse lines shown in Fig 3.5 exemplifying instances of detected CNVs in parental and F2 mice, but conspicuously missing from both F1 mice. The bottom x-axis represents autosomal position in Mb. The top x-axis denotes autosome number. The left y-axis represents the sample, named for individual number in the mouse line, mouse type, tissue type, sex. The right y-axis denotes the mouse line. Red boxes surround possible false negative F1 CNV groups, considered as such if called in both tissues from at least one parent, neither tissue from either F1, and at least two out four tissues from at least one F2.

Table 3.7 Total and recurrent CNVs detected by HD-CNV analysis for 96 samples from six three-generation mouse lines

Line	Total CNVs			Recurrent CNVs		
	Amount	Average Length (bp)	Average Number of Markers	Amount	Average Length (bp)	Average Number of Markers
1	643	148 095	45	122	210 745	63
2	483	142 781	41	109	214 380	58
3	532	152 241	44	101	208 964	62
4	538	134 647	42	103	191 179	59
5	613	132 996	40	123	200 816	60
6	670	145 720	43	127	226 668	67

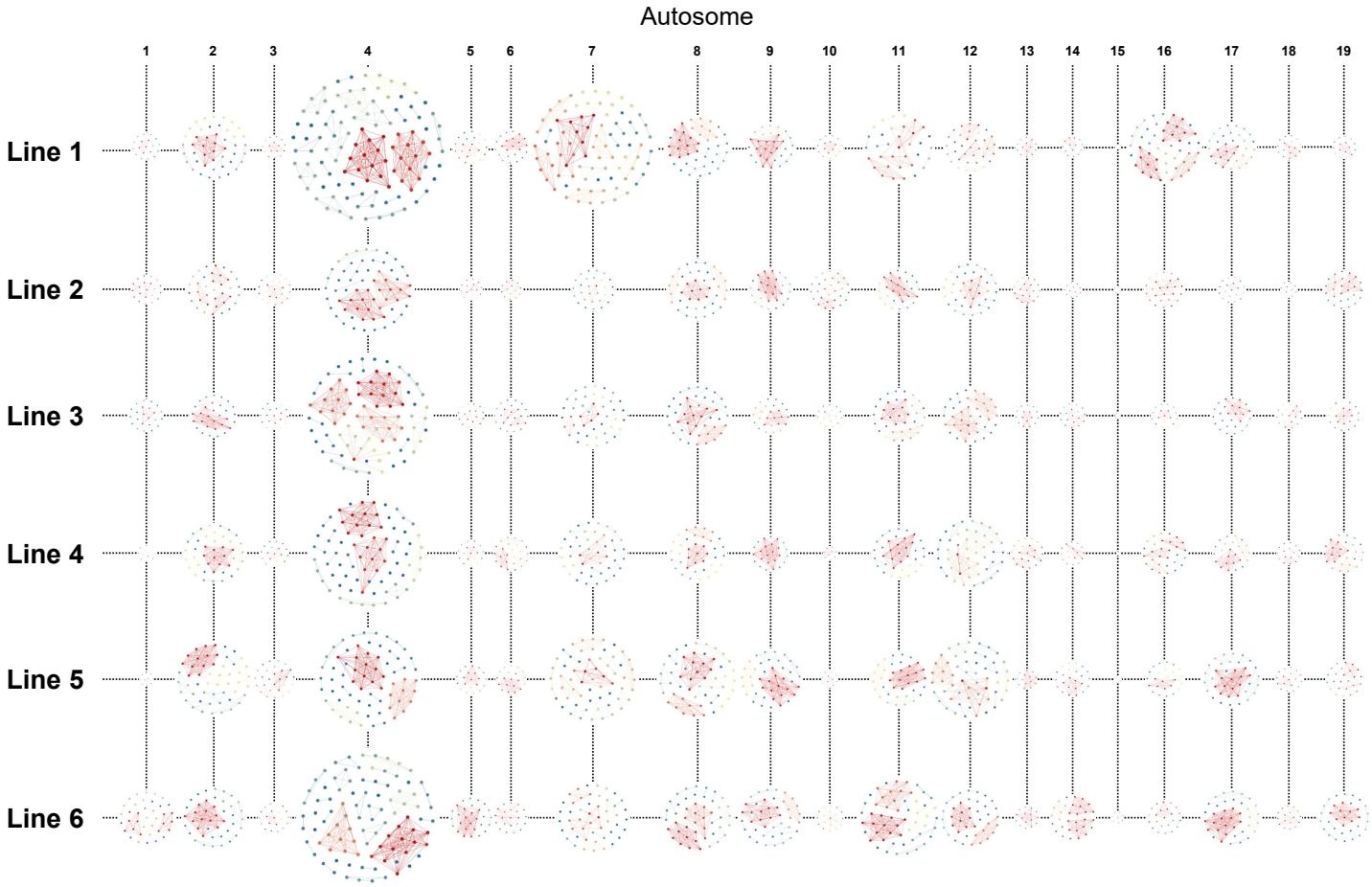

Figure 3.8: Gephi-based visualization of HD-CNV output for all 19 mouse autosomes for six three-generation mouse lines shows singleton and recurrent CNVs within lines. Each node (circle) represents a CNV and edges (lines) indicate CNVs that overlap reciprocally by at least 40%. Colour represents the number of CNVs involved in a merge region, with warmer (red) colours representing more CNVs in a merged region compared to cool colours indicating fewer CNVs in a merged region. CNVs were only merged for samples within lines and not between lines. The autosomes are ordered by number vertically left to right and the six lines are ordered horizontally top to bottom.

Putative *de novo* CNVs in the F2 mice are those CNVs found in at least two tissues from an individual F2 mouse and not in any tissue from the parental or F1 mice in its line (Table 3.9). There were a total of 43 *de novo* CNVs in the F2 mouse cohort across all six three-generation mouse lines. The average length of *de novo* F2 CNVs was 230,145 called by at least 16 probe markers.

3.11 F2 *de novo* copy number variants are proximally associated with F1 heterozygous single-nucleotide polymorphic loci

Of the 13 autosomes harbouring a *de novo* CNV in F1 mice, there were 7 proximal, 3 distal, and 3 both proximal and distal spatial associations between heterozygous SNP loci and each *de novo* CNV (α = <0.05, J statistic). All 13 autosomes had nonrandom spatial associations between heterozygous SNP loci and *de novo* CNVs (Fig. 3.9).

In the F2 mice, 43 *de novo* CNVs were found occurring across 36 different autosomes. Each *de novo* CNV was assessed individually and there were 7 proximal, 32 distal, and 1 both proximal and distal spatial associations between heterozygous SNP loci and *de novo* CNVs (α = <0.05, J statistic). As well, there were three instances of no association between heterozygous SNP loci and *de novo* CNVs in the F2 mice (Fig. 3.9).

Mapping the F2 *de novo* CNVs to the landscapes of their F1 parents yielded primarily proximal associations between F1 heterozygous SNP loci and F2 *de novo* CNVs. In the maternal heterozygous SNP landscape, the J statistic showed 33 proximal, 4 distal, and 6 both proximal and distal associations between heterozygous SNP loci and *de novo CNVs*. In the paternal heterozygous SNP landscape, the J statistic showed 32 proximal, 4 distal, and 7 both proximal and distal associations. All but one *de novo* CNV showed the same spatial association in either maternal or paternal F1 mouse heterozygosity landscapes (Fig. 3.9).

As an example, there was a *de novo* CNV on autosome four in the 9889 F2 mouse from line one starting at 118,364,287 bp and ending at 118,559,529 bp. Showing analysis for all mice

Table 3.8 *De novo* CNVs detected in 12 F1 mice from six three-generation mouse lines.

Line	Number of de novo CNVs	Average Number of Markers	Minimum Number of Markers	Average Length (bp)
1	0	0	0	0
2	3	48	31	178 501
3	2	42	19	193155
4	0	0	0	0
5	6	32	18	140 707
6	2	31	20	162 083

Table 3.9 *De novo* CNVs detected in 12 F2 mice from six three-generation mouse lines.

Line	Number of de novo CNVs	Average Number of Markers	Minimum Number of Markers	Average Length (bp)
1	13	55	22	297 131
2	11	57	17	259 934
3	3	34	28	249 236
4	4	44	32	145 774
5	9	49	16	154 995
6	3	41	32	149 504

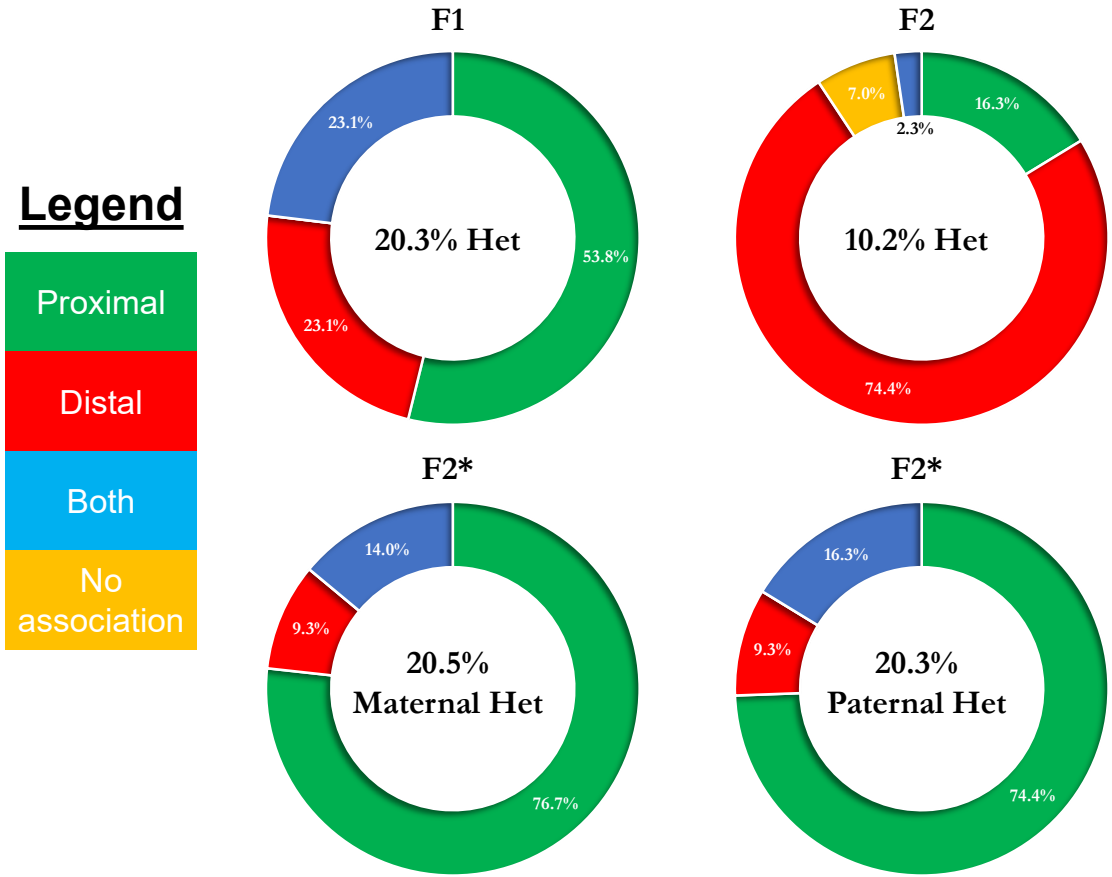

Figure 3.9: The majority of *de novo* CNVs on autosomes from the F2s of the three-generation mouse lines are proximally associated with F1 heterozyogous SNP loci As per the legend, bar colour represents the type of J statistical spatial association between heterozygous SNP loci and *de novo* CNVs ($\alpha = 0.05$, J statistic, Monte Carlo simulations = 1000) while bar size represents the percentage of autosomes assayed with at least one *de novo* CNV. F2* circle plots show the the J statistical spatial association between maternal or paternal heterozygous loci and F2 *de novo* CNVs. SNP heterozygosity of the assayed landscape is shown in the middle of the circle plots.

in line one, Rainfall plots for autosome four in all six mice show the spatial distribution of heterozygous SNP loci (Fig 3.10). Rainbow plots for autosome four in all six mice in line one show the distribution and proximity of total CNVs to heterozygous loci (3.11). The J statistic shows the spatial association between heterozygous SNP loci and total CNVs for all six mice in line one (3.12). Finally, Figure 3.14 shows the *de novo* CNV mapped to the heterozygous SNP landscape of the maternal 2309 F1 mouse with a proximal J statistic plot. Figure 3.13 shows the *de novo* CNV mapped to the heterozygous SNP landscape of the paternal 2314 F1 mouse with a proximal J statistic plot.

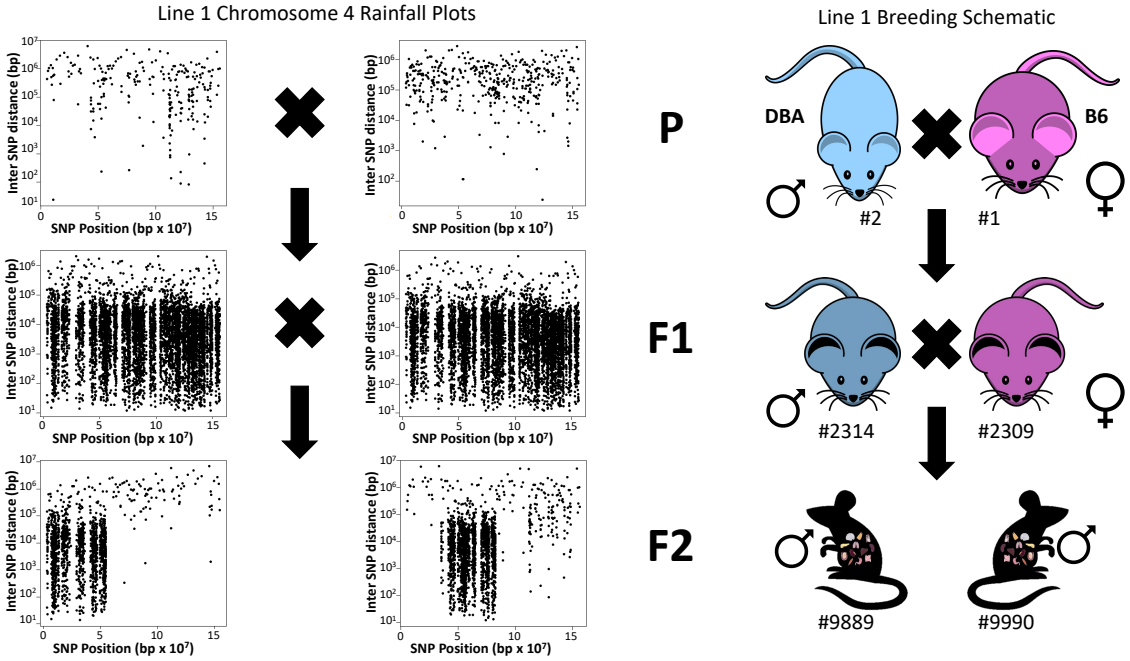

Figure 3.10: Rainfall plots for the first three-generation mouse line show the density and distribution of heterozygous SNP loci on autosome four. Rainfall plots correspond positionally to the individual mice from line one in the schematic on the right. The x-axis of each rainfall plot denotes autosomal position (bp). The y-axis of each rainfall plot shows the distance to the preceding heterozygous SNP locus (bp). The rainfall plots portray the landscape of interspacing between heterozygous SNP loci along autosome four for parental mice, F1 mice, and F2 mice.

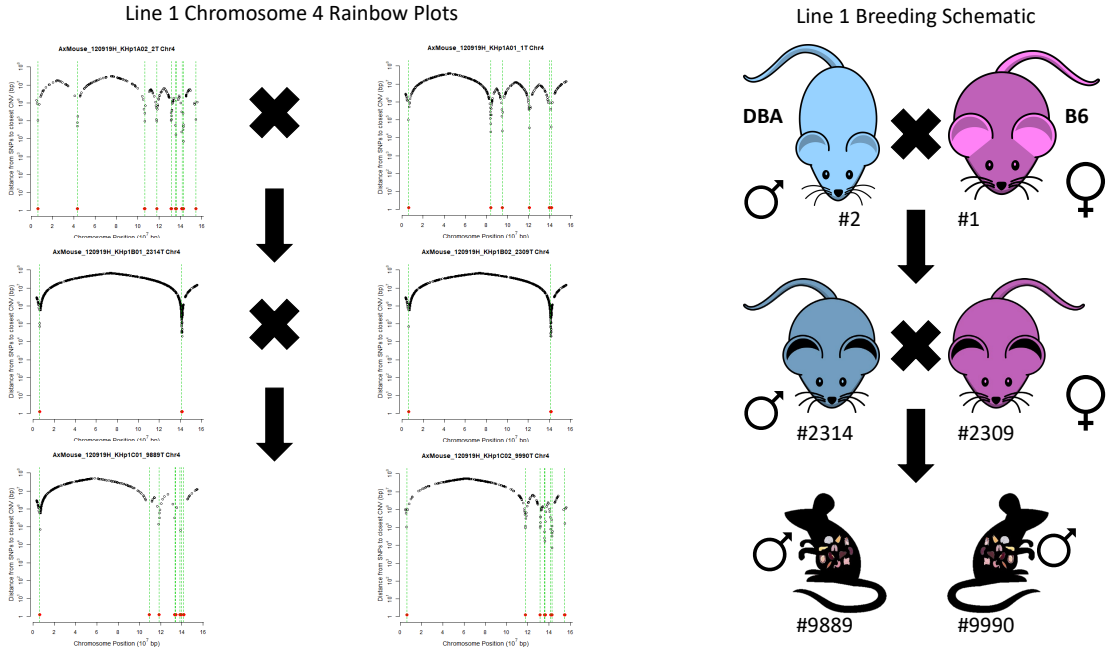

Figure 3.11: Rainbow plots for a three-generation mouse line show the distribution of CNVs and their proximity to heterozygous SNP loci on autosome four. Rainbow plots correspond positionally to the individual mice from line one in the schematic on the right. Black points represent heterozygous SNP loci and red points represent CNVs. The x-axis of each Rainbow plot denotes autosomal position (bp). The y-axis of each rainbow plot shows the distance of heterozygous SNP loci to their nearest CNV. Rainbow plots portray the landscape of heterozygous SNP loci in relation to CNVs along autosome four for parental mice, F1 mice, and F2 mice.

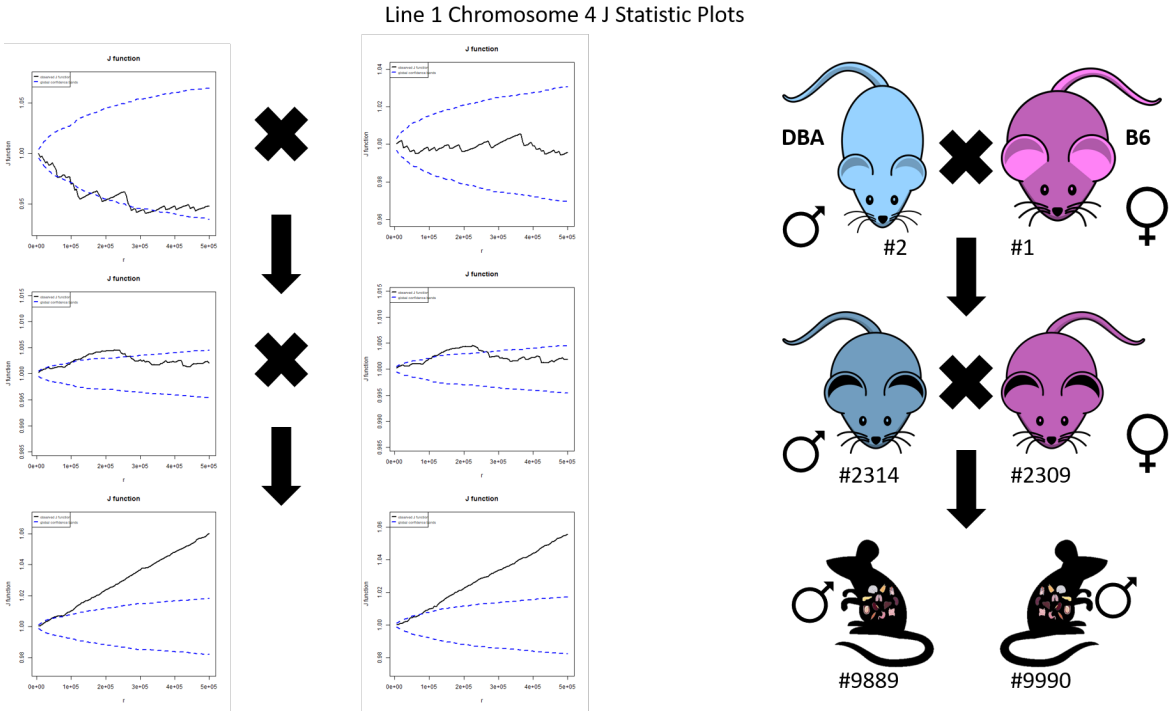

Figure 3.12: J statistic plots for a three-generation mouse line show the spatial relationship between heterozygous SNP loci and CNVs on autosome four. J statistic plots correspond positionally to the individual mice from line one in the schematic on the right. The blue lines represent global confidence bands (α <0.05, J statistic). The black line represents the observed J statistic generated by 1000 Monte Carlo simulations. The x-axis for each J statistic plot denotes the distance from CNV start and end positions in bp (r value). The y-axis for each J statistic plot shows the J function.

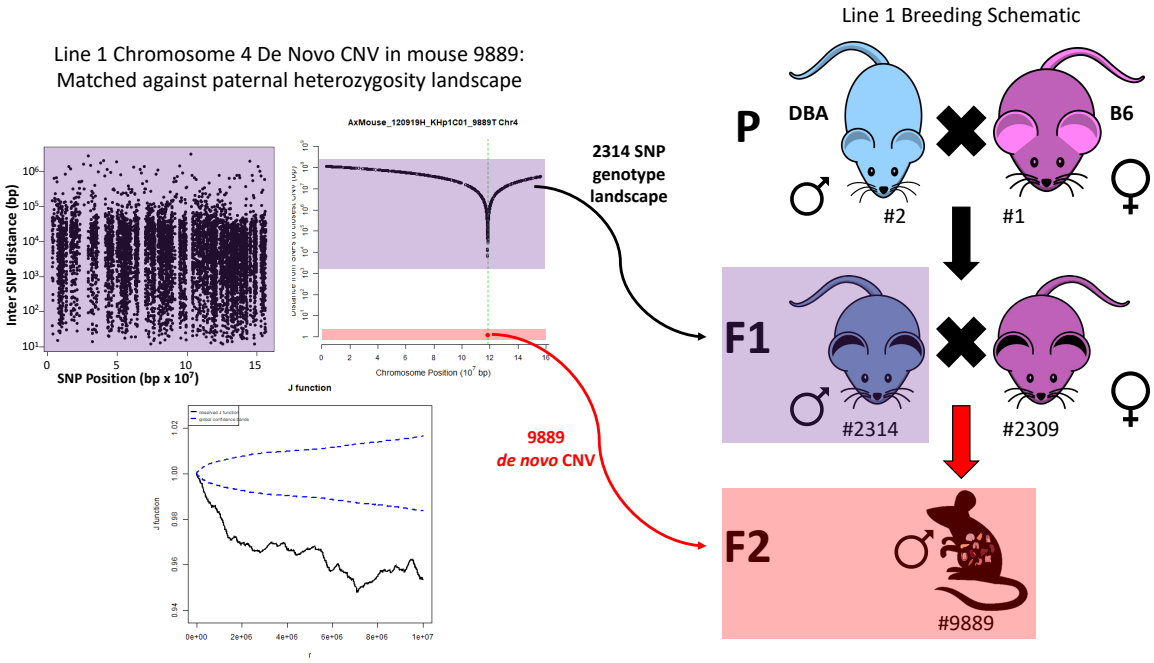

Figure 3.13: Rainfall, rainbow, and J statistic plots of a *de novo* CNV on autosome four in a F2 mouse from a three-generation mouse line matched against a paternal F1 heterozygous SNP loci landscape. The x-axis of the rainfall plot denotes autosomal position (bp) and the y-axis shows the distance to the preceding heterozygous SNP locus (bp). The rainfall plot shows the heterozygous SNP landscape of 2309 maternal F1 mouse. The x-axis of the rainbow plot denotes autosomal position (bp) and the y-axis shows the distance of heterozygous SNP loci to their nearest CNV, with points corresponding to heterozygous SNP loci from the 2314 paternal F1 mouse and red points corresponding to the *de novo* CNV from the 9889 F2 mouse. The J statistic shows the spatial relationship between the 2309 maternal F1 heterozygous SNP loci and the 9889 F2 mouse *de novo* CNV. The blue lines represent global confidence bands (α <0.05, J statistic). The black line represents the observed J statistic generated by 1000 Monte Carlo simulations.

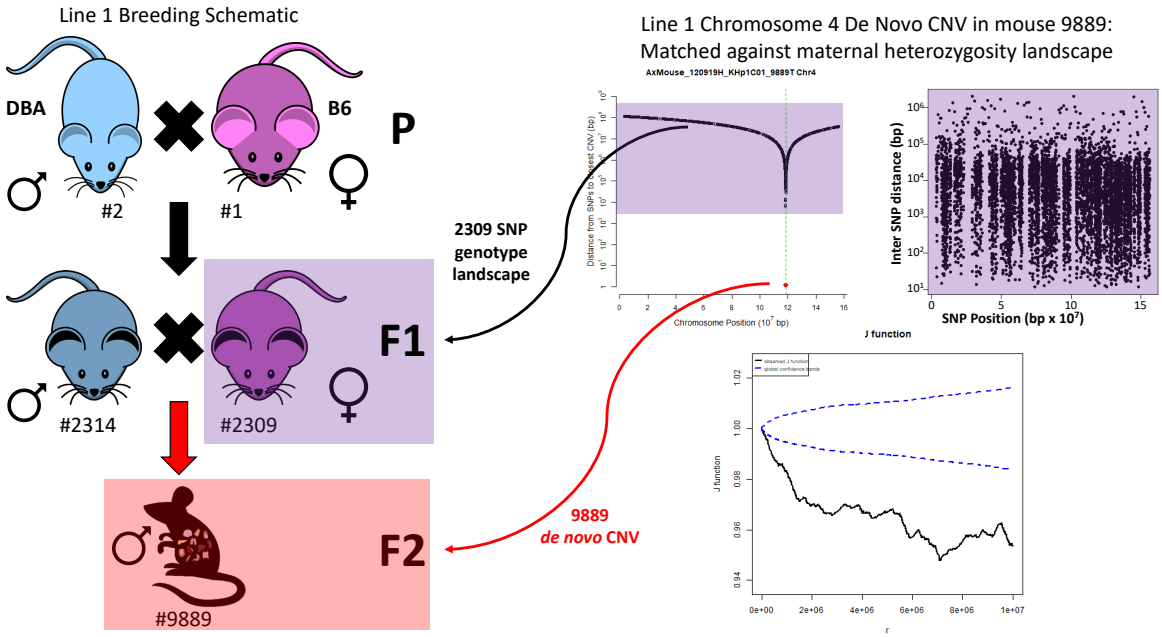

Figure 3.14: Rainfall, rainbow, and J statistic plots of a *de novo* CNV on autosome four in a F2 mouse from a three-generation mouse line matched against a maternal F1 heterozygous SNP loci landscape. The x-axis of the rainfall plot denotes autosomal position (bp) and the y-axis shows the distance to the preceding heterozygous SNP locus (bp). The rainfall plot shows the heterozygous SNP landscape of 2314 paternal F1 mouse. The x-axis of the rainbow plot denotes autosomal position (bp) and the y-axis shows the distance of heterozygous SNP loci to their nearest CNV, with points corresponding to heterozygous SNP loci from the 2314 paternal F1 mouse and red points corresponding to the *de novo* CNV from the 9889 F2 mouse. The J statistic shows the spatial relationship between the 2314 paternal F1 heterozygous SNP loci and the 9889 F2 mouse *de novo* CNV. The blue lines represent global confidence bands (α <0.05, J statistic). The black line represents the observed J statistic generated by 1000 Monte Carlo simulations.

Discussion

The recent discoveries of localized heterozygosity affecting the number of mutations within the genome of *Arabidopsis* [1] and peach plants [40] shed light on an inconspicuous endogenous factor potentially contributing to the development of genetic disease and the process of evolution in unanticipated ways. *Arabidopsis* and peach plant progeny derived from self-fertilization of high genomic heterozygosity F1 plants both exhibit 1.8- to 3.6-fold elevated point and indel mutations relative to progeny derived from self-fertilization of low heterozygosity parental plants [1, 40]. In *Arabidopsis*, the total number of point and indel mutations is associated wtih lower levels of genomic heterozygosity, as determined in the F3 and F4 selfed plants that exhibit a reduction in heterozygosity by one half for each generation. These mutations are found to be closer to regions of heterozygosity than would be expected based on random occurrence across the genome. These observations suggest that mutagenesis is enhanced by heterozygosity at least during gametogenesis and meiosis. However, these findings did not evaluate whether heterozygosity affects mutagenesis leading to the formation of larger genomic changes such as CNVs. Also, the universality of the phenomenon to mitosis and other cell types was not assessed. No evidence was provided for the existence of this phenomenon beyond plants. The nature of heterozygosity associated with *de novo* mutagenesis was not investigated. Properties of heterozygosity such as spatial distribution including clustering, density in clusters, number of clusters, cluster sizes and chromosomal distribution of clusters were not examined. Further, no attempt was made in either study to propose a possible mechanism by which heterozygosity could be mutagenic.

The breeding studies used in the plant models to track *de novo* mutagenesis with different levels and spatial distributions of heterozygosity exist for mice and in fact provide a greater

spectrum of known diverse landscapes of heterozygosity. More attributes of the chromosomal landscapes of heterozygosity can be examined in a context of meiosis to gain insight into mutational mechanisms regarding mutagenesis. If heterozygosity participates in the formation of *de novo* CNVs, genomes are predicted to have a greater number of CNVs and CNVs that are located in or near regions of heterozygosity. Similarly, comparisons of mice with breeding schemes yielding high and low levels of post-zygotic heterozygosity, but near-identical pre-zygotic heterozygosity, permit testing of a mitosis-linked mechanism of mutagenesis in offspring. If replication is prone to heterozygosity-induced formation of *de novo* CNVs, the mice with higher heterozygosity are hypothesized to have a greater number of CNVs and CNVs that are located in or near regions of heterozygosity.

This study sought to determine whether heterozygosity is associated with the formation of *de novo* CNVs in mice. The first experimental aim was to test whether there is an increased number of CNVs in relation to higher genomic heterozygosity during gametogenesis in mice. This was accomplished by evaluating number of CNVs in classical inbred mice, F1 mice, and wild-derived mice that are of low, high, and moderate heterozygosity, respectively, using a large dataset of publicly available MDGA data. The classical inbred and F1 mice both arise from low heterozygosity gametogenesis and showed no difference in their total number of CNVs, suggesting that heterozygosity does not affect *de novo* CNV formation post-zygotically. The wild-derived mice had significantly more CNVs, consistent with a meiosis-linked mechanism of mutagenesis, implicating heterozygosity.

Knowledge of the total number of CNVs per mouse is insufficient to demonstrate convincingly a correlation between heterozygosity and *de novo* CNVs. Therefore, the second aim of this study was to create an analysis pipeline by which two genomic features can be spatially co-profiled on thousands of autosomes. This was accomplished by applying a novel in-house R script to over 15,000 mouse autosomes. Spatial analysis of heterozygous SNP loci and total CNVs in classical inbred, wild-derived, and F1 mice showed no substantial elevation in the number of proximal associations between heterozygous SNP loci and CNVs in the F1 mice, but a greater proportion of proximal associations between heterozygous SNP loci and CNVs for the wild-derived mice, bolstering the evidence that heterozygosity is associated with CNV

formation during meiosis but not mitosis.

Direct comparisons of the heterozygosity of the mice in this breeding study with the heterozygosity found in the plant studies is challenging. With the missing inter-probe information for the mouse array, it is hard to provide an estimate of genomic heterozygosity adjusted for comparison with next-generation sequencing data. Comparison of whole genome sequencing data to quantitatively assess nucleotide diversity between B6 and DBA mice and subsequent comparisons with the plant models would be needed for direct comparisons at the same and highest resolution. These comparisons are possible but not yet done.

My final aim was to conduct a breeding experiment mimicking that of the *Arabidopsis* and peach studies. Low heterozygosity parental mice, high heterozygosity F1 mice, and moderate heterozygosity F2 mice were produced. If heterozygosity influences the number of *de novo* CNVs formed during gametogenesis in mice, it is expected that the F2 mice would have an elevated number of *de novo* CNVs that arise in or near regions of heterozygosity. Genotyping of these mice using the Axiom MouseHD array, no significant difference in total number of CNVs was found between the parental mice, the F1 mice, or the F2 mice, however, the number of CNVs in the F1 mice may have been underestimated due a limited, genetically homogenous training set for the PennCNV calling algorithm. The F2 mice were found to have 43 *de novo* CNVs in comparison to only 13 in F1 mice, while *de novo* CNVs could not be detected in parental mice. While This finding is consistent with a meiosis-linked hypothesis of mutagenesis, the difference in *de novo* CNVs may be in part due to false-negative CNV detection primarily in the F1 cohort. The spatial relationships of heterozygous SNP loci to CNVs were more often in the F1 and F2 cohorts, providing evidence against a mitosis-linked mechanism of mutagenesis. The spatial relationships of gametogenic heterozygous SNP loci in the F1 mice compared to *de novo* CNVs in the F2 mice provide strong evidence that heterozygosity is indeed affecting CNV formation during meiosis in mice.

4.1 Total number of copy number variants in classical inbred mice and F1 mice suggests that heterozygosity does not elevate mutagenesis during mitosis

Classical inbred mice, F1 mice and wild-derived mice have low, high, and moderate, levels of genomic SNP heterozygosity, respectively (Fig. 4.1). Consistent with the logical outcome of inbreeding for more than twenty generations and previous findings, MDGA assayed classical inbred mice had <1% SNP heterozygosity [91]. Wild-derived mice have a SNP heterozygosity range of 1 – 16%, a moderate increase with respect to classical inbred mice that is reflective of a breeding scheme with fewer inbred generations. F1 mice are specifically bred with the purpose of increasing genetic heterogeneity between two different inbred mouse strains and the observed range of SNP heterozygosity of 18 – 46% is consistent with this intended elevation of genetic heterogeneity. Importantly, the cells undergoing gametogenesis that give rise to these F1 mice are derived from classical inbred mice and are therefore minimally heterozygous.

Across these three mouse cohorts, there was no match to expectations associated with genomic heterozygosity levels and total number of CNVs per mouse, indicating that genomic heterozygosity alone is not a predictor of elevated CNV formation. Notably, no significant difference in the total number of CNVs per autosome for the classical inbred mice (1.22 per autosome) compared to the F1 mice (1.15 per autosome) suggests that heterozygosity does not substantially contribute to CNV formation post-zygotically in these mice. If heterozygosity were a mitosis-linked mutagen, it would be expected that the high heterozygosity F1 mice from a low heterozygosity gametogenesis environment would harbour a greater number of total CNVs than the classical inbred mice.

Further evidence against heterozygosity acting as a mitotis-linked mutagen is provided by the intra-cohort analysis of autosomal SNP heterozygosity and number of total CNVs per autosome. These data show no positive correlation between autosomal SNP heterozygosity and number of total CNVs per autosome, indicating there is no noticeable increase in somatic *de novo* CNVs.

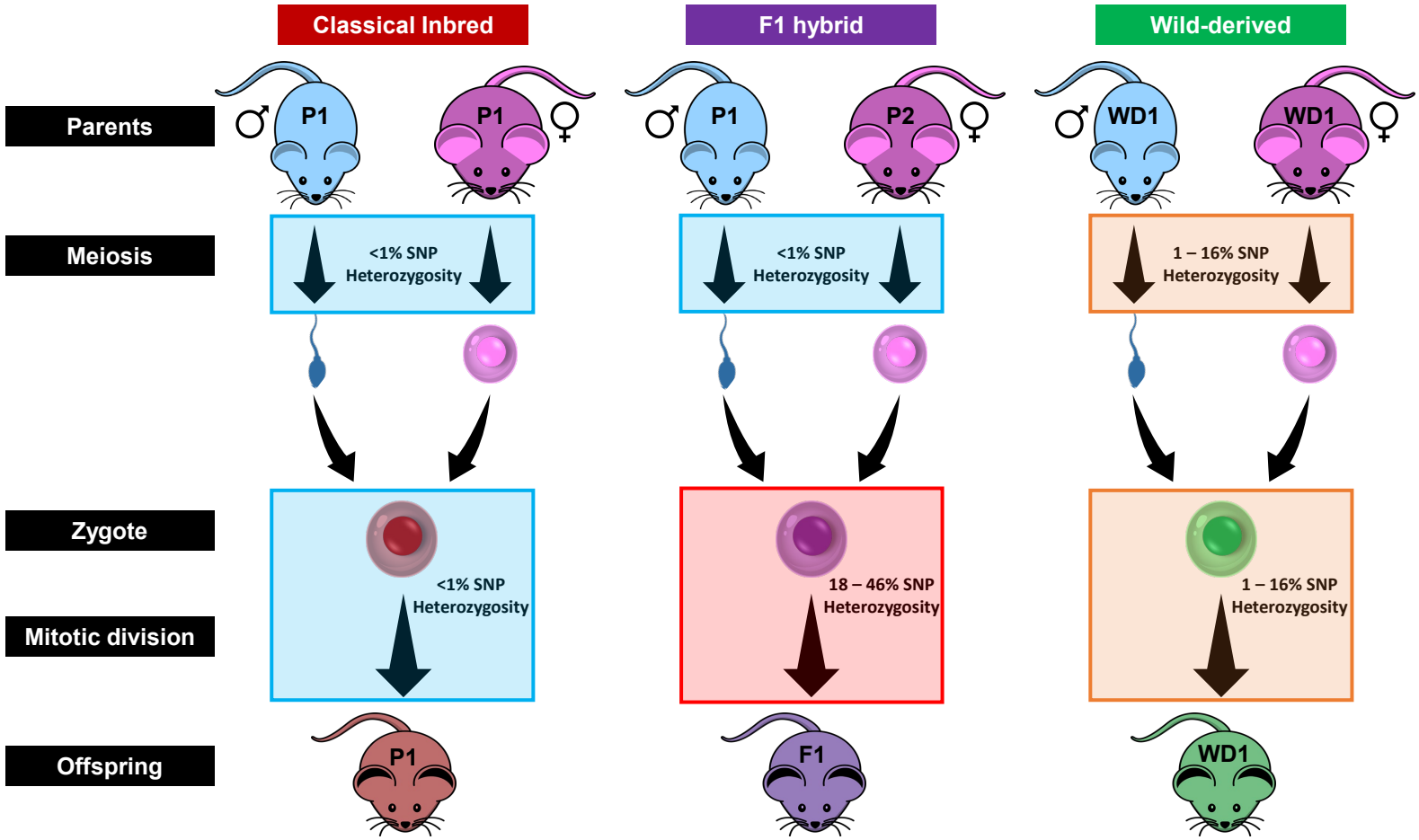

Figure 4.1: Relative single nucleotide polymorphic heterozygosity during gametogenesis and post-zygotically of classical inbred mice, F1 mice, and wild-derived mice. MDGA assayed SNP heterozygosity values are shown and characterized as low (blue boxes), moderate (orange boxes) and high (red box). During gametogenesis of classical inbred mice, F1 mice, and wild-derived mice, heterozygosity is low, low, and moderate, respectively. Post-zygotically, classical inbred mice, F1 mice, and wild-derived mice, heterozygosity is low, high, and moderate, respectively.

4.2 Total copy number variant spatial proximity to single nucleotide polymorphic heterozygosity in wild-derived mice is consistent with a meiosis-linked mechanism of mutagenesis

The hypothesis of heterozygosity contributing to CNV formation as a meoisis-linked mutagen in mice is supported by the relationship between total CNVs and heterozygous SNP loci in wild-derived mice. SNP heterozygosity during gametogenesis of cells that ultimately give rise to wild-derived mice is 1 – 16% compared to classical inbred mice and the assayed F1 mice that are typically <1%. Wild-derived mice had a more CNVs per autosome on average (3.13) than the classical inbred mice (1.22) and the F1 mice (1.15). Further, the heterozygous SNP loci on autosomes of wild-derived mice were more frequently close by CNVs than far away.

The spatial relationship between heterozygous SNP loci and CNVs in classical inbred mice is further evidence supporting a meiosis-linked hypothesis. Although the majority of autosomes with CNV events in these mice exhibited no relationship between heterozygous SNP loci and CNVs, of those that were spatially associated, 17.3% were proximal while only 0.5% were distal. These data suggest that, at least in some instances, *de novo* CNVs are associated with nearby heterozygosity in these mice.

The spatial relationship between heterozygous SNP loci and CNVs in F1 mice is not inconsistent with a meiosis-linked hypothesis. While initially counter-intuitive, it is not surprising that the majority of autosomes harbouring CNVs in F1 mice exhibited nonrandom spatial associations between heterozygous SNP loci and CNVs. This is because heterozygous SNP loci in these mice are discontinuously spaced in clusters, reflective of different homozygous allele haplogroups in the parental inbred strains. Since the vast majority of CNVs will be inherited rather than generated *de novo* and these clusters of heterozygous SNP loci were homozygous for many previous generations, it is unlikely that these heterozygous SNP loci represent a genomic context by which most CNVs found in these mice arose. CNVs that exist in a cluster of heterozygous SNP loci are more likely to be characterized as proximal by the J statistic while

CNVs that exist in a heterozygous SNP loci desert are more likely to be characterized as distal by the J statistic. CNVs that have been inherited for many generations may arbitrarily fall within a heterozygous SNP loci cluster or desert, therefore inflating significant results for an association.

Greater precision and information are achieved by examining those CNVs that arose *de novo* in an individual with empirical knowledge of chromosomal heterozygosity during gametogenesis. However, the publicly available MDGA data do not contain information with regard to the parentage of F1 mice and therefore precludes unambiguous assigning of *de novo* versus inherited status of the CNVs. This problem was remedied by conducting a breeding experiment with six three-generation mouse families wherein discovery of *de novo* CNVs was possible using the Axiom MouseHD array.

4.3 The Axiom MouseHD array accurately characterizes heterozygous single nucleotide polymorphic loci in six three-generation mouse lines

The Axiom MouseHD array is a recently developed, custom-made high-density microarray that, to my knowledge, has only been used in the methodology of one peer-reviewed paper [85]. Therefore, it was critical to evaluate carefully the genotyping calls made using this array to ensure accurate observations regarding heterozygous SNP loci and CNVs were obtained. The first step to ensure high quality genotyping data was to ensure high call rates of all samples. Call rate QC metrics of at least 97% have been obtained successfully for a number of other custom-made, high-density Axiom arrays, including the Axiom Apple 480K SNP array [92], the Axiom CicerSNP Array for chickpeas [93], the 600K Affymetrix Axiom HD Chicken Array [70], and many others. In this study, all 96 samples from the six three-generation mouse lines passed the 98.5% call rate QC threshold recommended by Axiom Analysis Suite software, indicating that the probes on the array are working as expected.

A further convincing indication of successful genotype calling was obtained by comparing

the predicted average SNP heterozygosity of the classical inbred parental mice (<1%), the F1 mice (approximately 20%), and the F2 mice (approximately 10%) to the relative average SNP heterozygosity of the parental mice (0.61%), the F1 mice (20.47%) and the F2 mice (10.09%). The DBA/2J and C57BL/6J inbred mice were expected to, and did have, <1% heterozygous SNP loci, in accordance with heterozygous SNP loci values for classical inbred mice called on the MDGA [91]. Importantly, predictions of the minimum number of heterozygous SNP loci in the F1 mice based on the homozygous genotypes of the parental mice were also correct. The F1 mice had little deviation in the range of number of heterozygous SNP loci between each mouse. Finally, as expected, the observed average number of heterozygous SNP loci in the F2 mice was approximately half that of the F1 mice and with greater inter-animal and inter-chromosomal variation than the F1 mice.

4.4 Total copy number variant occurrences and their spatial proximity to heterozygous single nucleotide polymorphic loci is not consistent with heterozygosity acting as a somatic mutagen

At genomic and chromosomal levels in the six three-generation mouse lines, the total number of CNVs per mouse was not indicative of a direct link between heterozygosity and *de novo* CNV mutation. There was no significant difference in the average number of CNVs per autosome between the low heterozygosity parental mice, the high heterozygosity F1 mice, and the moderate heterozygosity F2 mice. As well, similar to the previous finding in the publicly available classical inbred, wild-derived, and F1 mouse samples, the intra-cohort autosomal heterozygous SNP loci in each cohort from the six three-generation mouse families do not correlate with an elevation in CNV occurrences. This is once again consistent with the idea that heterozygosity of an autosome is not a predictor of the number of CNVs it will harbour. These findings provide further evidence that heterozygosity does not play a role in CNV formation post-zygotically.

Of note, the finding that there is no significant difference in the number of CNVs per autosome between the F1 mice and the F2 mice suggests that even if heterozygosity is a meiosis-linked mutagen, it is not introducing a large enough number of *de novo* CNVs for a significant increase to be detected between these mice. However, it is important to account for well-known methods of CNV formation such as NHEJ and retrotransposition that are unlikely to be affected by heterozygosity. Further, CNVs arise rarely in the germline of typical laboratory mice. A rough, conservative estimate of 0.6 per mouse per generation has been proposed [94]. Considering the average number of total CNVs on mouse autosomes has been found to be approximately 28.84 per individual [75], the expected modest increase in *de novo* CNVs per generation makes identifying factors affecting the rate of CNV formation challenging.

The spatial association between heterozygous SNP loci and CNVs on autosomes in the parental, F1, and F2 cohorts is not consistent with heterozygosity affecting CNV formation post-zygotically. If heterozygosity did affect CNV formation post-zygotically, it might be expected that the high heterozygosity F1 mice would show the most proximal associations whereas the low heterozygosity parentals would show the least. This was not observed. For the parental mice, the F1 mice, and the F2 mice, there were 37.9%, 30.4%, and 18.9% of autosomes demonstrating proximal associations between heterozygous SNP loci and CNVs, respectively. Importantly, these results are not inconsistent with the hypothesis that heterozygosity is a meiosis-linked mutagen. To uncover the role that heterozygosity plays in CNV formation, it is necessary to investigate features of *de novo* CNVs in the F2 cohort in greater detail.

4.5 The Axiom array can be adapted to successfully identify copy number variants

Although the Axiom MouseHD array was not designed with CNV calling in mind, the high probe density allows well established CNV detection software such as PennCNV to be used effectively. This is the first time that the Axiom MouseHD array has been used for CNV calling, but other Axiom array platforms have been successfully adapted for CNV detection with similar methodology [95, 96].

Strong evidence validating successful CNV detection by the Axiom MouseHD array was provided by finding 36 robustly called inherited CNVs across six three-generation mouse lines. These CNVs were independently called in at least six different tissues and serve as an effective internal check that the Axiom Array can indeed make reliable CNV calls across samples with differing genotypes.

One drawback to using the Axiom MouseHD array is the small pool of samples that have been genotyped using the array. Only one previous study has made available CEL files that were genotyped using the Axiom array [85]. Ideally, a training set for genotyping and CNV detection would consist of several hundred samples that are genetically similar to the test set [89]. However, the training set used for this study comprised only 88 samples that were all backcrossed F2 mice of C57BL/6J and BALB/cJ genetic background. This homogenous training set reduces the capacity for PennCNV to make high quality CNV calls reliably, particularly in the F1 mice that are much more heterozygous than those in the training set. This is illustrated by the finding that 130 CNVs were detected in both a parental mouse and an F2 mouse from the same line, but were conspicuously missing from either F1 mouse. The likelihood that a CNV exists in a parental mouse, is not inherited by either F1 mouse, and arises *de novo* in the F2 progeny is exceptionally low. It is likely that the majority of missing F1 CNV calls are representative examples of false-negatives that are a product of a training set that could be larger and more heterozygously diverse.

4.6 The Axiom MouseHD array is a reasonable alternative to the Mouse Diversity Genotyping Array for mouse genotyping and CNV detection

The Axiom MouseHD array is a practical addition to the toolkit of technologies for assaying genetic variation in the mouse genome. It is inexpensive and, as demonstrated by accurate calling of heterozygous SNP loci in the F1 mice, capable of high fidelity SNP genotyping. While the MDGA is an established technology that has been shown capable of accurately detecting CNVs in mice with a range of genetic backgrounds [75], the Axiom MouseHD array

had never been used for CNV detection prior to this study. As shown by the detection of 36 inherited CNVs called in three generations of mouse lineages, the Axiom array is capable of CNV detection. It is possible as a future next step to determine if the CNVs discovered in this thesis have been reported previously toward validation and assessment of the confidence of the biological existence of the CNV calls. However, CNV detection capacity of the Axiom array is limited in comparison to the MDGA. Without IGPs, the Axiom array relies only on SNP probes and therefore a higher inter-probe distance, meaning that smaller CNVs may not be detected.

Another severe drawback to the Axiom array is the lack of previous samples run using this technology. Without access to a large and diverse set of Axiom array CEL files, insufficient training of CNV calling algorithms reduces accuracy in calling, particularly for test samples that are too distant from the training set samples in terms of genetic diversity. This is best illustrated by the high apparent rate of false-negatives in the F1 cohort from this study.

Regardless of its shortcomings, the Axiom array provides high quality SNP genotyping of mice and serviceable CNV detection that will improve as more CEL files become available to improve the training set. Combined with it's low cost, the Axiom array is an indispensable addition to the mouse molecular toolkit.

4.7 *De novo* CNVs proximal to the heterozygous SNP landscape of F1 mice support the hypothesis that heterozygosity is a meiosis-linked mutagen

The relationship between heterozygous SNP loci and *de novo* CNVs in the F2 mice is further evidence against heterozygosity acting as a mitosis-linked mutagen. Of the 43 *de novo* CNVs found in this cohort, only 7 were proximally associated with heterozygous SNP loci whereas 32 were distally associated with heterozygous SNP loci. The landscape of heterozygosity in the F2 presents an interesting discontinuous pattern of regions of high heterozygosity bordered by deserts of heterozygosity. This alternating pattern in levels of heterozygosity creates a bias

for detection of non-random (i.e., associations) between *de novo* CNVs and heterozygosity. There is a tendency in this landscape for *de novo* CNVs to be either located within or nearby to clusters of heterozygous SNP loci or located in deserts and far away from clusters of heterozygous SNP loci. The spacing of clusters and deserts causes 'no association' calls by the J statistic to be underrepresented. Given that *de novo* CNVs were not often found to be proximal more frequently than distal to heterozygous SNP loci, an exclusively meiotic context for the phenomenon of heterozygosity-induced mutation is implicated.

Mapping those same 43 *de novo* CNVs to the landscape of heterozygous SNP loci in the F1 mice yielded a dramatically different pattern of spatial association. Altogether, 33 proximal associations and only four distal associations were found. This is consistent with the hypothesis that heterozygosity is a meiosis-linked intrinsic mutagen. It appears that *de novo* CNVs detected in the F2 mice tend to arise more often within or close to regions of heterozygous SNP loci during gametogenesis.

Although previous studies linking heterozygosity to elevated levels of mutation did not investigate large structural changes such as CNVs, parallels may still be drawn between their findings and those of this study. For instance, Yang *et al*. found that *de novo* point mutations in *Arabidopsis* were more likely to occur in a region of higher heterozygosity [1]. They also found a positive correlation between number and location of mutation and crossover events, suggesting that meiotic recombination is mutagenic. They speculated that the underlying mechanism of this phenomenon may be linked to poor pairing of homologous chromosomes during DSB repair. This is notable because it has recently been shown that there is a total of approximately 300 induced DSBs per meiosis in mice, of which most are repaired by HDR [97]. Failure of homology search could increase due to heterozygosity-induced mismatching between the strand to be repaired and template. If repair proceeds using the incorrect template, large CNVs could arise. The results by Yang *et al*. are consistent with the hypothesis that heterozygosity is a meiosis-linked mutagen.

4.8 Proposed mechanism of heterozygosity affecting meiosis-linked mutagenesis leading to copy number variant formation

The finding that *de novo* CNVs arise in proximity to clusters of heterozygous SNP loci during gametogenesis suggests that heterozygosity affects mutagenesis leading to CNV formation in mice. These results are consistent with a speculative mechanism of NAHR by which heterozygosity could be upregulating the rate of CNV mutagenesis during meiotic recombination (Fig 4.2).

During meiosis, 200 to 400 programmed DSBs are generated in mouse gametocytes, mediated by an initiator of meiotic double stranded breaks (SPO11) [98, 99]. These programmed DSBs are repaired almost entirely by HDR mechanisms; NHEJ is suppressed during meiosis [100, 101]. Repair of the DSBs begins with 5' end resection, followed by homology search of the 3' ssDNA ends. In most organisms, including mice, the homologous chromosome is primarily used as the template for repair as opposed to the sister chromatid [102, 103]. Homologous chromosomes are aligned during programmed DSB induction, facilitating preferential repair using the corresponding allele [104]. Once a template of sufficiently high sequence complementarity is found, a D-loop is formed and repair can proceed with either a crossover or a noncrossover event [105]. Noncrossover events result in simple gene conversion, the nonreciprocal exchange of a few hundred nucleotides from the template strand to the damaged strand [106]. Crossover events result in the reciprocal exchange of large portions of homologous chromosomes and are critical to the shuffling of genetic information in sexually reproducing organisms. Although noncrossover events are estimated to occur more frequently than crossover events by ten-fold in mice, there are still 1-2 crossovers estimated to occur per chromosome [107].

Importantly, NAHR has been shown to occur during the repair of SPO11-mediated DSBs [59]. After strand resection and during homology search by the ssDNA for its allelic template on the homologous chromosome, encountering a nearby non-homologous region of high sequence

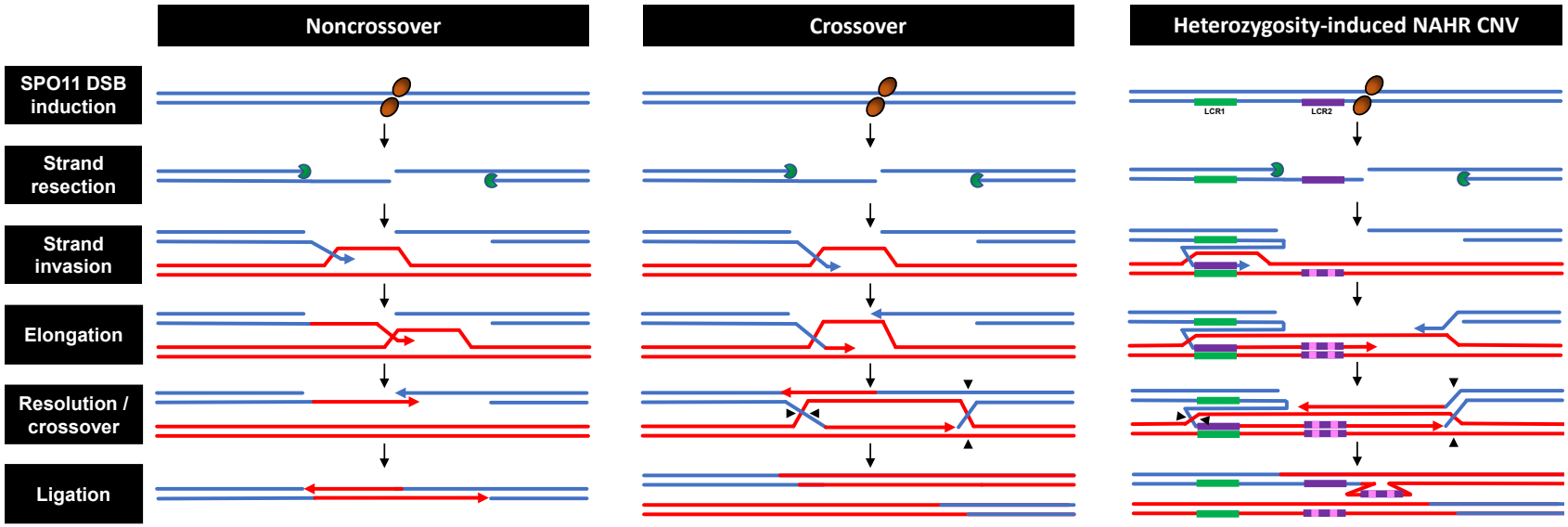

Figure 4.2: Proposed mechanism of heterozygosity-induced NAHR leading to CNV formation during meiotic recombination. Three pathways of repair of SPO11-induced DSBs are shown. For all pathways, 5' ends of strands are resected to expose 3' overhangs that may perform strand invasion of the homologous chromosome to form a D-loop. **Noncrossover:** Following elongation of 3' end, the D-loop is resolved and the remaining 3' end elongates using the newly synthesized complementary strand. Ligation seals the nicks. **Crossover:** Elongation of both 3' ends occurs simultaneously, using opposing strands of the homologue as a template. Crossover events result in the exchange of chromosomal material between homologues. Ligation seals the nicks. **Heterozygosity-induced NAHR CNV:** Green and purple rectangles represent LCRs. Strand invasion targeting the paralogous LCR may be upregulated by heterozygous mismatches between the homologous allele. Elongation and resolution follows that of a normal crossover event, resulting in a CNV duplication on the repaired chromosome.

complementarity can ultimately result in duplications or deletions to the damaged chromosome. In mice, an abundance of LCRs and other sequence repeats promote NAHR during meiotic recombination. Heterozygosity of homologous chromosomes may accidentally promote the likelihood of NAHR. Sequence divergence between alleles can dramatically reduce recombination rates. For example, as little as two nucleotide mismatches of sequence between the damaged strand and the template can reduce the rate of recombination in mouse cells by up to 20-fold [108]. If another high sequence identity template exists nearby, large scale duplications or deletions may occur. If DSBs are introduced into regions with LCRs close by, delays in homology search caused by heterozygosity may increase the rate at which erroneous strand invasion of repeat sequences occurs. The spatial association between heterozygous SNP loci in F1 mice and *de novo* CNVs in F2 mice is consistent with such a mechanism of NAHR mutagenesis leading to the formation of CNVs in mice during meiosis, although definitive evidence for the specifics of such a mechanism are lacking and are an ideal next step for future studies.

4.9 Study limitations

Whereas the MDGA and Axiom array are high-density genotyping arrays, they do not provide information concerning inter-probe heterozygosity. Therefore, this study is limited to sampling only a fraction of the heterozygosity present within the mouse genome. Additionally, high-density genotyping arrays suffer from sensitivity limitations in that they are more likely to detect CNVs that are in the majority of cells in a tissue sample. Therefore, post-zygotic *de novo* CNVs that do not occur early in development or in rapidly dividing persistent cell subpopulations may not be detected by this technology. These problems of resolution and sensitivity could be remedied by performing whole-genome sequencing with high read depth, although the cost of such an endeavor might be prohibitively expensive for non-human organisms with large genomes such as the mouse.

The use of publicly available MDGA data is also a limitation. While strain-specific CNVs could be validated using another mouse of the same strain, the tissue used to generate the available CEL files is not available for independent confirmation and therefore mouse-specific CNVs cannot be validated. As well, without parent-progeny information, it is impossible to

distinguish between inherited and *de novo* CNVs from the publicly available data. Further, grouping strains of similar breeding schemes does not necessarily account for fixed genetic differences between strains. To account for these fixed genetic differences, increasing sample size and separating analysis by strain would be ideal to ensure reproducibility.

Detection of CNVs using microarrays is prone to false positive and negatives. The choice of CNV-calling algorithm can result in significant differences in CNV detection, with one study showing less than 50% concordance in CNV calls between two algorithms [109]. The density of the array also affects false negative rates, owing at least in part to smaller CNVs not being called due to large inter-probe distances. A recent study showed a 15-fold increase in the number of CNVs detected by a high-density 700k bovine array in comparison to a medium density 50k bovine array [110]. Another limitation of this study was the inability of PennCNV to make reliable CNV calls for the sex chromosomes. In particular, the CNVs detected in samples processed by the Axiom array were far below QC cutoffs on the sex chromosomes, thus requiring this study to focus only on autosomes. Unfortunately, this means that sex differences in gametogenesis could not be investigated. If several hundred more Axiom array CEL files become available that are of similar genetic background to the test set, it is anticipated that CNV calling reliability will improve for both autosomes and sex chromosomes.

Finally, CNVs detected in this study, in particular putative *de novo* CNVs, have not yet been validated by another method such as ddPCR or high read depth whole genome sequencing. While CNV filtering steps such as requiring a high marker count per CNV call were used, the false positive rate for the Axiom array remains unknown. This could mean that some erroneous CNV calls in the six three-generation mouse lines have been reported.

4.10 Study contributions and future directions

The preeminent finding of this study is that the majority of *de novo* CNVs found in F2 mice from six three-generation mouse families are proximally associated with heterozygous SNP loci in F1 mice, providing strong evidence that heterozygosity is indeed playing a role in the formation of these CNVs during meiosis. This is significant because it is the first study to

show a link between heterozygosity and CNV formation in any organism. It is also the first study to investigate the link between heterozygosity and elevated numbers of nearby mutations in mammals. Additionally, this study contributes new evidence supporting a meiosis-linked mechanism of mutagenesis by heterozygosity as well as evidence that does not support a mitosis-linked mechanism. These findings are consistent with those of studies into *Arabidopsis* and peaches [1,40]. A proposed mechanism of NAHR exacerbated by strand invasion inhibition by heterozygosity is consistent with the findings of this study.

This study contains a number of novel accomplishments using high-density genotyping arrays. For example, advances in computing power allowed 707 MDGA samples to be genotyped together, increasing genotyping accuracy, up from the previous largest group of 351 by Locke *et al*. [75]. This study also verified the utility of the Axiom MouseHD array for SNP genotyping and CNV detection. Another accomplishment is the development and application of the automated analytical pipeline for assessing the spatial relationship between two genomic features. Doing so for a total of 15,247 autosomes, this study provides a rich database of genotype and CNV information for a wide variety of mice that may be used in future studies. For example, more in-depth analysis of CNV length, location, and spatial relation to other genomic features such as recombination hotspots, genic regions, and repeat regions may shed more light on the necessary environment by which heterozygosity acts as a mutagen. Finally, the breeding experiment yielded a number of tissue samples uniquely valuable for the eventual tracking of *de novo* and inherited genetic variation in the context of specific tissues and cell types. Future researchers can increase the resolution and sensitivity of the experimental design to validate and extend the findings of this study. Specifically, confirmation of putative *de novo* CNVs could then be followed by sequence analysis of breakpoints to determine whether CNVs did indeed arise by NAHR. As well, future in-depth analysis of sex chromosomes from the six three-generation mouse lines could investigate the role of oogenesis versus spermatogenesis and decipher whether there is a variable role in male or female gametogenesis on the rate at which heterozygosity affects mutagenesis leading to CNV formation.

Conclusion

Evidence of elevated mutagenesis leading to the formation of CNVs near heterozygous SNP loci was found in moderately heterozygous wild-derived mice, but not high heterozygosity F1 mice, indicating that the difference in heterozygosity must exist pre-zygotically during gametogenesis for a mutagenic effect to be observed. Evidence supporting a heterozygosity-associated, pre-zygotic mechanism of mutagenesis was provided by the observed proximity of *de novo* CNVs in the F2 mice to the heterozygous SNP loci in the genomic landscape of their F1 parents. It may be concluded that heterozygosity affects the local number of *de novo* CNVs formed in mice pre-zygotically. These findings are consistent with an hypothesized mechanism by which heterozygosity could act as a direct mutagen by reducing the capacity for successful allelic strand invasion to repair SPO11-induced DSBs generated during meiotic recombination, promoting NAHR with LCRs or other repetitive sequences adjacent to DSBs that result in large duplications or deletions of genetic material. In the future, whole genome sequencing with high read depth of the six three-generation mouse lines could be used to validate *de novo* CNVs, characterize the genomic features surrounding their breakpoints and assess whether repetitive elements such as LCRs are indeed required to promote mutageneis. As well, future studies may seek to determine with more precision the underlying likelihood of heterozygosity inducing genomic instability, what other genetic factors are required for CNV formation to occur, and if this phenomenon is pervasive beyond plants and mice. Heterozygosity is an endogenous mutagen or mutagen-associated genetic factor that should be considered for understanding the risk of genetic disease development and a fuller appreciation of the genetic diversity contributing to the process of evolution.

Bibliography

[1] Yang, S. *et al.* Parent–progeny sequencing indicates higher mutation rates in heterozygotes. *Nature* **523**, 463–467 (2015).

[2] Lupski, J. R. Structural variation mutagenesis of the human genome: Impact on disease and evolution. *Environmental and Molecular Mutagenesis* **56**, 419–436 (2015).

[3] Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nature Reviews Genetics* **7**, 85–97 (2006).

[4] Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nature Reviews Genetics* **10**, 551–564 (2009).

[5] Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nature Reviews Genetics* **12**, 363–376 (2011).

[6] Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. *Nature Reviews Genetics* **16**, 172–183 (2015).

[7] Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nature Reviews Genetics* **13**, 565–575 (2012).

[8] Kloosterman, W. P. *et al.* Characteristics of de novo structural changes in the human genome. *Genome Research* **25**, 792–801 (2015).

[9] Campbell, C. D. & Eichler, E. E. Properties and rates of germline mutations in humans. *Trends in Genetics* **29**, 575–584 (2013).

[10] Cheng, Z. *et al.* A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**, 88–93 (2005).

[11] Takumi, T. & Tamada, K. CNV biology in neurodevelopmental disorders. *Current Opinion in Neurobiology* **48**, 183–192 (2018).

[12] Marshall, C. R. & Scherer, S. W. Detection and characterization of copy number variation in autism spectrum disorder. *Methods in Molecular Biology* **838**, 115–135 (2012).

[13] Marshall, C. R. *et al.* Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nature Genetics* **49**, 27–35 (2017).

[14] Helbig, I. *et al.* 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. *Nature Genetics* **41**, 160–162 (2009).

[15] Shao, L. *et al.* Identification of chromosome abnormalities in subtelomeric regions by microarray analysis: A study of 5,380 cases. *American Journal of Medical Genetics, Part A* **146**, 2242–2251 (2008).

[16] Rovelet-Lecrux, A. *et al.* APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nature Genetics* **38**, 24–26 (2006).

[17] Chartier-Harlin, M. C. *et al.* α-synuclein locus duplication as a cause of familial Parkinson's disease. *Lancet* **364**, 1167–1169 (2004).

[18] Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434–1440 (2005).

[19] Sandegren, L. & Andersson, D. I. Bacterial gene amplification: Implications for the evolution of antibiotic resistance. *Nature Reviews Microbiology* **7**, 578–588 (2009).

[20] Maron, L. G. *et al.* Aluminum tolerance in maize is associated with higher MATE1 gene copy number. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 5241–5246 (2013).

[21] Patterson, E. L., Pettinga, D. J., Ravet, K., Neve, P. & Gaines, T. A. Glyphosate Resistance and EPSPS Gene Duplication: Convergent Evolution in Multiple Plant Species. *Journal of Heredity* **109**, 117–125 (2018).

[22] Sohrabi, S. S., Mohammadabadi, M., Wu, D. D. & Esmailizadeh, A. Detection of breed-specific copy number variations in domestic chicken genome. *Genome* **61**, 7–14 (2018).

[23] Revilla, M. *et al.* A global analysis of CNVs in swine using whole genome sequence data and association analysis with fatty acid composition and growth traits. *PLoS ONE* **12**, e0177014 (2017).

[24] Da Silva, V. H. *et al.* Genome-wide detection of CNVs and their association with meat tenderness in Nelore cattle. *PLoS ONE* **11**, e0157711 (2016).

[25] Gao, Y. *et al.* CNV discovery for milk composition traits in dairy cattle using whole genome resequencing. *BMC Genomics* **18**, 1–12 (2017).

[26] Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).

[27] Itsara, A. *et al.* De novo rates and selection of large copy number variation. *Genome Research* **20**, 1469–1481 (2010).

[28] Nguyen, D.-Q., Webber, C. & Ponting, C. P. Bias of Selection on Human Copy-Number Variants. *PLoS Genetics* **2**, e20 (2006).

[29] Guryev, V. *et al.* Distribution and functional impact of DNA copy number variation in the rat. *Nature Genetics* **40**, 538–545 (2008).

[30] Stankiewicz, P. & Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends in Genetics* **18**, 74–82 (2002).

[31] Chen, L., Zhou, W., Zhang, L. & Zhang, F. Genome Architecture and Its Roles in Human Copy Number Variation. *Genomics & Informatics* **12**, 136 (2014).

[32] Shaffer, L. G. & Lupski, J. R. Molecular Mechanisms for Constitutional Chromosomal Rearrangements in Humans. *Annual Review of Genetics* **34**, 297–329 (2000).

[33] Shaw, C. J. & Lupski, J. R. Non-recurrent 17p11.2 deletions are generated by homologous and non-homologous mechanisms. *Human Genetics* **116**, 1–7 (2005).

[34] Westemeier, R. L. *et al.* Tracking the long-term decline and recovery of an isolated population. *Science* **282**, 1695–1698 (1998).

[35] Markert, J. A. *et al.* Population genetic diversity and fitness in multiple environments. *BMC Evolutionary Biology* **10**, 205 (2010).

[36] Spielman, D., Brook, B. W. & Frankham, R. Most species are not driven to extinction before genetic factors impact them. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 15261–15264 (2004).

[37] Wang, D. G. *et al.* Large-scale identification, mapping, and genotyping of single- nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).

[38] Doran, A. G. *et al.* Deep genome sequencing and variation analysis of 13 inbred mouse strains defines candidate phenotypic alleles, private variation and homozygous truncating mutations. *Genome Biology* **17**, 167 (2016).

[39] Luzzatto, L. Sickle cell anaemia and malaria. *Mediterranean Journal of Hematology and Infectious Diseases* **4** (2012).

[40] Xie, Z. *et al.* Mutation rate analysis via parent– progeny sequencing of the perennial peach. I. A low rate in woody perennials and a higher mutagenicity in hybrids. *Proceedings of the Royal Society B: Biological Sciences* **283** (2016).

[41] Boateng, K. A., Bellani, M. A., Gregoretti, I. V., Pratto, F. & Camerini-Otero, R. D. Homologous Pairing Preceding SPO11-Mediated Double-Strand Breaks in Mice. *Developmental Cell* **24**, 196–205 (2013).

[42] Lieber, M. R., Ma, Y., Pannicke, U. & Schwarz, K. Mechanism and regulation of human non-homologous DNA end-joining. *Nature Reviews Molecular Cell Biology* **4**, 712–720 (2003).

[43] Lieber, M. R. & Karanjawala, Z. E. Ageing, repetitive genomes and DNA damage. *Nature Reviews Molecular Cell Biology* **5**, 69–75 (2004).

[44] Martin, G. M., Smith, A. C., Ketterer, D. J., Ogburn, C. E. & Disteche, C. M. Increased chromosomal aberrations in first metaphases of cells isolated from the kidneys of aged mice. *Israel journal of medical sciences* **21**, 296–301 (1985).

[45] Symington, L. S. & Gautier, J. Double-Strand Break End Resection and Repair Pathway Choice. *Annual Review of Genetics* **45**, 247–271 (2011).

[46] Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).

[47] Mao, Z., Bozzella, M., Seluanov, A. & Gorbunova, V. Comparison of nonhomologous end joining and homologous recombination in human cells. *DNA Repair* **7**, 1765–1771 (2008).

[48] Downs, J. A. & Jackson, S. P. A means to a DNA end: The many roles of Ku. *Nature Reviews Molecular Cell Biology* **5**, 367–378 (2004).

[49] Grawunder, U., Zimmer, D., Fugmann, S., Schwarz, K. & Lieber, M. R. DNA ligase IV is essential for V(D)J recombination and DNA double-strand break repair in human precursor lymphocytes. *Molecular Cell* **2**, 477–484 (1998).

[50] Li, Z. *et al.* The XRCC4 gene encodes a novel protein involved in DNA double-strand break repair and V(D)J recombination. *Cell* **83**, 1079–1089 (1995).

[51] Conover, H. N. & Argueso, J. L. Contrasting mechanisms of de novo copy number mutagenesis suggest the existence of different classes of environmental copy number mutagens. *Environmental and Molecular Mutagenesis* **57**, 3–9 (2016).

[52] Symington, L. S. End resection at double-strand breaks: Mechanism and regulation. *Cold Spring Harbor Perspectives in Biology* **6** (2014).

[53] Jasin, M. & Rothstein, R. Repair of strand breaks by homologous recombination. *Cold Spring Harbor Perspectives in Biology* **5** (2013).

[54] Nassif, N., Penney, J., Pal, S., Engels, W. R. & Gloor, G. B. Efficient copying of non-homologous sequences from ectopic sites via P-element-induced gap repair. *Molecular and Cellular Biology* **14**, 1613–1625 (1994).

[55] Pâques, F. & Haber, J. E. Multiple Pathways of Recombination Induced by Double-Strand Breaks in Saccharomyces cerevisiae. *Microbiology and Molecular Biology Reviews* **63**, 349–404 (1999).

[56] Dittwald, P. *et al.* NAHR-mediated copy-number variants in a clinical population: Mechanistic insights into both genomic disorders and Mendelizing traits. *Genome Research* **23**, 1395–1409 (2013).

[57] Frank-Vaillant, M. & Marcand, S. Transient stability of DNA ends allows nonhomologous end joining to precede homologous recombination. *Molecular Cell* **10**, 1189–1199 (2002).

[58] Haenel, Q., Laurentino, T. G., Roesti, M. & Berner, D. Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. *Molecular Ecology* **27**, 2477–2497 (2018).

[59] Sasaki, M., Lange, J. & Keeney, S. Genome destabilization by homologous recombination in the germ line. *Nature Reviews Molecular Cell Biology* **11**, 182–195 (2010).

[60] Peng, Z. *et al.* Correlation between frequency of non-allelic homologous recombination and homology properties: evidence from homology-mediated CNV mutations in the human genome. *Human molecular genetics* **24**, 1225–33 (2015).

[61] Morse, H. C. Building a Better Mouse: One Hundred Years of Genetics and Biology. In James G. Fox *et al.* (eds.) *The Mouse in Biomedical Research*, vol. 1, chap. 1, 1–11 (Academic Press, 2007), 2 edn.

[62] Davisson, M. T. Rules and guidelines for genetic nomenclature in mice: Excerpted version. *Transgenic Research* **6**, 309–319 (1997).

[63] Beck, J. A. *et al.* Genealogies of mouse inbred strains. *Nature Genetics* **24**, 23–25 (2000).

[64] Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).

[65] Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).

[66] Lilue, J. *et al.* Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nature Genetics* **50**, 1574–1583 (2018).

[67] Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genetics* **40**, 1253–1260 (2008).

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

[69] Matukumalli, L. K. *et al.* Development and Characterization of a High Density SNP Genotyping Assay for Cattle. *PLoS ONE* **4**, e5350 (2009).

[70] Kranis, A. *et al.* Development of a high density 600K SNP genotyping array for chicken. *BMC Genomics* **14**, 1–13 (2013).

[71] Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME) - Toward standards for microarray data. *Nature Genetics* **29**, 365–371 (2001).

[72] Jakubek, Y. A. & Cutler, D. J. A model of binding on DNA microarrays: Understanding the combined effect of probe synthesis failure, cross-hybridization, DNA fragmentation and other experimental details of affymetrix arrays. *BMC Genomics* **13**, 737 (2012).

[73] Komura, D. *et al.* Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Research* **16**, 1575–1584 (2006).

[74] Lin, C.-F., Naj, A. C. & Wang, L.-S. Analyzing Copy Number Variation Using SNP Array Data: Protocols for Calling CNV and Association Tests. *Current Protocols in Human Genetics* **79**, 1–17 (2013).

[75] Locke, M. E. O. *et al.* Genomic copy number variation in Mus musculus. *BMC Genomics* **16**, 497 (2015).

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

[77] Eitutis, S. Array-based genomic diversity measures portray Mus musculus phylogenetic and genealogical relationships, and detect genetic variation among C57Bl/6J mice and between tissues of the same mouse. *The University of Western Ontario Electronic Thesis and Dissertation Repository* (2013).

[78] The Jackson Laboratory Center for Genome Dynamics - CEL Files Available at: ftp://ftp.jax.org/petrs/MDA/.

[79] Domanska, D. *et al.* The rainfall plot: Its motivation, characteristics and pitfalls. *BMC Bioinformatics* **18**, 264 (2017).

[80] Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).

[81] Luo, B. *et al.* Spatial statistical tools for genome-wide mutation cluster detection under a microarray probe sampling system. *PLOS ONE* **13**, e0204156 (2018).

[82] Luo, B., Boehler, N. & Villani, S. Spatial statistical analyis pipeline (N.D.). Unpublished.

[83] Foxall, R. & Baddeley, A. Nonparametric measures of association between a spatial point process and a random set, with geological applications. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **51**, 165–182 (2002).

[84] Luo, B. *Statistical tools for assessment of spatial properties of mutations observed under the the microarray platform*. Ph.D. thesis, The University of Western Ontario (2018).

[85] Becirovic-Agic, M., Jönsson, S. & Hultström, M. Quantitative trait loci associated with angiotensin II and high-salt diet induced acute decompensated heart failure in Balb/CJ mice. *Physiological Genomics* **51**, 279–289 (2019).

[86] Wang, K. *et al.* PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* **17**, 1665–1674 (2007).

[87] Thermo Fisher Scientific. Affymetrix Power Tools. Available at: https://www.thermofisher.com/ca/en/home/life-science/microarray-analysis/

microarray-analysis-partners-programs/affymetrix-developers-network/affymetrix-power-tools.html.

[88] PennAffy. Available at: http://www.openbioinformatics.org/penncnv/penncnv_download.html.

[89] PennCNV. PennCNV-Affy: Additional topics. Available at: http://penncnv.openbioinformatics.org/en/latest/user-guide/affy/ (2017).

[90] Bastian, M., Heymann, S. & Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. *International AAAI Conference on Weblogs and Social Media,* Conference paper (2009).

[91] Rau, C. D. *et al.* High-density genotypes of inbred mouse strains: Improved power and precision of association mapping. *G3: Genes, Genomes, Genetics* **5**, 2021–2026 (2015).

[92] Bianco, L. *et al.* Development and validation of the Axiom® Apple480K SNP genotyping array. *Plant Journal* **86**, 62–74 (2016).

[93] Roorkiwal, M. *et al.* Development and evaluation of high-density Axiom® *CicerSNP* Array for high-resolution genetic mapping and breeding applications in chickpea. *Plant Biotechnology Journal* **16**, 890–901 (2018).

[94] Cutler, G. & Kassner, P. D. Copy number variation in the mouse genome: Implications for the mouse as a model organism for human disease (2009).

[95] Salomon-Torres, R., Villa-Angulo, R. & Villa-Angulo, C. Analysis of copy number variations in Mexican Holstein cattle using axiom genome-wide Bos 1 array. *Genomics Data* **7**, 97–100 (2016).

[96] Bai, H. *et al.* Genome-wide detection of CNVs associated with beak deformity in chickens using high-density 600K SNP arrays. *Animal Genetics* **49**, 226–236 (2018).

[97] Li, R. *et al.* A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. *Nature Communications* **10**, 1–15 (2019).

[98] Baudat, F. & De Massy, B. Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis (2007).

[99] Li, R. *et al.* A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. *Nature Communications* **10**, 1–15 (2019). URL https://doi.org/10.1038/s41467-019-11675-y.

[100] Valencia, M. *et al.* NEJ1 controls non-homologous end joining in Saccharomyces cerevisiae. *Nature* **414**, 666–669 (2001).

[101] Joyce, E. F., Paul, A., Chen, K. E., Tanneti, N. & McKim, K. S. Multiple barriers to nonhomologous DNA end joining during meiosis in Drosophila. *Genetics* **191**, 739–746 (2012).

[102] Humphryes, N. & Hochwagen, A. A non-sister act: Recombination template choice during meiosis (2014).

[103] Zickler, D. & Kleckner, N. Recombination, pairing, and synapsis of homologs during meiosis. *Cold Spring Harbor Perspectives in Biology* **7**, 1–28 (2015).

[104] Finsterbusch, F. *et al.* Alignment of Homologous Chromosomes and Effective Repair of Programmed DNA Double-Strand Breaks during Mouse Meiosis Require the Minichromosome Maintenance Domain Containing 2 (MCMDC2) Protein. *PLoS Genetics* **12** (2016).

[105] Youds, J. L. & Boulton, S. J. The choice in meiosis - Defining the factors that influence crossover or non-crossover formation (2011).

[106] Drouaud, J. *et al.* Contrasted Patterns of Crossover and Non-crossover at Arabidopsis thaliana Meiotic Recombination Hotspots. *PLoS Genetics* **9**, e1003922 (2013).

[107] Gray, S. & Cohen, P. E. Control of Meiotic Crossovers: From Double-Strand Break Formation to Designation (2016).

[108] Waldman, A. S. & Liskay, R. M. Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology. *Molecular and Cellular Biology* **8**, 5350–5357 (1988).

[109] Xu, L., Hou, Y., Bickhart, D., Song, J. & Liu, G. Comparative Analysis of CNV Calling Algorithms: Literature Survey and a Case Study Using Bovine High-Density SNP Data. *Microarrays* **2**, 171–185 (2013).

[110] Rafter, P., Gormley, I. C., Parnell, A. C., Kearney, J. F. & Berry, D. P. Concordance rate between copy number variants detected using either high- or medium-density single nucleotide polymorphism genotype panels and the potential of imputing copy number variants from flanking high density single nucleotide polymorphism haplotypes in cattle. *BMC Genomics* **21**, 205 (2020).

Appendix A

Supplementary figures and tables

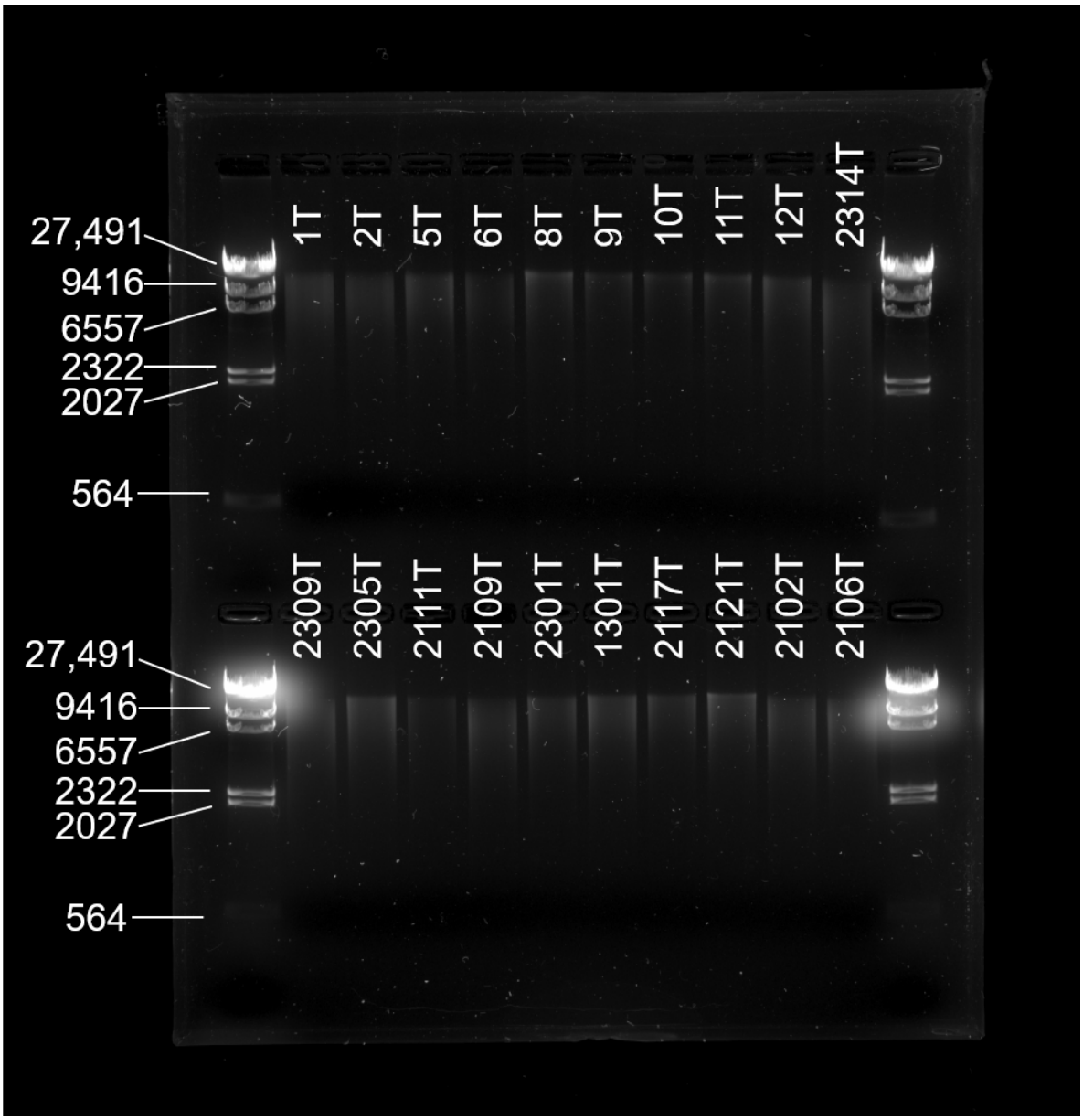

Figure A.1: DNA degradation evaluated by agarose gel electrophoresis for extracted genomic DNA from six three-generation mouse lines, gel 1 Hind III DNA ladder is in columns 1, 12, 13, and 14 where columns 1-12 are the top row and 13-24 are the bottom row. Columns 2-11 and 14-23 contain genomic DNA and are labeled according to sample identifier from six three-generation mouse lines. Gel is 0.8% agarose, run at 100V for 75 minutes in TBE 0.5X buffer and stained with SYBR Safe DNA gel stain.

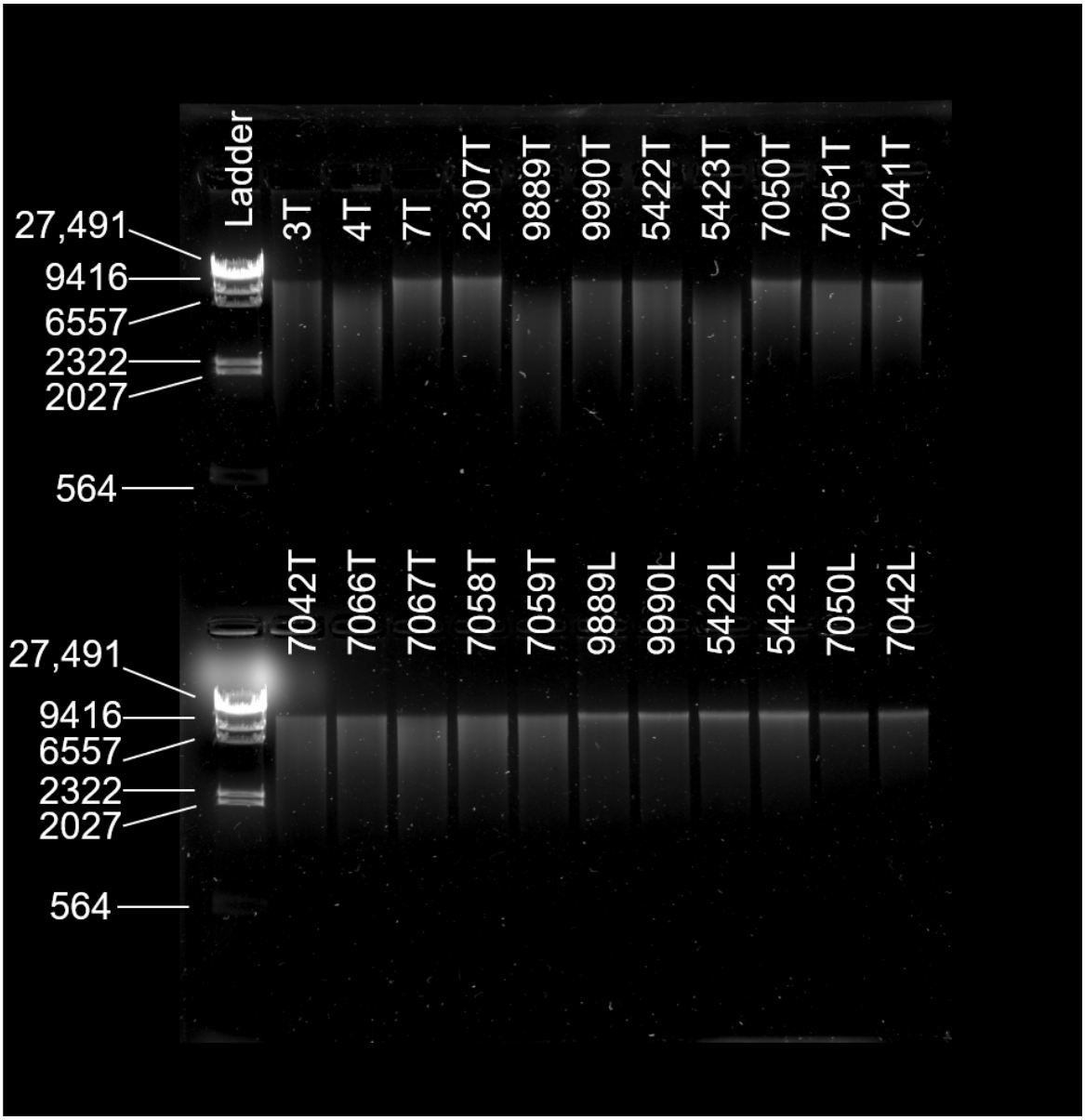

Figure A.2: DNA degradation evaluated by agarose gel electrophoresis for extracted genomic DNA from six three-generation mouse lines, gel 2 Hind III DNA ladder is in columns 1 and 13, where columns 1-12 are the top row and 13-24 are the bottom row. Columns 2-12 and 14-24 contain genomic DNA and are labeled according to sample identifier from six three-generation mouse lines. Gel is 0.8% agarose, run at 130V for 50 minutes in TBE 0.5X buffer and stained with SYBR Safe DNA gel stain.

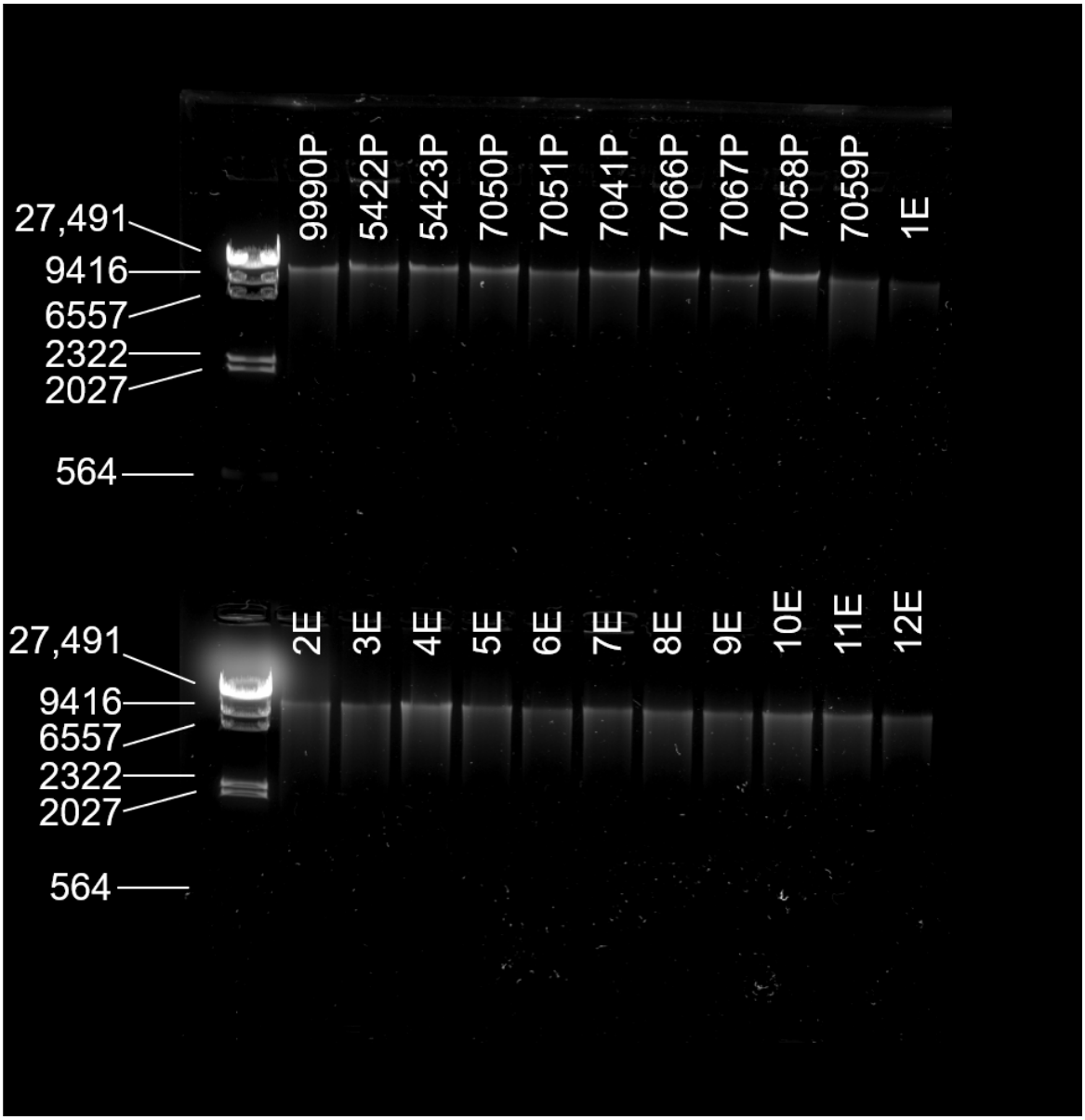

Figure A.3: DNA degradation evaluated by agarose gel electrophoresis for extracted genomic DNA from six three-generation mouse lines, gel 3 Hind III DNA ladder is in columns 1 and 13, where columns 1-12 are the top row and 13-24 are the bottom row. Columns 2-12 and 14-24 contain genomic DNA and are labeled according to sample identifier from six three-generation mouse lines. Gel is 0.8% agarose, run at 130V for 55 minutes in TBE 0.5X buffer and stained with SYBR Safe DNA gel stain.

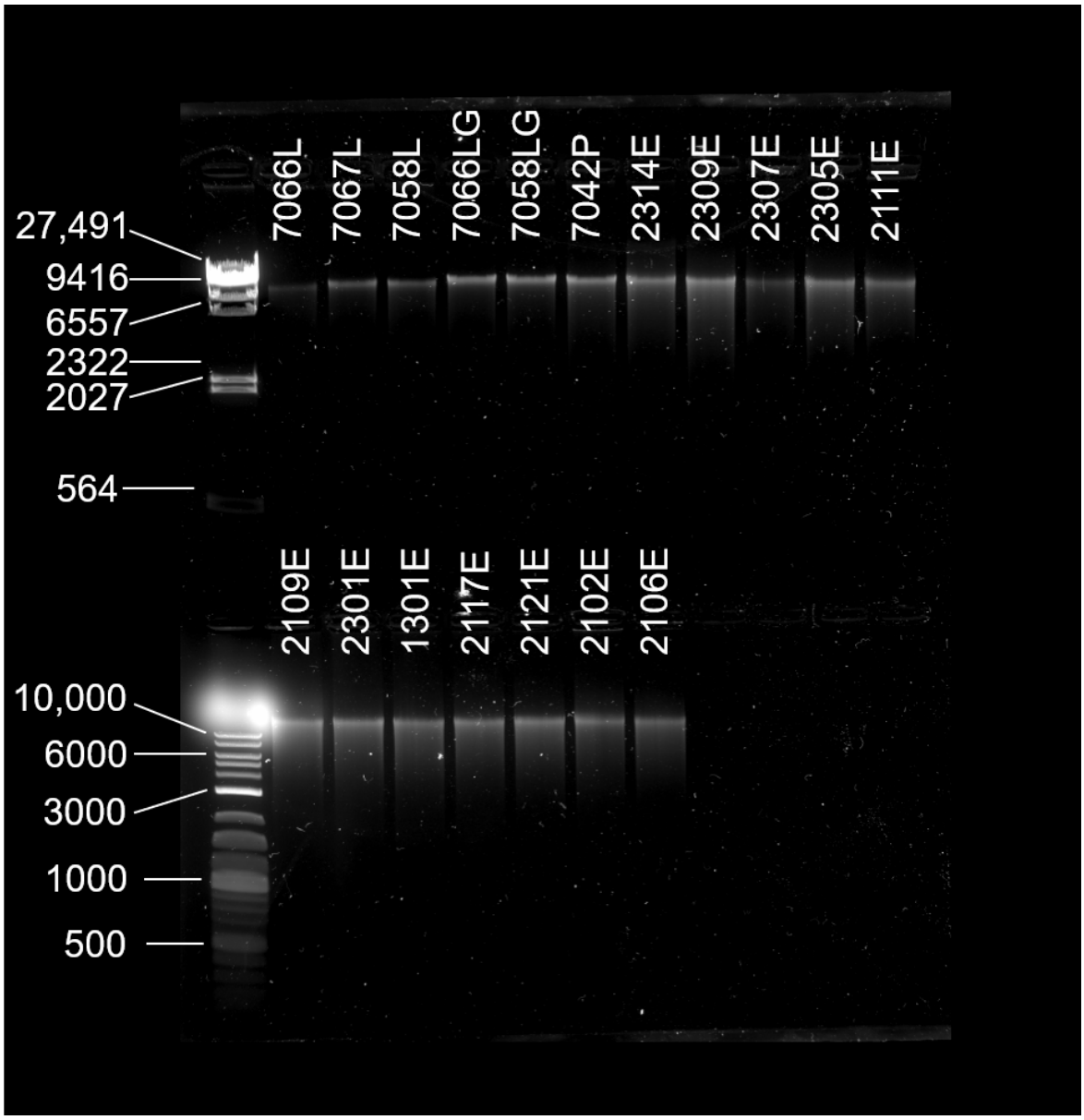

Figure A.4: DNA degradation evaluated by agarose gel electrophoresis for extracted genomic DNA from six three-generation mouse lines, gel 4 Hind III DNA ladder is in column 1 and Quick-Load Purple 1 Kb Plus ladder is in column 13, where columns 1-12 are the top row and 13-24 are the bottom row. Columns 2-12 and 14-24 contain genomic DNA and are labeled according to sample identifier from six three-generation mouse lines. Gel is 0.8% agarose, run at 140V for 50 minutes in TBE 0.5X buffer and stained with SYBR Safe DNA gel stain.

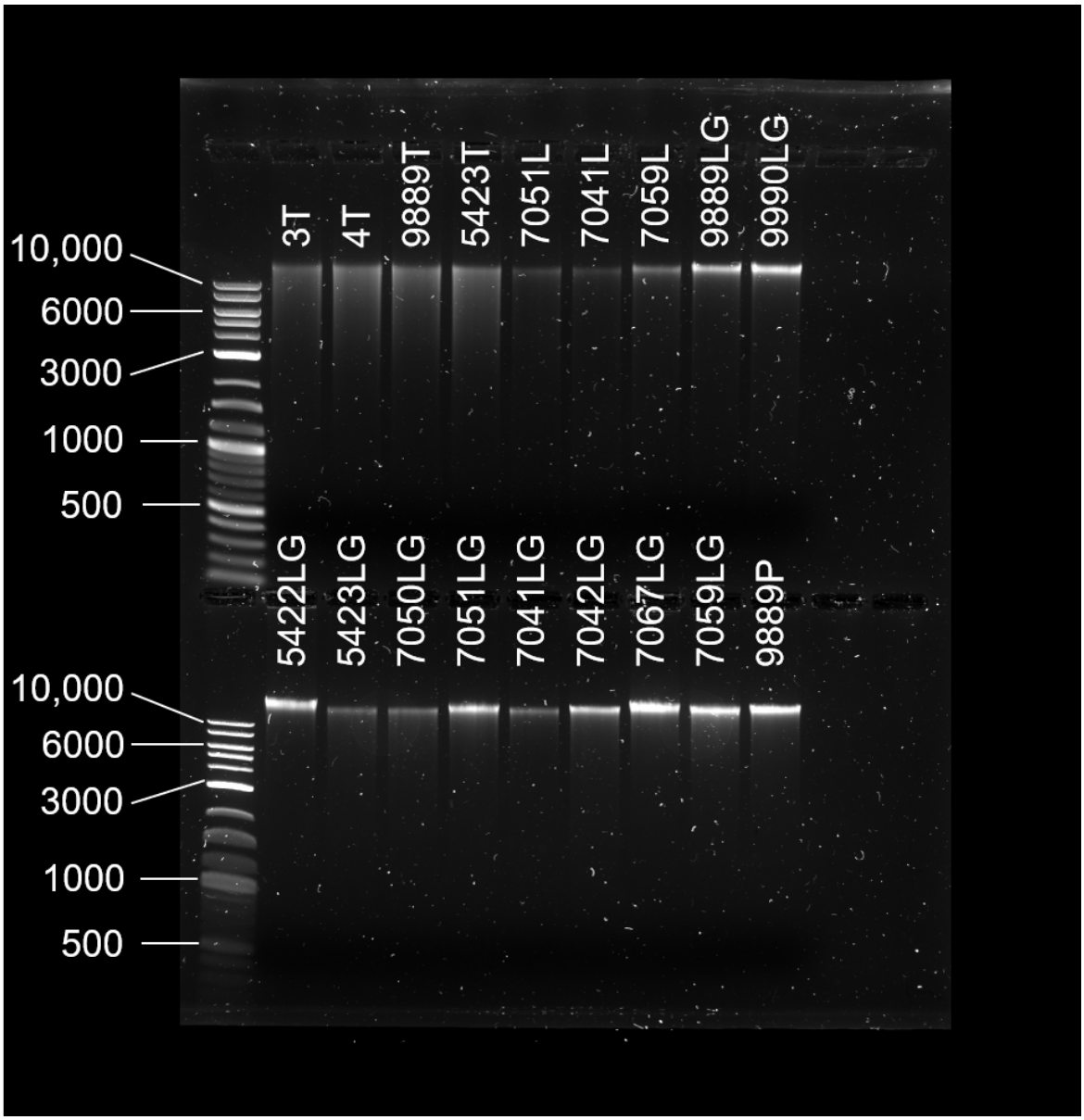

Figure A.5: DNA degradation evaluated by agarose gel electrophoresis for extracted genomic DNA from six three-generation mouse lines, gel 5 Quick-Load Purple 1 Kb Plus ladder is in columns 1 and 13, where columns 1-12 are the top row and 13-24 are the bottom row. Columns 2-12 and 14-24 contain genomic DNA and are labeled according to sample identifier from six three-generation mouse lines. Gel is 0.8% agarose, run at 122V for 60 minutes in TBE 0.5X buffer and stained with SYBR Safe DNA gel stain.

Appendix B

Online supplementary material

This is a clickable hyperlink to my online supplementary material in OneDrive

Name:	Nicholas Boehler
Post-Secondary Education and Degrees:	University of Western Ontario London, Ontario, Canada 2011-2017 Hons. B.Sc
Honours and Awards:	Queen Elizabeth II Graduate Scholarship for Science and Technology 2018-2019
	NSERC Undergraduate Research Award 2017
	Dean's Honour List 2015, 2016, 2017
	The Western Scholarship of Excellence 2011
Related Work Experience:	Teaching Assistant The University of Western Ontario 2017-2020

Peer-reviewed Published Abstract Publications:

Boehler N, Luo B, Milojevic M, Pavanel H, Dean CB, Kulperger R, Hill KA. Heterozygosity: A meiosis-linked intrinsic mutagen in mice. International conference at *Environmental Mutagenesis and Genomic Society*, Virtual annual meeting, (2020).

Boehler N, Luo B, Milojevic M, Locke MEO, Dean CB, Kulperger R, Hill KA. Heterozygosity: An underappreciated meiosis-linked intrinsic mutagen in mice. International conference at *Environmental Mutagenesis and Genomic Society*, San Antonio, TX, USA, (2018).

Boehler N, Qi FW, Luo B, Milojevic M, Locke MEO, Tolg C, Turley E, Dean CB, Kulperger R, Hill KA. Application of genomic spatial statistic tools to evaluate higher mutation rates associated with regions of heterozygosity. International conference at *Environmental Mutagenesis and Genomic Society*, Raleigh, NC, USA (2017).

Oral Presentations:

Heterozygosity: A meiosis-linked intrinsic mutagen in mice. International conference presenting as a talk for the Environmental Mutagenesis and Genomics Society at their Virtual Annual Meeting, Sept 16, 2020.

Heterozygosity: An underappreciated meiosis-linked mutagen in mice. International conference presented as a flash talk for the Environmental Mutagenesis and Genomics Society at the Hyatt Regency, San Antonio, TX, USA, Sept 23, 2018.

Mutant or variant? Perspective is paramount. Presented as a lightning talk for the 3 Minute Thesis competition at the University of Western Ontario, London, ON, Canada, Feb 13, 2018.

Elevated mutation rates spatially linked to heterozygosity in mice. Presented as a lightning talk for Biology Graduate Research Forum at the University of Western Ontario, London, ON, Canada, Oct 20, 2017.

Elevated mutation rate associated with heterozygosity in mice: Do opposites attract mutations? Presented at Ontario Biology Day at Laurentian University. Sudbury, ON, Canada, Mar 18, 2017.

Poster Presentations:

Boehler, N, Luo B, Milojevic M, Locke MEO, Dean, CB, Kulperger, R, Hill KA. *Heterozygosity: An underappreciated meiosis-linked intrinsic mutagen in mice.* International conference at Environmental Mutagenesis and Genomic Society, San Antonio, TX, USA, Sept 23, 2018.

Boehler, N, Qi FW, Luo B, Milojevic M, Locke MEO, Tolg C, Turley E, Dean CB, Kulperger R, Hill KA. *Application of genomic spatial statistic tools to evaluate higher mutation rates associated with regions of heterozygosity.* International conference at Environmental Mutagenesis and Genomic Society, Raleigh, NC, USA, Sept 12, 2018.

Boehler, N, Qi, FW, Luo B, Milojevic M, Locke MEO, Tolg C, Turley E, Dean CB, Kulperger R, Hill KA. *Spatial statistical tools for mutation cluster detection tested using the case of F1 heterozygosity and a cancer model.* Presented at Oncology Research and Education day at Best Western Lamplighter Inn, London, ON, Canada, June 16, 2017.