4-15-2014 12:00 AM

Somatic Copy Number Mosaicism Contributes to Genomic Diversity in Mus musculus

Andrea E. Wishart
The University of Western Ontario

Supervisor
Dr Kathleen A Hill
The University of Western Ontario

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

SOMATIC COPY NUMBER MOSAICISM CONTRIBUTES TO GENOMIC DIVERSITY IN *MUS MUSCULUS*

(Thesis format: Monograph)

by

Andrea E.Wishart

Graduate Program in Biology

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

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

Abstract

Copy number variants (CNVs) are a source of genomic variation associated with altered phenotypes. Somatic copy number mosaicism results when different populations of cells in an individual differ due to *de novo* copy number changes (CNCs). Tissue-specific patterns of CNCs resulting in mosaicism have yet to be characterized in the mouse, an organism frequently used to model human diseases. Here, DNA was sampled from spleen, liver, and cerebellum of eight highly related mice selected from a familial unit. CNVs and CNCs were detected using the Mouse Diversity Genotyping Array with three computational methods (ConsecN, Partek, and PennCNV). Tissue-specific patterns of CNCs were revealed, including genomic regions of putative recurring CNCs. Genetic distance estimated using CNVs and CNCs recapitulated genealogical relationships. The novel framework can thus be used to identify and analyze tissue-specific CNCs, and the results establish the need to account for CNCs in model organisms.

Keywords: copy number variation, copy number change, copy number variable region, mutation, single nucleotide polymorphism, genotyping, genetic distance, genetic variation, genotyping array, structural variation, somatic mosaicism, spleen, cerebellum, liver, *Mus musculus*, Mouse Diversity Genotyping Array

Co-Authorship Statement

Andrea E. Wishart performed the following work under the supervision and financial support of Dr Kathleen A. Hill. Andrea Wishart performed the experimental work presented in this thesis. Dr. Kathleen Hill will be a co-author on all publications stemming from this work for her involvement in experimental design and review. Susan T. Eitutis and Marjorie E. Osborn Locke may also be included in future papers related to contributions of filtered SNP probe lists, and PennCNV pipeline for analyses, respectively.

Acknowledgements

For a long time, the placeholder text for this section simply read “Thank you Starbucks and Bombay Sapphire.” Now that I have completed this thesis, it is time to pay proper tribute to all of the people who have supported me throughout this project.

First and foremost I must thank my supervisor, Dr Kathleen Hill, for her dedicated mentorship and support over the past four years. Thank you for seeing potential in me and having faith that I would succeed in an ambitious project, while also giving me the opportunity to participate in as many scholarly pursuits as possible, including teaching, side projects, and conferences. I have been incredibly fortunate to work with supervisor who invests so much time into each of her students.

Thank you to all of the faculty and staff here at Western for their advice, constructive criticism, and technical help along the way. Special thanks to my advisors, Drs Graham Thompson and Bryan Neff, for invaluable guidance and clear focus. Dr Shiva Singh, for expertise on copy number variation, and for emphasizing the importance of looking at the variation around the mean in all results. Dr Andre Lachance, for suggestions about comparing tree topologies. Dr Ben Rubin, for statistics assistance both on this and other projects, and for normalizing R usage in the department (stats pun intended). Dr Mark Daley, for a critical eye for methods and data, and for reminding me that there are many routes to and definitions of success. Jenna Butler and Beth Locke, for teaching me so much in the realm of computer science and for technical help with the bioinformatics of this project. David Carter, for the efforts in getting the MDGA project rolling with Partek in the LRGC. Christina Castellani, for teaching me how to teach, and for your CNV guidance and humour.

Thank you to all of the Hill lab members, past and present. Special thanks to Anita Prtenjaca, for welcoming me onto the nocturnal behaviour project and waging the thermostat war with me; Alex Laliberte, for showing me the ropes and introducing me to Friday beer o’clock;

Matt Edwards, I miss our morning coffee routine from that first year, but we've kept it up best we could! Heather Garner, for moral support, exam feedback, and mini cheesecakes. Last but not least, Susan T. Eitutis, also known in the literature now as Susan E. Eitutis (I'm SO sorry) —you've been on the MDGA journey with me from the start. From that very first time we tried to open 623,124 rows of genotyping results in Excel, we've come such a long way and learned a lot together. I could never have done this without our teamwork and your salted caramel cookies.

Thank you to all my friends for doing their best to get me out on weekends but for being understanding when I couldn't. Thank you to the wonderful scientists, grad students, and writers I have met through Twitter for your 140 character bursts of support and encouragement, even though I have yet to meet many of you face to face. I would also like to acknowledge the support from the staff of Western's Student Development Center. Grad school can be tough, and I owe my success in part to those who sat down with me on a regular basis to help me organize my thoughts and goals to help me reach my potential.

Finally, thank you to my family. To my parents for their unwavering support, for always reminding me that I am more than my CV, and for putting such a high value on my education and happiness. Thanks to my little brother Brian for being a constant companion even with several provinces between, and almost as importantly, the Netflix password.

This research was supported by funding to Dr Kathleen Hill from the Natural Sciences and Engineering Research Council of Canada and the Canadian Institute of Health Research. Additional support was funded to Andrea Wishart from the Department of Biology as a Graduate Thesis Research Award.

Contents

List of Figures

List of Tables

List of Appendices

List of Abbreviations, Symbols, and Nomenclature

aCGH	Array comparative genomic hybridization
BAF	B-allele frequency
BEDdetail	browser extensible data detail [*file extension*]
bp	Base pairs
BIR	Break-induced replication
BRLMM-P	Bayesian Robust Linear Modeling with Mahalanobis distance (Perfect match)
.CEL	Cell intensity file [*file extension*]
.CDF	Chip definition file [*file extension*]
.chrvar	Chromosome variation file
.csv	comma separated values [*file extension*]
CGD	Center for Genome Dynamics
CGH	Comparative genomic hybridization
CL	Cerebellum
CNC	Copy number change
CNP	Copy number polymorphism
CNV	Copy number variant
CNVR	Copy number variable region
CRLMM	Corrected Robust Linear Modeling with Mahalanobis distance
DMBA	7,12-dimethylbenz[a]anthracene
DNA	Deoxyribonucleic acid
DSB	Double stranded breaks
dsDNA	Double-stranded DNA
FISH	Fluorescent *in situ* hybridization
GWAS	Genome-wide association study
HD-CNV	Hotspot Detector for Copy Number Variants
HMM	Hidden Markov Model

IGP	Invariant genomic probe
kb	Kilobases
LI	Liver
LRR	Log R ratio
Mb	Mega bases
MDGA	Mouse Diversity Genotyping Array
MM	Mismatch
MMBR	Microhomology-mediated break-induced replication
NAHR	Non-allelic homologous recombination
NCBI	National Center for Biotechnology Information
NHEJ	Non-homologous end joining
PCR	Polymerase chain reaction
PM	Perfect match
QC	Quality control
RMA	Robust multichip average
ROMA	Representational oligonucleotide microarray analysis
rpt	replicate (for tissues sampled from mouse 911.50)
SNP	Single nucleotide polymorphism
SP	Spleen
SV	Structural variation
ssDNA	single stranded DNA
.tsv	tab separated values [*file extention*]
UCSC	University of California Santa Cruz

Chapter 1

Introduction

1.1 General introduction

Genomic variability underlies many of the phenotypic differences between organisms. Over evolutionary time, mutations arise and can be passed on to future generations through the germline, introducing genomic variation among individuals in a population. Additionally, somatic mutations can occur within an individual that cause there to be genomic variation across different tissues, termed *somatic mosaicism*. Such mosaicism can have phenotypic effects, including genetic diseases that do not appear to have been inherited from either parent. Identifying such genomic alterations is necessary to understand the etiology of diseases that occur without a pre-existing family history, to assess the accuracy of studying DNA derived from accessible tissues (such as blood) as a proxy for studying less accessible tissues (such as the brain), and to understand mechanisms of tissue-specific rates of mutagenesis. Despite the popularity of using the laboratory mouse as a model of human disease, somatic mosaicism is not nearly as well studied in murine tissues. Identifying the types of mutations that contribute to somatic mosaicism in mice and determining if there are tissue-specific patterns in accumulation of mutations is crucial to our understanding of genomic variation in current mouse models of human diseases and phenotypes.

At the population level, major sources of genomic variation between individuals have been

identified. The completion of the Human Genome Project in 2001 ushered in the new era of genomics, the study of all of the genetic content of an organism in the nuclear and mitochondrial DNA (as well as chloroplast DNA in plants).[1] The successful compilation of the human reference sequence marked a major expansion to accessible biological information. Initially, the high degree of sequence similarity observed between individual humans was interpreted as evidence of minimal genetic variation within our species (any two human genomes share approximately 99.9% sequence identity).[2] The differences in sequence that were observed were primarily attributed to single nucleotide polymorphisms (SNPs). SNPs are single base pair positions in the genome at which the sequence can differ between healthy individuals, with the less common nucleotide, or allele, maintained in the population at a frequency of at least 1%.[3] It is estimated that there are approximately 10 million SNPs in the human genome distributed over all autosomes, sex chromosomes, and the mitochondrial genome.[1,4] As of 2007, the International HapMap Consortium had successfully genotyped over 3.1 million human SNPs.[5] The abundance of SNPs in the genome, as well as their potential for phenotypic significance, makes SNPs ideal genetic markers for genome-wide association studies (GWAS),[1] the first of which was performed on human SNP data, and uncovered an association between SNPs and macular degeneration.[6] Despite the large number of SNPs in the genome, researchers have found that SNPs alone are insufficient to be the sole source of genetic-based phenotypic variation, including genomic variation between individual tissues.

1.1.1 Copy number variants are a significant source of genetic variation that can impact phenotype

Just a few years after the initial drafts of the human genome sequence were published, a class of previously undetected genomic variants emerged: structural variants (SVs) larger than single nucleotide polymorphisms, yet smaller than chromosome aberrations observable under the microscope.[7–9] Despite the focus on SNPs as a major source of genetic variation, SVs were

found to affect a greater proportion of the genome than SNPs.[10] Grouped under this category of submicroscopic SVs are copy number variants (CNVs), first reported independently by two studies in 2004.[11,12] A CNV is characterized as a gain or loss of a segment of genomic DNA. This is in contrast to inversions and balanced translocations, in which the total amount of DNA present remains unchanged after the rearrangement.[13] CNVs are typically defined as structural variants 1 kb or larger, although continued improvements to the methods used to detect CNVs have led to the suggestion that SVs as small as 500 bp may be classified as CNVs,[14] and one review suggests the lower size boundary to be 50 bp.[15] In light of the recently published ENCODE project that reported on previously unknown functions of non-coding sequence,[16] the work presented here in this thesis defines a CNV based on length alone, regardless of the function of the sequence between its breakpoints. The term copy number polymorphism (CNP) specifically refers to a variant that occurs in a population at a frequency of 1% or higher.[12] Although the term CNP was originally used for all copy number variation, use of the term in reference to population frequency is now considered to be analogous to that of SNP, while the term copy number variant is now applied more widely.[8]

Despite the relatively short history of CNVs, they are already recognized as providing a significant source of genomic variation between individuals.[17,18] Up to 13% of any one human genome may be affected by copy number variation. CNVs may be contributing to genomic diversity among human populations that may previously have gone undetected at the sequence level.[19] Cumulatively, more nucleotides are affected by CNVs in the human genome than are affected by SNPs.[19] One study estimates that the genomes of two unrelated people may differ by approximately 0.4%,[20] as opposed to the 0.01% estimated by sequence identity alone.[2] To date, over 25,000 CNV loci have been identified in healthy human subjects.[21,22] CNVs, like any mutation, can be inherited.[19,23] This has been demonstrated in a family study that showed the majority of the CNVs detected in a child were inherited from either parent.[23] Also similar to other mutations, CNVs can form *de novo* within individuals. The frequency of *de novo* CNV formation is especially interesting not only because of the large amount of genetic material

involved in CNVs, but also because the frequency of *de novo* CNV formation is higher than the frequency of point mutations.[24] In humans, the rate of *de novo* CNV formation is estimated to be 1.2×10^{-2} events per transmission per generation,[25] while the rate of point mutations is estimated to be 2.5×10^{-8} events per transmission per generation.[26]

The bulk of association studies between CNVs and disease has focused on human subjects, but CNVs are by no means limited to *Homo sapiens*. That CNVs are found in our close primate relatives, including chimpanzees and rhesus macaques,[27] offers the opportunity to peer into our own evolutionary history. CNVs are present in the genomes of other mammals including rats,[28] dogs,[29] cattle,[30] sheep,[31] goats,[32] and pigs.[33,34] Other classes of animals have been found to have CNVs, such as class Aves (including chickens,[35,36] turkeys,[35] and Pekin ducks[37]) and class Insecta (most notably *Drosophila*).[38,39] CNVs have been catalogued in barley[40] and *Saccharomyces cerevisiae*.[41] The discovery of CNVs across kingdoms demonstrates that CNVs are pervasive sources of genomic variation in the tree of life and can be examined in many divergent lineages.

Of all the species in which CNVs have been identified, the laboratory mouse (*Mus musculus domesticus*) is a popular model organism for studying mechanisms of human disease. We have an extraordinary amount of data available pertaining to the biology of the mouse, including very specific genetic and phenotypic characterizations of the inbred strains that have been developed and maintained over many generations.[42] Mice and humans are both species of eutherian mammals that diverged from each other approximately 75 million years ago, and currently share approximately 85% sequence identity in coding regions.[43] Long used to model human phenotypes and genetics, mice are an ideal alternative to humans in research for a number of reasons. Compared to the generation time of humans (estimates range from 20 to 35 years),[44,45] mice have a very short generation time (about 12 weeks),[46] which allows researchers to quickly obtain several generations of one strain for a given study. Researchers are able to control both environment and genetic background to minimize confounding variables that would normally exist if using human subjects. Control of genetic background through selective breeding pro-

grams has led to a wide availability of specialized strains and transgenic models with specific traits. Access to all tissues of interest in mice increases the variety and specificity of studies that can be designed. Additionally, our knowledge of murine genetics and breeding histories is extensive. The initial draft sequence of the laboratory mouse genome was assembled from female C57BL/6J mice in 2002, soon after the first draft of the human genome.[43] From this information, similarities between the mouse and human genomes have been identified. For example, the GC content of the mouse genome is about 42%, slightly higher than the GC content of the human genome (41%).[43] Databases cataloguing all known SNPs in the mouse genome for most inbred strains provide researchers with easy access to information useful in association studies. The Center for Genome Dynamics SNP Database (CGDSNPdb) is a large database of murine SNP information that can be queried on a number of factors such as strain or genomic position.[47] All of these factors contribute to the mouse maintaining an invaluable position as an ideal model organism in which to research mutational mechanisms across multiple mammalian tissues that would otherwise be inaccessible in humans, while controlling for genetic background and environment. Given the popularity of using the mouse as a model of human disease, extensive efforts have been made to map CNVs in the mouse genome.[48–50] As the identification of murine SNPs facilitated genotype-phenotype associations, so too can the identification of common murine CNVs. Additionally, by knowing what variants commonly exist in a particular inbred mouse strain, it is possible to assess the accuracy of copy number calls in new datasets by looking for CNVs expected in that strain.[21,51,52]

1.2 Differences in copy number between tissues exemplify somatic mosaicism

Although all tissues within a given animal are derived from the same zygote, mutations can accumulate over the organism's lifetime as a result from the stress of DNA replication as well as from exposure to both endogenous and exogenous mutagens. Cell turnover rates differ

not only between tissues, but also between cell types within a tissue.[53] This variation in cell replication rate can contribute to the overall genomic variation between tissues resulting from replication-dependent mutational mechanisms. Mutation frequencies differ not only with age, but with tissue type as well.[54]

Mutations that occur early on in development, if not lethal to the cell, can be passed on to daughter cells as the original cell in which the mutation occurred divides, a process called clonal expansion. Such expansion can lead to the presence of multiple cell populations that are genetically distinct from one another within the tissues of a single organism. This phenomenon is referred to as somatic mosaicism.[55] Depending on the mutation and genes affected, somatic mosaicism may have no observable phenotypic effect, or mosaicism may be the underlying cause of disease, such as Sturge-Weber syndrome in humans.[56] Somatic copy number mosaicism is the specific case when populations of somatic cells are genetically distinct from one another as a result of a copy number change.

From the early days of CNV research, it has been acknowledged that tissues within an organism are not identical in the genetic variants they harbour. Sebat *et al.* performed genomic analysis in multiple human tissues, including blood and lymphoblastoid cell lines, and observed that differences in somatic copy number were restricted to gene clusters encoding T-cell receptors or immunoglobins.[12] An earlier study in inbred strains of laboratory mice found evidence for somatic copy number mosaicism at multiple loci.[57] The extent to which somatic copy number mosaicism affects the genome in different tissues is unknown. Somatic copy number mosaicism has been implicated in cancer, with evidence for mosaicism being detectable prior to cancer diagnosis.[58] Outside of tumour analysis, genetic analysis in humans is usually performed on samples drawn from the blood, which is a mesoderm-derived tissue. If the variant of interest was inherited, it will be present in all tissues, including the blood. However, if somatic copy number mosaicism is at play, ascertaining copy number from a blood sample may not be accurately capturing the genetic variation underlying a particular condition that affects a tissue other than hemopoietic tissues.[59] The interpretation of gene expression analysis may be

questionable if a CNV precipitates a gene-dosage effect in a specific tissue (such as the brain), yet the variant is not present in the tissue being sampled (such as the blood).[60]

Since CNVs are defined as population-level variants (analogous to the definition of SNPs as having two alleles that are both present at least 1% of the population), a different moniker is necessary to distinguish *de novo* alterations in copy number that differ between tissues within an individual organism from those variants that differ between individuals in a population. The term copy number change (CNC) specifically refers to the class of large structural variants that differ in copy number between somatic tissues within an individual organism.[61] It is possible that CNVs previously reported as *de novo* germline mutations may instead be somatic mutations that formed in the particular tissue that was analyzed.[62]

1.2.1 CNVs contribute to phenotypic variation, including complex phenotypes and diseases

The large amount of DNA that is involved in a given CNV, whether copies of sequence are gained or lost, begs investigation into the functional impact these variants may have on phenotype.[8] CNVs can directly encompass entire genes or parts thereof, and can span regions that regulate gene expression. Even CNVs that lie completely outside of coding or gene regulatory regions may influence gene expression: CNVs can alter the expression of nearby genes up to 450 kb beyond the breakpoint of the variant.[63] Perhaps the most well known case of gene duplications giving rise to phenotypic variation among individuals is that of the major histocompatibility (MHC) complex. Contained within the human MHC complex coding region is the RCCX locus, which is named for the four genes *RP* (since renamed STK19), *C4*, *CYP21*, and *TNX*, which appear in tandem to form a module. The number of modules (and thus the number of copies of the genes) varies between individuals, and this gives rise to phenotypic variation in immunity between individuals.[64] Common CNVs in humans are enriched for genes involved in immunity (including the MHC complex), as well as olfaction genes and

genes encoding secreted proteins.[65,66] These biases towards adaptive phenotypes in CNV genes are indicative of selective pressures acting on the associated phenotypes.

Any variation in the genome that influences gene expression has the potential to contribute to phenotypic variation, and this includes dysfunction. Aneuploidy (aberration in whole chromosome copy number) is well known to lead to severe phenotypic effects, such as trisomy 21 in humans.[67] Likewise, the smaller aberrations in copy number that occur due to CNVs can result in segmental aneusomies (disorders from inappropriate gene dosage in a segment of a chromosome).[68,69] CNVs have been associated with many diseases including Charcot-Marie-Tooth neuropathy,[70] schizophrenia,[71,72] cerebral palsy,[73] Crohn's disease,[74] psoriasis,[75] and osteoporosis,[76] to name a few. CNVs also contribute to less severe phenotypes, such as differential changes in amylase in response to diet[77] and differences in HIV susceptibility.[78] In the case of copy number gains, an increase in the number of copies of a given gene can lead to an increase in the expression of that gene.[77] Copy number losses tend to be observed outside of gene regions, which may result from selection acting against negative effects that may arise from the deletion of essential genes.[10] Despite the potential for deleterious effects, copy number losses tend to be more numerous than copy number gains.[14,29,48,79]

Perhaps one of the most well-known examples of *de novo* tissue-specific genomic structural abnormalities associated with a disease phenotype is cancer.[80–82] The first GWAS linking CNVs with cancer susceptibility was performed by analyzing the genomes of patients who were predisposed to cancer due to Li-Fraumeni syndrome.[81] Patients with Li-Fraumeni syndrome were found to have an increased number of CNVs in comparison to healthy individuals, particularly in regions with genes identified to be associated with cancer pathways.[83] Since then, tumour genomes have been analyzed in many cancers in an effort to better understand the contribution of copy number variation to oncogenesis. Some CNVs are themselves a root cause of cancer, as identified in one study of breast cancer tumour tissue.[84] Other variants only arise after tumourigenesis has begun, which is a period during which there is significant rearrangement of the genome.[85] Copy number variation thus plays a complex role in cancer, whether a

CNV encompasses a dose-sensitive gene or not.

If a given CNV affects the phenotype of the organism, it does not necessarily do so in a direct Mendelian manner. Rather than a given phenotype occurring as a result of a genetic variant at a single locus, the phenotype instead arises from the contributions of and interactions among several loci, and/or interactions between genes and environmental factors. Such complex phenotypes, where no one locus alone is responsible for the phenotype, have been associated with the presence of CNVs. Psychiatric disorders that have challenged geneticists for decades, such as schizophrenia and autism, exemplify complex phenotypes that are now being linked to CNVs in some instances.[86,87] Pleiotropy (multiple phenotypic effects stemming from a single gene) is yet another mechanism by which CNVs can influence complex phenotypes. In one human study, a duplication at 16p11.2 was associated with an elevated risk for several disorders, including schizophrenia and congenital malformations.[88] Differences in measures of "CNV burden" (which include total length, average length, and number of CNV events in the genome) implicate departures from population averages with disease phenotypes in the absence of a direct association with a single CNV locus.[89] The phenotypic effect of CNVs has even been observed to be haplotype-specific, as in the case of human killer immunoglobulin-like receptor (KIR) genes.[90] Additionally, deleterious CNVs may be masked by heterozygosity, so a CNV in a typically dose-sensitive gene may ultimately have no effect on phenotype at all.[91] Thus, the phenotypic impact of CNVs can range from no observable effect to severe disease. Phenotypic changes may result from direct changes in gene dosage, from indirect changes via gene regulatory regions, or from interactions with other genomic factors such as local haplotype. It is therefore necessary that we continue to identify CNVs and CNCs, ascertain phenotypic associations, and elucidate the mechanisms by which CNVs and CNCs can occur.

1.3 CNVs and CNCs are variable in origins and mechanisms

DNA replication is a cellular process that is high-risk for the formation of *de novo* CNVs by a number of different mechanisms. Non-allelic homologous recombination (NAHR) is widely considered to be one of the primary mechanisms of CNV formation, as it has a moderate-to-high mutation rate and is associated with segmental duplications.[92] Fork stalling and template switching (FoSTeS),[93] *Alu*-mediated recombination,[94] and strand slippage can also contribute to the accumulation of CNVs in replicating cells. Sequence homologies at CNV breakpoints may give researchers insight into the mechanism by which the CNV was formed, particularly in cases of repair-mediated mechanisms that occur in response to double-stranded breaks (DSBs).[24,95] Repair-mediated mechanisms include break-induced replication (BIR), microhomology-mediated break-induced replication (MMBR), and non-homologous end joining (NHEJ). Even within the relatively short time in which CNVs have been intensely studied, new research is constantly updating what we know of their mutational mechanisms. For example, although NHEJ was once considered to be a major mechanism of *de novo* CNV formation, a study of induced DSB repair in mouse embryonic stem cells demonstrated that NHEJ was not a main source of CNVs.[96]

In addition to endogenous factors such as DNA replication, certain exogenous chemical agents are known to induce CNC formation. These include clastogens that induce chromosome breaks, which opens up the possibility of repair-mediated mechanisms leading to faulty DNA repair. Rats injected with 7,12-dimethylbenz[a]anthracene (DMBA) were found to have increased copy number in genomic regions associated with tumourigenesis.[97] The presence of hydroxyurea (used to treat sickle cell disease) in cultured human cells induces CNVs that have breakpoints characterized by microhomologies, suggesting a replication stress-mediated mechanism.[98] Considering that CNCs may result from factors including replication stress and environmental mutagens, it is necessary that copy number assessment is performed when analyzing tissues for acquired mutations, and identify whether tissue type is a factor in CNC

formation.

1.4 A multi-tissue comparison is necessary to detect somatic copy number mosaicism

Embryonic development has been well established as a period of mutagenesis due to the high replication rate and the opportunity for clonal expansion of mutations to occur. Copy number variants are included in the list of genomic changes that can occur during this time; chromosome instability during embryonic development has been shown to produce somatic copy number mosaicism in human tissues.[99] Murine embryonic stem cell lines can differ in CNCs, and even cells derived from the same inbred mouse strain may harbour different CNCs.[100] Germ layers are formed during a process called gastrulation, which begins to occur at embryonic day 6.5 in the mouse.[101] It is at this stage that cells in the single-layered blastula begin to reorganize into the three distinct germ layers (the endoderm, mesoderm, and ectoderm) that will form a multi-layered organism. Gastrulation is a critical timepoint in development, so much so that embryologist Lewis Wolpert stated "It is not birth, marriage, or death, but gastrulation, which is truly the most important time in your life." Tightly controlled mechanisms maintain ordered development of certain tissues from each of the three germ layers.[102] Tissues derived from the endoderm include the liver, intestines, thyroid, and pancreas. Mesoderm-derived tissues include the spleen, circulatory system, kidneys, and skeletal muscle. Finally, the ectoderm gives rise to the central nervous system, including the brain. The detection of CNCs that differ between tissues may depend on the germ layer from which the tissues were derived. If a *de novo* copy number event occurs prior to gastrulation, it is more likely to be present in multiple tissues derived from different germ layers. On the other hand, if a *de novo* event occurs after gastrulation, it will be separated in space from cells in other germ layers, and thus may be present only in tissues derived from that germ layer.

It has been well established that mutations accumulate over the lifetime of an animal at

different rates in different tissues.[54,103–107] The different biological function of each tissue may contribute to variation in mutation accumulation between tissues within an animal, with some tissues being exposed to higher levels of reactive oxygen species or increased replication, two sources of cellular stress that may lead to DNA damage.[108,109] Although research has been executed regarding the mutation spectra in each tissue type, little is known about the accumulation of CNCs over the course of an organism's lifetime. Mutation types have been found to be similar across tissues despite different mutation rates, but these analyses do not test for larger structural variations such as CNCs.[106]

1.4.1 The genome of the spleen is subject to somatic hypermutation

The spleen, like the blood, is a mesoderm-derived tissue. In mice, development of the spleen begins at embryonic day 12.5.[110] Unlike the human spleen, the murine spleen is hematapoietic, meaning that it is involved with blood cell production. The heterogeneous cellular composition includes dendritic cells, granulocytes, lymphocytes, and macrophages. Many of these cells are related to the spleen's role in the immune system. Immunological function renders the spleen subject to somatic hypermutation, including V(D)J recombination and base substitutions.[111,112] Additional mutagenesis in the spleen may result from a high rate of cell turnover in the adult mouse (between 1 and 21 days, depending on cell type)[53] CNVs have been identified in the murine spleen previously, although it is not known whether these are tissue-specific or present in other tissues.[48,113]

1.4.2 The genome of the cerebellum is more plastic than previously thought

The murine cerebellum is located at the back of the brain, which itself is an ectodermal tissue that begins development at embryonic day 10.5. The cells of the cerebellum consist

largely of neurons (accounting for about 90% of cells), including Purkinje cells and granule neurons.[114] The largely post-mitotic nature of many cell types, including the aforementioned Purkinje cells and granule neurons, in the cerebellum may exempt the DNA from replication-based mutagenesis compared to other tissues in the body; however, recent evidence suggests that the genome of the brain, including in the cerebellum, may be more plastic than previously thought.[115] Aneuploidy is known to exist in the human brain, and it has been suggested that such mosaicism may be biologically typical and not necessarily associated with any pathology.[116,117] Similarly, aneuploidy is known to occur in the mature murine cerebellum.[117] The authors of this study rejected the possibility that CNVs may be contributing to their observations of variation in fluorescent *in situ* hybridization (FISH) signals for two reasons: firstly, the assumption that CNVs are constitutive and thus would be detected in all cerebellar cells; and secondly, the FISH probes used in this study did not overlap with any CNV loci that had been documented at that time. We now know that large structural changes can indeed occur during brain development.[118] Advances in technology that permit single-cell genome analysis have recently revealed copy number mosaicism among human neurons.[79] The *de novo* variation observed in the brain thus makes the detection of CNCs as compared with other tissues likely.

1.4.3 The liver experiences genome instability with age

The liver is an ideal tissue for genome analysis as it provides a good source of large quantities of high molecular weight DNA.[119] The liver develops from the endoderm at embryonic day 9.5. Aside from the earlier time of development, the murine liver offers similarities to each of the spleen and cerebellum in several ways. Like the cerebellum, the liver has also shown to be affected by aneuploidy,[120] and is largely homogeneous in its cell type composition, with approximately 70% of cells being hepatocytes.[121] The cell turnover rates of hepatocytes range between 480 and 620 days.[53] CNVs have been observed in the liver in both humans and

mice.[122–125] Genome stability in the murine liver changes with age, with genomic rearrangements implicated in the observed increase in mutant frequency with age.[107] These findings lead us to expect *de novo* CNCs to be observed in the murine liver.

1.5 Genomic regions containing multiple CNV events across individuals are copy number variable regions (CNVRs)

In contrast to an inherited CNV that occurs at a fixed chromosomal position, a copy number variable region (CNVR) is a region of the genome in which different individuals harbour different variants that are close in proximity and/or overlap with each other.[19] Estimating the breakpoints of variants in these regions is difficult as the CNVs themselves are different between individuals. Assigning an overlap threshold when analyzing CNVs across a given set of samples is useful as it can identify CNVRs. Previous studies have found an overlap of 40% of the length of the CNVs to be sufficient to identify CNVRs.[21] When identifying CNCs from multiple tissues, CNVR analyses may reveal genomic regions prone to acquiring *de novo* CNCs, otherwise known as "hotspots". First identified in one of the seminal 2004 publications on copy number variation, it was noted that despite the genome-wide distribution of variants detected in human subjects, some genomic regions harboured multiple variants.[12] The authors suggested that these regions could be "hotspots" for the formation of CNVs. CNV hotspots in the human genome have since been observed in several discovery studies,[66,126,127] and have been associated with disease.[128] The genomes of other primates, including chimpanzees and rhesus macaques, show clusters of CNVs at sequences similar to those found in humans, with some hotspots occurring across all three of these primate species, once sequence alignment has been accounted for.[66] The authors of this study cited high levels of gene expression in genes occurring within these hotspots of recurring CNVs across primates as evidence that these regions are under positive selection, which is perhaps preventing certain individual mutational events from becoming fixed in a population.

Mechanisms underlying the existence of CNV hotspots can differ between species, even within the class *Mammalia*. A case study illustrating this point is one that investigated the genomes of humans and domestic dogs to understand the role of the human positive-regulatory domain containing 9 (PRDM9) protein in recombination. The product of the gene *PRDM9* has been implicated in the initiation of the double-stranded breaks (DSBs) that occur during meiosis in both the human and mouse genomes.[129] In humans, the 13-bp binding motif for this protein is associated with recombination hotspots,[130] yet in dogs, PRDM9 is inactive. The result is that in dogs, CNV hotspots are shifted away from regions containing the PRDM9 binding motif, and are instead found in regions with GC content peaks.[29] A link between GC content and CNVs has been noted in other mammalian genomes besides the dog. Increased genomic GC content is driven by recombination in the human genome.[131] Regions in the human genome with high GC content are known to be prone to CNV formation, with higher GC content associated with larger CNVs.[132] CNVs detected in the bovine genome have an average GC content percentage that is elevated above the average for the rest of the genome.[30]

Human CNV hotspots are enriched in regions of segmental duplications, which are stretches of DNA with high sequence identity interspersed throughout the genome.[133] The high sequence identity between these genomic regions makes them prime candidates for the NAHR mechanism of CNV formation in humans, and indeed, CNVs are found to be enriched in regions of segmental duplications.[134] CNV hotspots are also present in the mouse genome, with rates of recurrent CNV formation varying across hotspot loci.[135] As in humans, murine CNVs are enriched in regions of segmental duplications, most likely due to the high degree of sequence identity in these regions. Similar to the dog and human genomes, regions with higher GC content in mice are associated with increased recombination in the mouse genome.[131] Additionally, there is a positive correlation between GC content and gene density in both mice and humans,[136] making GC content a prime metric to assist in the identification of CNVs and to characterize CNVs for potential phenotypic impact.

While CNVRs are indicative of copy number events present in multiple samples, singletons

are identified as CNVs that are detected only once in a set of samples in a given study.[19,137] Within a study, singletons tend to account for a small percentage of the total number of variants identified, [138,139] although they have a high validation rate when confirmed using secondary CNV detection methods, giving confidence that initial singleton calls are likely not false positives.[19] Singletons can severely impact phenotype, and have been associated with diseases such as Alzheimer's.[138] Singleton losses of copy number in particular, as opposed to singleton gains, contribute to elevated risk of developing bipolar disorder.[140,141] In addition to copy number state, the length of singletons appears to be an important factor as well, with larger events associated with disease phenotypes such as intellectual disabilities and bipolar disorder.[141–144] The relative frequency of singletons as well as their potential to impact phenotype make it necessary to include singletons in CNV analyses as they may indicate novel variants at a population level, or if in a study investigating copy number across multiple tissue types, they may be indicative of *de novo* CNC formation.

1.6 Copy number can be used to estimate genetic distance and distinguish between individuals and between populations

Genetic distance in the phylogenetic sense refers to a quantitative measure of the divergence between two evolutionary divergent populations descended from a common ancestor. There are many different methods that can be used to calculate genetic distance between two populations, most of which use sequence identity to estimate divergence and construct a phylogenetic tree. Since CNVs can be inherited as well as form *de novo*, CNVs can be employed to calculate genetic distance. The identification of CNVs in humans from multiple Chinese populations revealed that some CNVs were shared across populations, while other CNVs were population-specific.[139] By generating phylogenetic trees from CNV data, the researchers were able to capture genomic diversity at the population level, and reveal that genetic distance based on CNVs was closely correlated with linguistic patterns among the populations. Similarly,

evolutionary examination of CNVs in primates reveals that CNVs diversity is correlated with SNP diversity, and that divergence in primate lineages can be estimated using CNV losses.[145]

In mice, the carefully documented breeding history of inbred strains permit accurate assessment of approaches for calculating genetic distance.[146] Cutler *et al* demonstrated that SNPs could be used to produce phylogenetic trees that successfully capture the known relatedness of the inbred strains.[49] In that study, murine CNVs were used to construct trees that were found to be topologically similar to SNP-based trees, showing not only that CNVs contribute to divergence between inbred strains, but that genetic distance can be estimated using CNVs as a metric. CNVs can also be used to distinguish between mice even within the same inbred strain, most notably in C57BL/6J,[147] the most commonly used inbred strain and the source of the mouse reference genome (The Jackson Laboratory; http://jaxmice.jax.org/strain/000664.html).

Since CNVs can be used to distinguish between individuals and between populations, it is expected that CNCs may be used to distinguish between tissues in an individual. Given our knowledge of the inherited and *de novo* nature of copy number events, as well as given evidence that there is variation in tissue biology and genome instability throughout the body of the mouse, we can make use of breeding records and tissue availability to extend the analysis of copy number variation from the levels of species and populations to a much finer scale of identifying differences between tissues within a single mouse. Such investigation into somatic copy number mosaicism can be difficult to do in isolation from confounding variables, given the degree to which mice from the same inbred strain can differ.[100] Therefore, minimizing external sources of variation while maintaining a large enough sample size of mice requires the use of closely related mice, much like CNV studies in humans have made use of parent/child families.[10,81,148–151] By studying mice from the same litter as well as their parents, we can make use of multiple tissues from a given mouse, in addition to the next most closely related samples attainable in order to better comprehend the frequency and nature of CNVs and *de novo* CNCs. Additionally, the known breeding records of mice can be used as known metrics against which estimates of genetic distance using CNV loci can be compared as has previously been done in

humans and other primates.[139,145]

1.7 The trend towards increasing genetic diversity among laboratory mice requires high-resolution technology to capture genetic variation

In theory, mice from the same inbred strain are genetically identical to one another at every base in the genome (save for some known regions of residual heterozygosity), due to many generations of brother-sister mating.[152] In practice, genetic variation within an inbred strain does arise.[153] Spontaneous mutations occur via a multitude of mechanisms, and any mutation may be passed on to the next generation if it were to occur in the germline. Much like in the case of clonal expansion of a mutation within a proliferating tissue, the genomes of the mice descended from the originating mutant will also harbour the mutation, rendering individuals within the inbred line to be non-identical. The type of mutation and where the mutation occurs in the genome will largely dictate whether or not it will have any phenotypic influence; some mutations will be lethal, while others will have no observable impact on the phenotype and so may never be discovered unless sequencing happens to be performed at that locus.

In contrast to the practice of breeding of inbred strains, in which the goal has been to decrease population-level genetic diversity by approaching almost complete genome-wide homozygosity, there have been recent efforts to breed mice in the opposite direction, towards increasing genome-wide heterozygosity and increasing population-wide genetic diversity. The goal in these breeding programs is to generate populations of laboratory mice in which the genetic diversity mirrors the genetic diversity of the human populations they are often used to model. Such diversity-driven breeding programs include the Collaborative Cross[154,155] and the Diversity Outbred mice.[156] Mouse strains developed by these projects are the result of strategic crossings of many inbred strains plus wild-caught and wild-derived mice. In the case of the Diversity Outbred program, the breeding program is designed so that ultimately, every mouse will be genetically distinct from all others. Non-traditional breeding programs such as these bring

forth the need for a genome-wide approach for genotyping laboratory mice, as researchers can no longer assume that the genomes of their murine subjects are mostly homozygous. In addition, the discussion surrounding genomic variation in mice thus far has been largely focused on single base pair differences, namely point mutations and SNPs. For researchers concerned with genetic homogeneity in mouse models, the genetic variation that arises due to structural variants such as CNVs and CNCs is essential to characterize. To assess the genotypes of thousands of SNPs present in the genome, as well as permit analysis of newly recognized structural variants, it is necessary to develop technology that is not only high-resolution (distinctly assays a large number of features in the genome) but also affordable in order to capture genetic diversity in the mouse genome for many samples.

1.8 Array-based technology permits affordable genome-wide copy number discovery

The ability to sequence whole genomes of individuals in many species has allowed researchers to survey nearly all genetic content for variation. Whole genome sequencing (WGS) methods coupled with bioinformatic analysis can detect CNVs in addition to providing the sequenced genome.[157] Similarly, exome sequencing can be used to identify CNVs in protein coding regions while drastically reducing the amount of data through which researchers must sift when compared to WGS.[158] Designing an experiment to include multiple tissues from closely related mice requires methods of copy number detection and analysis that are not only capable of genome-wide assessment, but are also economically and temporally feasible. Array-based methods can be employed for copy number detection for a fraction of the cost of both WGS and exome sequencing, permitting affordable genome-wide copy number discovery and the opportunity to increase sample size.

There are two main types of array-based methods for CNV discovery: array-comparative genomic hybridization (aCGH) and genotyping arrays.[159,160] One of the fundamental differ-

ences between these two array types is the reference to which a sample on the array is compared, with aCGH arrays co-hybridizing clones of control DNA as reference with the sample, while genotyping arrays accept one labelled DNA sample per array and comparisons to a reference are made computationally after scanning the resulting hybridization fluorescence. This difference in array design leads to the necessity of different downstream procedures in the interpretation of fluorescent data obtained from each array type. Because aCGH directly compares two labelled samples on the same array, aCGH is frequently used in cancer research for discovering structural variants that differ between tumour and non-tumour tissue taken from the same patient. Each of the two samples is tagged with a different fluorescent label, so that quantitative differences between the samples can be detected via fluorescence intensity.[159] The theory behind using array-based CNV discovery assays is based on the assumption that loci close together are likely to have the same copy number.[161]

Alternatively, genotyping arrays provide the ability to assess both SNP genotypes and copy number using the fluorescence intensity from the same set of oligonucleotide probes that are tethered to the array.[162,163] Genotyping arrays typically comprise sets of oligonucleotide probes, with each set designed to interrogate both the A and B alleles for one SNP locus in the genome of the organism of interest (the organism for which the array is specifically designed).[164] The added benefit of obtaining SNP genotypes is important when identifying copy number losses in highly heterozygous genomes (such as the human genome), as the loss of a segment of DNA will typically lead to loss of heterozygosity at the SNPs contained within that region, since there isn't a second strand from which to genotype the alleles for those SNPs. Genotyping arrays are much more economically feasible than whole genome sequencing, and allow for a genome-wide approach for assessing copy number by interrogating thousands of loci in parallel on a single array. Compared to aCGH methods, the protocols for SNP genotyping arrays may also call for less sample per experiment, thus allowing for future confirmation of putative copy number events.[165] The high-resolution, low-cost, genome-wide assessment offered by genotyping array platforms enables researchers to design experiments using multiple

tissue types to explore somatic copy number mosaicism.

Not all array platforms for detecting CNVs are equal. The different designs and capabilities of CNV detection platforms lead to different results. Numerous studies have performed cross-platform comparisons of copy number data, identifying differences in resulting calls and assessing reproducibility and accuracy.[51,166] Many discovery studies have used two or more platforms as complementary methods to identify CNVs.[19,52] It is thus necessary to critically assess new CNV detection platforms for performance prior to investigating the biology in question.

1.9 The Mouse Diversity Genotyping Array is a high-resolution array-based genotyping platform that can also detect copy number in the mouse genome

Genotyping arrays for human CNV research remain the gold standard, with high-resolution arrays such as the Affymetrix® Genome-Wide Human 6.0 Array, providing both genotype and copy number data.[74] High-resolution technology for analysis of the human genome is available and affordable, but comparable technology for analysis of the mouse genome has lagged behind in development. For example, the Affymetrix® Genome-Wide Human 6.0 Array tests over 900,000 SNP loci and contains an additional 900,000 invariant probes for detecting copy number, while until recently the highest resolution SNP genotyping array for the mouse was limited to just 10,000 SNP loci.[167] The Affymetrix® Mouse Diversity Genotyping Array (MDGA) was first introduced in 2009 as the highest-resolution genotyping array available to date for analysis of the mouse genome.[168] The MDGA was initially developed by The Jackson Laboratory (Bar Harbor, ME) and manufactured commercially by Affymetrix® Inc. for research use.

Starting at approximately $700 CAD per array, the MDGA is a relatively affordable option for genome-wide analysis in comparison to WGS methods (approximately $35,000 per

sample). The MDGA contains oligonucleotide probes for 623,124 SNPs and an additional set of 916,296 invariant genomic probes (IGPs) specifically for defining copy number. Since the original 2009 publication, The Jackson Laboratory has provided some revisions to the annotation of the SNP probes represented on the array and have recommended certain probes on the array be eliminated from analysis based on performance criteria.[169] To date, over 300 mice have been genotyped using the MDGA, providing a rich source of genotyping information for laboratory mice that is specific to this array.[169,170]

The Affymetrix® GeneChip© family of arrays (which includes the Affymetrix® Genome-Wide Human 6.0 Array, the MDGA, as well as a variety of expression arrays) is popular amongst genetics researchers, and so there are many computational pipelines for analysis established for these arrays. The similar design of the GeneChip© permits the use of these established pipelines; however, it is essential to clarify here that the novelty of the MDGA design differentiates it from typical Affymetrix® genotyping arrays. The unique design thus limits the number of analysis pipelines that can be used to call copy number from this array, as the design affects fundamental fluorescent signal corrections and summarizations necessary in the analysis of array data.

The MDGA design is based upon the SNP-CNV hybrid model initially developed with the Affymetrix® Human 6.0 array. Both of these arrays combined the traditional genotyping approach of including probe sets designed to interrogate hundreds of thousands of SNP loci with a novel approach for interrogating non-polymorphic genomic sequences.[74] Probes for these non-polymorphic sequences, numbering in the 900,000 range, were included to provide better coverage of the human genome and interrogate regions that would otherwise be undetected by SNP probes only, as SNPs are not uniformly distributed throughout the genome.[74,171] The ultimate intention to use a combination of probe types was to increase the resolution of the array to increase the narrowing down of CNV breakpoints, as well as to better estimate the number of copies of a particular sequence actually present in the genome.

The physical MDGA itself is a small glass square, on which unlabelled probes are synthesized.[172] The area on the array surface that contains a particular probe sequence is called a feature. Each 5 μm x 5 μm feature on the array contains approximately 1.66^{-6} picomoles of identical 25 bp long single-stranded DNA probes (calculated from data communicated through personal correspondence with an Affymetrix® representative). The sequence that is synthesized in each feature is documented in an accompanying annotation computer file, along with the *(x,y)* coordinates of the feature's position on the surface of the array. This annotation file is used to determine which target sequence a particular feature is reporting via fluorescence intensity.

Typically, Affymetrix® oligonucleotide array platforms (both genotyping arrays and expression arrays) contain two types of probes for each locus assayed: perfect match (PM) probes and mismatch (MM) probes. A sequence of interest in the genome (target sequence) is interrogated by both PM and MM probes: PM probes to detect the fluorescent signal pertaining to the sequence of interest, and mismatch (MM) probes to estimate background fluorescent noise specific to the 25 bp sequence. For each target sequence, the MM probe is identical in sequence to the PM probe except for the 13th position in the sequence, at which the MM probe will differ by a single nucleotide. PM and MM probes for a given target sequence are physically adjacent to each other on the array. Measuring and comparing the fluorescent intensities of both PM and MM probe sets for each target sequence allows for probe-specific noise correction steps to be taken.[173] In contrast to this traditional Affymetrix® GeneChip© array design, the MDGA is unique in that it contains only PM probes. Although correction steps using the MM probe signals in conjunction with PM signals are common, they cannot be employed for data acquired from the MDGA due to the absence of MM probes, imposing a limitation on which existing strategies can be employed to analyze this array.

It is necessary to use the most up-to-date information possible when performing microarray studies of any kind. Arrays and their associated kits, documentation, libraries, and software are often marketed as packages that provide a straightforward pipeline for analyzing biologi-

cal information. Given the amount of research being performed on the genome of a particular organism at any given time, and compounded with the rate of increasing accuracy in technology, the accuracy of probe set design and annotation at the time of manufacturing may be undermined by an update in the information of even just a single gene. Dai *et al.* analyzed and reorganized probe set definitions for Affymetrix® GeneChip expression arrays, and found significant departures in gene and transcript results from the original probe set definitions.[174] Post-hoc removal of failing probesets may lead to lost information on real biological variation and lead to ascertainment bias.[169] Selecting probe sets that have been analyzed for performance across many arrays and samples, as well as for correct annotation, is essential to limit sources of error before genotyping and copy number calling are performed.

1.10 Array data can be used to infer SNP genotypes and copy number

A genotyping array provides data about a sample of genomic DNA in the form of raw fluorescence intensity for each feature, which is contained in a CEL file (in the file format .CEL).[175] Each array type has a corresponding chip definition file (.CDF) which provides the annotation for the features on the array using the *(x,y)* coordinates of the 2-dimensional array surface. Raw fluorescence intensity from the array is converted into a CEL file using a specially designed array scanner coupled with image processing methods.[176] It is this CEL file that is used to infer SNP genotypes and CNVs after correction and normalization procedures have been applied.

Normalization procedures are usually employed to account for differences between arrays that are non-biological, including differences in sample labelling and array production.[177] For example, batch effects are technical artefacts that can arise when comparing arrays that were processed at different times, or in different "batches", for which mathematical adjustments can often be made.[178] Within-array variation can also stem from the GC content of the probes. Peaks in fluorescence intensity acquired from genotyping arrays correlate with increased GC

content.[179] This "GC waviness" can be minimized by applying mathematical corrections to array data to account for potential differences in hybridization based on probe sequence.[180] The normalized and corrected fluorescence data can then be used to call genotypes and infer copy number.

Genotyping arrays such as the MDGA require an algorithm to determine the genotype of each SNP locus interrogated by the microarray. These algorithms take into account the fluorescence intensity of all probes used to interrogate each locus (in the case of the MDGA this is eight probes per locus), and calculate the most probable genotype, returned by the algorithm as AA (homozygous A), BB (homozygous B), or AB (heterozygous), where A represents the allele present in the C57BL/6J (reference) genome, and B represents an alternative allele. The algorithm recommended for use with the MDGA is a variation on BRLMM (Bayesian Robust Linear Modeling using Mahalanobis Distance) called BRLMM-P (the P stands for perfect-match, as the MDGA has only perfect-match probes).[181] This algorithm looks at the clustering properties of each genotype for each probeset and calls the most likely genotype based on the summarization of each allele's fluorescent intensities for a particular SNP. Most probesets yield fluorescence intensities that can be graphed as three distinct clusters, with one cluster for each genotype call (AA, AB, and BB). Although the shape and position of each cluster is predicted before the algorithm is run (the prior cluster), the algorithm will re-adjust the clusters as it runs with a given set of CEL files (creating posterior clusters). From these posterior clusters, silhouette scores are calculated for each data point (which represents a single sample's summarized fluorescence intensity value for the SNP locus in question) to assess the clustering.[182] A data point that has a silhouette score below a predetermined threshold is considered to have failed genotyping, as it was not placed near enough to one of the genotyping clusters to confidently call a genotype. When a silhouette score falls below this threshold, the algorithm will return a "NoCall", indicating that the algorithm was unable to genotype the SNP in question for a particular sample.

1.10.1 Runs of consecutive "NoCalls" may indicate deletions

In some instances, the SNP-calling algorithm is unable to return a genotype, and in these cases the result "NoCall" is returned instead. There are several reasons that a "NoCall" may be returned instead of a genotype. The genotyping algorithm itself will have some error rate associated with it; this also means that some of the SNPs that appear to have genotyped well (assigned either AA, AB, or BB) may in reality be a different genotype or a "NoCall". Typically, a "NoCall" results from low fluorescence intensities from both allele probes (private correspondence with Affymetrix®). A "NoCall" may also indicate hemizygosity at that locus, as the genotyping algorithms are not capable of returning a hemizygous call.[183] Another possibility that a "NoCall" may occur is a two-copy deletion, which in a diploid genome would indicate a copy number of zero. These calls are returned for a single base pair position in the genome (the SNP of interest that is interrogated by a given probeset), but it is possible that two or more "NoCalls" that occur consecutively along a chromosome may indicate a larger deletion, based on the principle that genetic markers that are physically close together on a chromosome are likely to have the same underlying copy number.[161,184] Similarly, a recent effort by Standfuß *et al* used the MDGA for copy number analysis of tumour tissue, but did so in the absence of genome-wide algorithms designed for the array.[85] First, the authors identified normalized SNP probe intensities that differed from a reference intensity calculated from normal tissue samples. Next, they identified groups of consecutive SNPs that reported similar normalized SNP probe intensities, which were interpreted as relative copy number for each SNP. The advantage of using consecutive genotype calls, as opposed to raw fluorescence intensities, is that it allows for the detection of putative deletions in a genome-wide discovery study that does not have the benefit of a case-control study design, as would be necessary with the Standfuß *et al.* approach.

1.11 CNVs can be detected from genotyping array data using algorithms

Computational methods to detect copy number events using microarray platforms range from commercially available software packages that allow researchers to input and analyze data using prescribed workflows (such as Partek® Genomics Suite and Golden Helix® SNP and Variation Suite), to open-source bioinformatics programs that can be downloaded from the Internet and invoked through the command line. The open-source programs tend to be more suited for more advanced bioinformatics users of R, Perl, Python, and other such programming languages, while the commercial programs offer streamlined workflows with graphical user interfaces that are more user-friendly. When choosing which tool to use for a particular study, there are important considerations.

A number of algorithms are designed for specific platforms, such as CNstream for Illumina arrays and TAPS (Tumor Aberration Prediction Suite) for tumour samples on Affymetrix arrays.[185,186] `MouseDivGeno`, an open-source R package, is the only software package designed specifically to handle the unique design of the MDGA. `MouseDivGeno` uses an algorithm called SimpleCNV, which detects CNVs by comparing the sample CEL files with a single user-selected CEL file designated as the reference. The problem with this one-to-one comparison is that it is highly subject to any variation present in the reference. Conversely, other algorithms require over a hundred CEL files to be run simultaneously with the test samples in order to call genotypes accurately and account for interchip variation. Given the lack of MDGA copy number work in the literature, `MouseDivGeno` was not deemed suitable for copy number detection at this time.

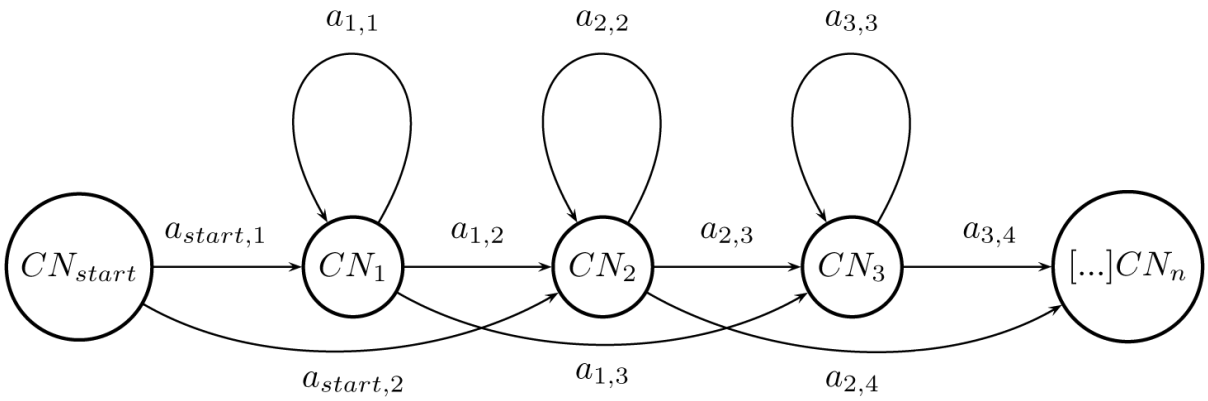
Other algorithms are capable of analyzing data from multiple platforms, permitting researchers to choose from a number of tools for a given array. But, like there is variation across platforms for detecting copy number, there is variation across the bioinformatic tools available to analyze data from each platform.[21] Different tools employ different strategies to handle the

same data, and the differences in the resulting copy number calls reflect these different strategies. Numerous studies have aimed to establish factors affecting the concordance of CNV calls made with different tools.[19,51,187,188] Due to the complex nature of CNVs and the differences in calling methods, they are difficult to ascertain with identical breakpoints across algorithms, even given the same set of array data.[51] In cases where SNP genotypes are highly concordant between different genotyping algorithms on a given set of arrays, the concordance of CNV calls derived from the same fluorescent signals may still be very low between algorithms.[189] For these reasons, many concordance studies recommend using multiple algorithms or software packages to call copy number in any given study.[34,165,189–191] Baross *et al.* recommend a minimum of two software packages, and to proceed using only the union of the two datasets in any downstream analysis.[189] Analyzing only CNVs found in common across multiple algorithms has been found to lower the rate of obtaining false-positive calls.[51,141]

1.11.1 CNVs can be detected with Hidden Markov Model-based algorithms

Bioinformatics is a relatively new discipline in biology, but many of its core principles are built upon well-established foundations in mathematics and statistics. A widely used example of the application of statistical approaches in bioinformatics is the Hidden Markov Model (HMM). When applied to CNV analysis, HMM is a stochastic approach for detecting regions that depart from the normal diploid state of the mouse genome (two for autosomes) and operates under the assumption that there exists a predetermined probability of a state transition; that is, the probability of moving from state A to state B given a certain set of circumstances (Figure 1.1). In copy number detection, the Hidden Markov Model is a profiling method to detect copy number based on established probabilities of state observations and state transitions, where state is the number of copies of a particular region present in the genome. In the application of HMM to copy number detection in a diploid genome, state transition probability

FIGURE 1.1: **Schematic diagram of a Hidden Markov Model for detecting copy number variation.** Beginning with the starting copy number (CN) state in the sequence CN_{start}, the transition to any other copy number state n has a transition state probability a.


$a_{1,1}$
$a_{2,2}$
$a_{3,3}$
CN_{start}
$a_{start,1}$
CN_1
$a_{1,2}$
CN_2
$a_{2,3}$
CN_3
$a_{3,4}$
$[...]CN_n$
$a_{start,2}$
$a_{1,3}$
$a_{2,4}$

can be understood as follows: given the current copy number state of 2, there is a strong probability that the next state will remain 2, equal probability that the state will move to either 3 or 1, and a lower probability that the next state will jump to 4 or 0. There are other factors that may influence the algorithm's decision to call a different copy number state as it moves along the genome. The collection of fluorescence intensities representing genomic loci interrogated by an array are understood by HMM-based algorithms as individual observations, which are ordered according to genomic position to provide a sequence of observations. Sufficiently high fluorescence intensity readings for SNP probes that interrogate a particular genomic region may push the algorithm to call a state transition to a higher copy number, even in the presence of a lower state transition probability.

Partek® Genomics Suite™ (herein referred to as Partek) offers two different algorithms for detecting copy number from array data: genomic segmentation and HMM. Genomic segmentation is Partek's proprietary method that is based on circular binary segmentation.[192] Partek recommends genomic segmentation when analyzing heterogeneous tissue samples such as tumour tissue (personal communication with a Partek representative).[193] For DNA samples expected to be more homogenous (that is, non-cancerous), the HMM method is recommended. Partek only became available for the MDGA in April 2011. At the time that the present experimental work was completed, Partek did not officially support copy number analysis using the MDGA. The unique array design and format of the accompanying annotation for the MDGA renders the invariant genomic probes (IGPs) currently unreadable by the software. Therefore, the ability to use Partek to call copy number using the MDGA is currently limited to using SNP probes only.

PennCNV is an open source HMM-based program to detect copy number from array data.[194] The first publication introducing PennCNV noted that the program is capable of detecting smaller CNVs.[194] Eckel-Passow *et al.* performed a CNV-calling software comparison for the Affymetrix® Human 6.0 array, assessing the performance of PennCNV, Affymetrix® Power Tools, Aroma.Affymetrix, and CRLMM.[188] PennCNV was ultimately recommended by

the authors for the analysis of locus-level copy number as well as summarized segmented data, noting that PennCNV displayed minimal bias and variability when compared with the other methods.[188] In a previous microarray study, the HMM model in PennCNV and the genomic segmentation model Partek produced similar results,[137] while in a different study, Partek was found to call the fewest and largest events.[195] It is thus expected that using the HMM model in Partek will produce results that are highly concordant with the PennCNV model as well, although it is also expected that Partek events will be longer than PennCNV events. Using genotyping results as well as the two HMM-based algorithms in Partek and PennCNV will provide multiple assessments of copy number across the same set of arrays to offer intensive and robust measures of somatic copy number mosaicism.

1.11.2 CNV detection algorithms often call more losses than gains

The increased frequency of copy number losses relative to copy number gains has been documented in several CNV discovery studies.[14,29,48,50,79] Mathematically, it is easier for an algorithm to detect a loss in copy number from the fluorescence signals received from microarrays. This is because for a diploid genome, to have a one-copy deletion requires the intensity to be halved relative to the diploid reference signal, while a one-copy gain requires the intensity to be increased by 50%. Some arrays are also more biased towards calling more losses than gains.[51] Previous studies using aCGH methods have found more losses than gains, but have cited this detection bias as the likely cause of these results.[74,113] However, the more recent sequencing-based CNV detection methods employed by Quinlan *et al.* confirmed that losses are more frequent in the mouse genome than gains.[92]

1.12 Resolving CNVRs from events called from array data

A common problem when using array-based techniques lies in identifying copy number events that are found to occur across multiple samples at the same CNV locus (a fixed chromosomal position).[7,19] Array-based CNV detection platforms and their associated algorithms do not have high enough resolution to give definitive CNV breakpoints down to base-pair resolution. Instead, the start and end positions returned with each CNV call are breakpoint estimates. For example, a CNV that is inherited within a family should theoretically have the exact same start and end position in the genomes of all family members that possess it; that is, the CNV should occur at the identical CNV locus in each person. But due to the noisiness of breakpoint estimates, the start and end positions for the CNV in each person may vary, making the CNV appear larger, smaller, or shifted in location among related family members. Similarly, the start and end positions of variants that are not inherited, but occur in CNVRs, may also appear to have different breakpoints.

It is therefore necessary to assess overlap of genomic location in the copy number calls made in multiple samples. This can be done manually by assessing the start and end positions of each event in relation to the start and end positions of nearby events in other samples; however, this approach is time consuming, labour-intensive, and prone to human error (citing personal experience). Computational automation of this step would provide quick reporting of overlap across samples. This solution is achieved by the novel program Hotspot Detector for Copy Number Variants (HD-CNV).[196] This custom software was developed by the authors tailored to my requirements for source files and the ability to adjust settings for the present study. HD-CNV takes information about CNV events detected by other CNV calling methods, such as Partek or PennCNV, scans across the start and end positions of each event, and then defines a "merge" if the smaller event overlaps the larger event by a user-specified percentage threshold (the default is 40% of the length of the smaller event).[21] The program will also define a "family" based on a user-specified percentage threshold higher than that for a merge. Events

that are found to be contained within a merge as defined by HD-CNV are referred to as merge-associated events. Events that do not overlap with events in any other samples are identified as singletons, as they represent the only copy number events that occur at that genomic location in the sample set.

Although the large amount of genetics research that has been carried out in the mouse has permitted more CNV analysis in *Mus musculus* than any other organism after humans, the amount and accessibility of data pertaining to murine CNVs still lags behind that of human CNVs. Previously reported structural variants in the mouse genome are publicly available in NCBI'S Database of Genomic Structural Variation (dbVar);[197] however, dbVAR does not allow researchers to upload custom tracks to compare in-house data alongside database entries in the viewer. Converesely, the University of California Santa Cruz (UCSC) Genome Browser allows users to upload custom tracks to the browser, but as of January 2014, the browser did not provide the option of accessing data on previously reported CNVs in the mouse genome (although CNV data are available for the human genome).[198,199] Current limitations in analyzing genomic data present challenges that, despite other benefits of using the mouse as a study animal, must be overcome to move our understanding of genomic variation forward.

Approaches taken to calculate false-positive rates for data from the Affymetrix® Genome-Wide Human 6.0 Array inspect SNPs contained within putative deletions for heterozygous genotyping calls, which should not be present in a deletion segment.[188,189] For genomes with a high degree of heterozygosity (such as the human genome), identifying loss of heterozygosity is informative. The same approach is not applicable to organisms with a high degree of homozygosity, such as inbred laboratory mice which are expected to be homozygous at nearly every SNP. A false-positive rate has yet to be established for copy number events called from the MDGA. Estimates of false positive rates made using data from the Affymetrix® Genome-Wide Human 6.0 Array vary between calling algorithms. False-positive rates using CRLMM and PennCNV have been estimated at 26% and 24%, respectively.[188] The false-negative rate of the array is difficult to quantify, as the amount of a given genome that is not called as being

copy number variable is much greater than the amount of the genome that is affected by CNVs. Despite this difficulty, researchers can make use of the tools and data available from previous studies. Here, the in-house construction of an annotated, accessible, and interactive database of previously discovered mouse CNVs provides a list of common variants that could be expected to be present in the strain of laboratory mouse under investigation.

1.13 Purpose and specific aims

For the first time, the contribution of somatic copy number mosaicism to the genomic diversity among a group of highly related mice will be investigated using the Mouse Diversity Genotyping Array. Given the novelty of the array and the limitations both in existing copy number data for the mouse in general, and in the available copy number analysis pipelines that can assess MDGA data, the work presented in this thesis will establish and evaluate the concordance of three methods for calling copy number and ultimately detect putative somatic copy number mosaicism. Computational approaches will be developed to identify tissue-specific recurrent events in two somatic tissues. The contribution of somatic mosaicism to overall genomic diversity among samples taken from multiple tissues in a group of highly related mice will be assessed using CNV- and CNC-based approaches to estimate genetic distance. Due to the inherited nature of many CNVs, it is expected that genetic distance as estimated using CNVs will reflect the known relationships between individual mice. If somatic mosaicism exists within individuals, it is expected that tissues within a given individual will have an estimated genetic distance greater than zero despite being derived from the same zygote. The analysis of CNVRs identified across samples will identify putative genomic regions of recurrent tissue-specific CNC formation.

Aim 1: Identify and characterize copy number events with three calling methods for the Mouse Diversity Genotyping Array

To date there has been no genome-wide analysis of CNVs using the MDGA, in part due

to the relatively recent development and novelty of the MDGA, and because there are few methods capable of handling the unique design of this array. Before analysis of somatic copy number variation can take place, it is the first goal of this study to develop and compare three independent approaches for detecting putative copy number events using the SNP probes on MDGA. I will develop the "ConsecN" calling method, which uses runs of consecutive "No-Call" genotypes identified in the genotyping results. This method will be used for the first time to detect putative deletions in the absence of other CNV detection methods for this array. The second method will be the application of HMM Region Detection available in Partek® Genomics Suite™. The third method for copy number detection will be the HMM-based algorithm used by PennCNV. The data collected through these three methods will be compared to assess concordance between independent CNV detection methods. Characteristics of events including event length and GC content will be assessed for each calling method to determine relative biases in copy number calls.

Aim 2: Develop an accessible resource of previously discovered murine CNVs

In the absence of an accessible and interactive database of known CNVs in the mouse genome, I will complete a literature survey and gather data on previously detected CNVs and CNVRs reported for different mouse strains and tissues. I will compile these data into a format that can be uploaded into a publicly available genome browser for further analysis. I will then assess each of the CNV calling methods used in the present study for any overlap with entries documented in the newly constructed database as well as identify previously found CNVs and CNVRs that the methods in this study are able to detect.

Aim 3: Determine if the putative events detected by each method capture the known relatedness between samples

I will assess the degree to which CNV events called with each of the three CNV calling methods capture the known relatedness between samples. To do this, I will develop a method

of estimating genetic distance between any two samples based on CNV data collected for each sample. Using this measure of genetic distance, I will generate phylogenetic-style trees from copy number data collected by each of the three calling methods, and compare these trees to the known relationships between samples documented in a pedigree.

Aim 4: Assess tissue-specific differences in copy number for evidence of somatic copy number mosaicism

Events will be analyzed for overlap across samples to distinguish singletons and merges. The tissue-specific nature of events will be assessed by performing pairwise analyses between tissue samples taken from the same mouse, identifying putative CNVs and putative CNCs. Across mice, putative CNCs occurring in merges will shed light on genomic regions prone to harbouring CNCs in a tissue-specific manner.

Chapter 2

Materials and Methods

2.1 Experimental design

2.1.1 Selecting related mice bred on two inbred mouse strain genetic backgrounds

All protocols were approved by The Canadian Council on Animal Care (Appendix A). Eight related mice from the breeding colony maintained by our laboratory at the University of Western Ontario were selected for analysis and denoted "Hill Laboratory samples". Each mouse was individually identified by a numeric label following an in-house numbering system specific to our laboratory, and the known relationships between the mice based on in-house breeding records were documented in a pedigree (Figure 2.1).[200] Founder mice (mice 300.6, 900.3, 300.7, and 900.5) were originally obtained from The Jackson Laboratory (Bar Harbor, ME). Mice 300.6 and 300.7 were bred on a pure C57BL/6J inbred strain background. Mice 900.3 and 900.5 were F1 hybrids between inbred mice from strains C57BL/6J and CBA/CaJ. The selected mice, in addition to founders 300.6 and 900.3, were chosen to include three pairs

of brothers (904.11 and 904.9, 911.49 and 911.50, and 911.143 and 911.148) that have varying expected genetic background ratios between C57BL/6J and CBA/CaJ, calculated based on their pedigree and documented in Table 2.1. The first pair of brothers (904.9 and 904.11) are the offspring of mouse 300.6 and mouse 900.3 and are expected to be 75% C57BL/6J and 25% CBA/CaJ. The second pair of brothers (911.49 and 911.50) are third-generation offspring of mouse 904.9 and are expected to be 7% C57BL/6J and 93% CBA/CaJ. The third pair of brothers (911.143 and 911.148) are third cousins to 911.49 and 911.50 and are also expected to be 7% C57BL/6J and 93% CBA/CaJ.

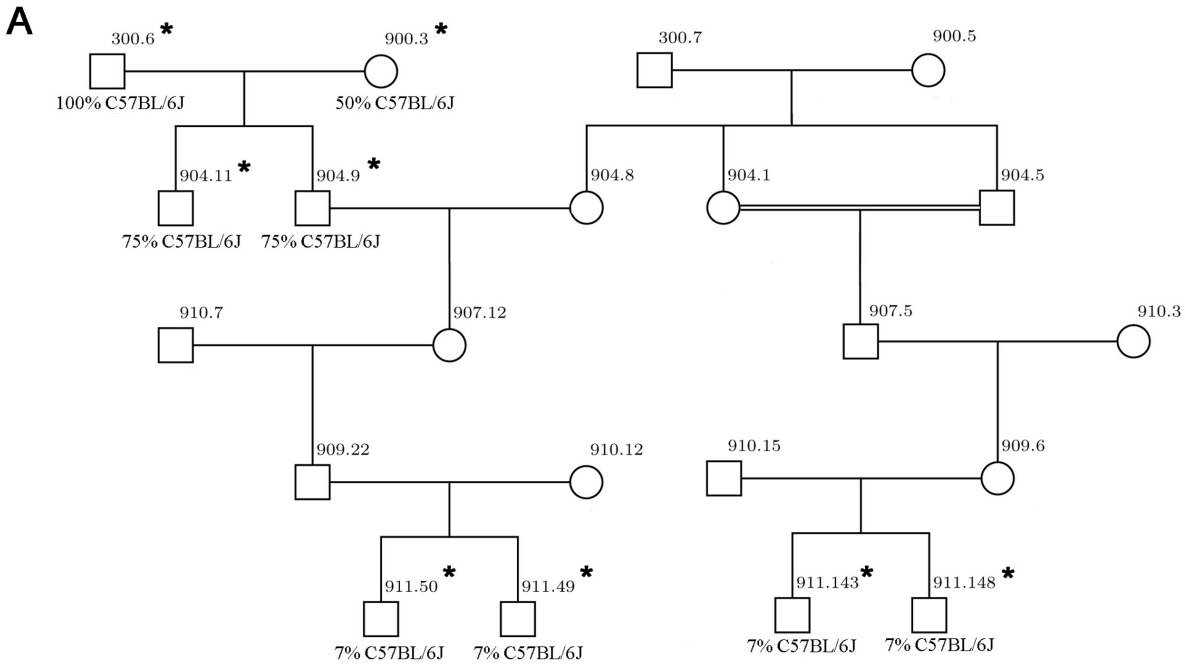
2.1.2 Selecting paired somatic tissues for detecting somatic copy number mosaicism

Wherever possible, two tissues from each mouse were selected to enable between-tissue comparisons (Table 2.1). Both the spleen (SP) and cerebellum (CL) were selected for mice 300.6, 904.9, 904.11, 911.49, and 911.50. Replicates of both tissues from mouse 911.50 (denoted by *rpt*) were used to assess reproducibility. These replicates represent different sections from the same tissue. For mouse 900.3, spleen and liver (LI) were used due to the unavailability of the cerebellum. For the remaining pair of brothers (911.143 and 911.148), only the cerebellum was available for each mouse. The three mice for which only one tissue was available were excluded from between-tissue analyses.

2.2 Obtaining raw data from the Mouse Diversity Genotyping Array associated files

A description of DNA extraction and hybridization procedures is described in Appendix B (Supplementary Methods). All analyses presented in this thesis are based on data obtained from the Mouse Diversity Genotyping Array (MDGA; Affymterix®, Santa Clara, CA and The

FIGURE 2.1: **Pedigree of highly related mice bred on genetic backgrounds of two inbred strains.** **(A)** Mice used in this study are indicated by asterisks (*). The expected percentage of C57BL/6J genetic background is indicated below the representation of each mouse in the pedigree. The remaining percentage (if any) is assumed to be from CBA/CaJ background based on breeding records. **(B)** Mice used in this study are plotted on a scale that ranges between pure C57BL/6J genetic background and pure CBA/CaJ genetic background, based on breeding records.


A
300.6 *
100% C57BL/6J
900.3 *
50% C57BL/6J
300.7
900.5
904.11 *
75% C57BL/6J
904.9 *
75% C57BL/6J
904.8
904.1
904.5
910.7
907.12
907.5
910.3
909.22
910.12
910.15
909.6
911.50 *
7% C57BL/6J
911.49 *
7% C57BL/6J
911.143 *
7% C57BL/6J
911.148 *
7% C57BL/6J

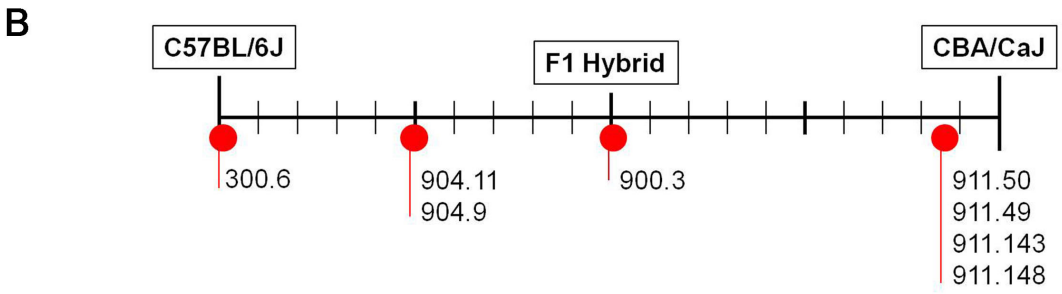

B
C57BL/6J
F1 Hybrid
CBA/CaJ
300.6
904.11
904.9
900.3
911.50
911.49
911.143
911.148

TABLE 2.1: Origin and genetic background of the sixteen Hill Laboratory tissue samples analyzed with the MDGA.

Sample ID[a]	Tissue Type	Sex[b]	Mouse Age (months)	Ratio of C57BL/6J:CBA/CaJ
300.6SP	Spleen	M	10.4	100:0
300.6CL	Cerebellum	M	10.4	100:0
900.3SP	Spleen	F	11.3	50:50
900.3LI	Liver	F	11.3	50:50
904.11SP	Spleen	M	7.7	75:25
904.11CL	Cerebellum	M	7.7	75:25
904.9SP	Spleen	M	7.7	75:25
904.9CL	Cerebellum	M	7.7	75:25
911.50SP	Spleen	M	8.7	7:93
911.50CL	Cerebellum	M	8.7	7:93
911.50SPrpt	Spleen	M	8.7	7:93
911.50CLrpt	Cerebellum	M	8.7	7:93
911.49SP	Spleen	M	8.7	7:93
911.49CL	Cerebellum	M	8.7	7:93
911.143CL	Cerebellum	M	15.2	7:93
911.148CL	Cerebellum	M	15.2	7:93

[a] Alphanumeric identifier for each sample. Numeral identifies the mouse, letters identify tissue (SP=spleen, CL=cerebellum, LI=liver), *rpt* identifies a tissue replicate

[b] M = male, F = female

Jackson Laboratory, Bar Harbor, ME). The raw data obtained from each array were in the form of a CEL file (with file extension .CEL). CEL files contain the fluorescence intensity value in relative fluorescent units (RFUs) for each array feature represented in the array image acquired from optically scanning the physical array. CEL files were obtained for analysis as described in Appendix B (Supplementary Methods).

The Center for Genome Dynamics (CGD) at The Jackson Laboratory (Bar Harbor, ME) digitally provides through its website 351 CEL files obtained from the Center's in-house projects using the MDGA.[201] These CEL files were obtained from arrays hybridized with DNA extracted from the tail tissue of mice representing several categories of mice that differ in the degree of genetic diversity between mice. These categories range from low genetic diversity between mice (classical laboratory strains of *Mus musculus domesticus*), to higher genetic diversity between mice (including wild-derived strains, wild-caught mice, and mice of different subspecies such as *Mus musculus spretus*).[169] Genotyping work using the 351 CEL files completed in the Hill Laboratory defined a subset of 335 CGD CEL files that passed a minimum genotyping call rate of at least 97%, defined as at least 97% of SNPs being successfully assigned a genotype call (AA, AB, or BB).[202] For the work presented here, this subset of 335 CEL files (referred to as the 335 CGD set) was downloaded from the CGD website.

All associated MDGA files and subsequent analyses were based on the most current *Mus musculus* reference sequence (NCBI Build 37, UCSC version mm9), as this was the reference for which the MDGA probes were designed. Associated files included a list of 526,162 SNPs represented on the MDGA selected for strict probeset design for optimal performance in genotyping and copy number analysis.[202] The MDGA library files (CD_MOUSEDIVm520650_rev2) and MDGA annotation files were downloaded from the Affymetrix® website.[172] FASTA files, sequence files, and cytoband files for UCSC Genome Browser mouse genome version mm9 (Golden Path) were downloaded from the UCSC Genome Browser database[203] and accessed locally.

2.3 Genotyping DNA samples hybridized to the Mouse Diversity Genotyping Array with a filtered SNP list

All samples hybridized to the Mouse Diversity Genotyping Array were genotyped using the corresponding CEL files with Affymetrix® Genotyping Console v. 4.1.2 (Affymetrix®) with the AGCC library files.[204] The CEL files for the sixteen samples were imported alongside the 335 CGD set. Gender information was imputed for each CEL file for computation of genotypes on the X and Y chromosomes. For the 526,162 autosomal, sex, and mitochondrial SNPs that were present in the filtered SNP list provided by S. Eitutis,[202] SNP genotyping was performed using the BRLMM-P algorithm (default settings, confidence score threshold = 0.1).[181,202] Samples that had genotype calls assigned to at least 97% of SNP loci were considered to be passing samples and were retained for analysis. Samples that had genotyping call rates below 97% were identified as failing the genotyping step, and these samples were removed from further analysis.[205] The call rate for sample 300.6SP failed the 97% threshold and was removed from further analysis. The passing samples were genotyped a second time. After the second genotyping round, all SNPs that did not have a genotyping call rate of at least 97% were removed. This removal step resulted in 470,339 remaining SNPs that were genotyped with an overall call rate of at least 97%. Affymetrix® Genotyping Console permits visualization of how each SNP is genotyped for a given set of CEL files by graphing the log ratio for each sample against the strength of the fluorescence intensity. For a sample to be genotyped successfully at a particular SNP, the probesets interrogating the SNP must have sufficient fluorescence intensity to be clustered, as well as have a silhouette score above the predetermined threshold (here, the silhouette score must be greater than 0.1).[181,182]

Of these 470,339 SNPs, the SNPs that map to the autosomes were selected to create a final filtered list containing 451,538 autosomal SNPs. The filtered autosomal SNP list was used for all further analysis. The genomic positions of the SNPs were plotted across the 19 autosomes

and visualized using Circos (Canada's Michael Smith Genomic Sciences Centre, Vancouver, BC; Figure 2.2).[206] The average distance between SNPs on the autosomes was calculated to be 5.3 kb ± 0.074 SEM. The smallest distance between two SNPs was 12 bp, and this was the smallest inter-SNP distance noted for all nineteen autosomes. The largest inter-SNP distance was 7.5 Mb and is evident by a visible gap in SNP coverage on chromosome 7.

2.4 Detecting putative copy number variation using three approaches

Three methods were employed to independently call copy number events from CEL files obtained for the total of fifteen Hill Laboratory samples from eight mice (Figure 2.3). The three methods were ConsecN (genotyping-based), Partek® Genomics Suite™ (Hidden Markov Model-based), and PennCNV (also Hidden Markov Model-based).

2.4.1 Calling putative copy number events with the novel "ConsecN" approach

All analyses using genotyping "NoCalls" were performed using Microsoft® Excel™ 2010 (Microsoft®, Redmond, WA). The genotyping results from Affymetrix® Genotyping Console for the fifteen Hill Laboratory samples that passed the genotyping step were ordered by chromosome position. Within each sample, occurrences of two or more consecutive SNP loci that failed to return a genotype call (instead returned as "NoCall") were scored as a single putative copy number event. All ConsecN events were classified as deletions due to the inability to identify increased fluorescence intensity from genotyping calls alone. Event lengths were calculated by subtracting the genomic location of the last SNP in the event from the genomic location of the first SNP in the event.

FIGURE 2.2: **The distribution of SNP loci interrogated by the Mouse Diversity Genotyping Array across the 19 murine autosomes. Outer circle:** Ideogram approximates locations of bands observed on Geimsa-stained chromosomes. **Inner circle:** Individual SNP loci are plotted as red dots against the corresponding genomic location on the respective chromosome. Regions that are SNP-dense appear as red spikes. Gaps in SNP coverage are evident by areas on the inner circle lacking red dots. The black arrow next to chromosome 7 indicates the largest gap between SNPs found on the autosomes (7.5 Mb).

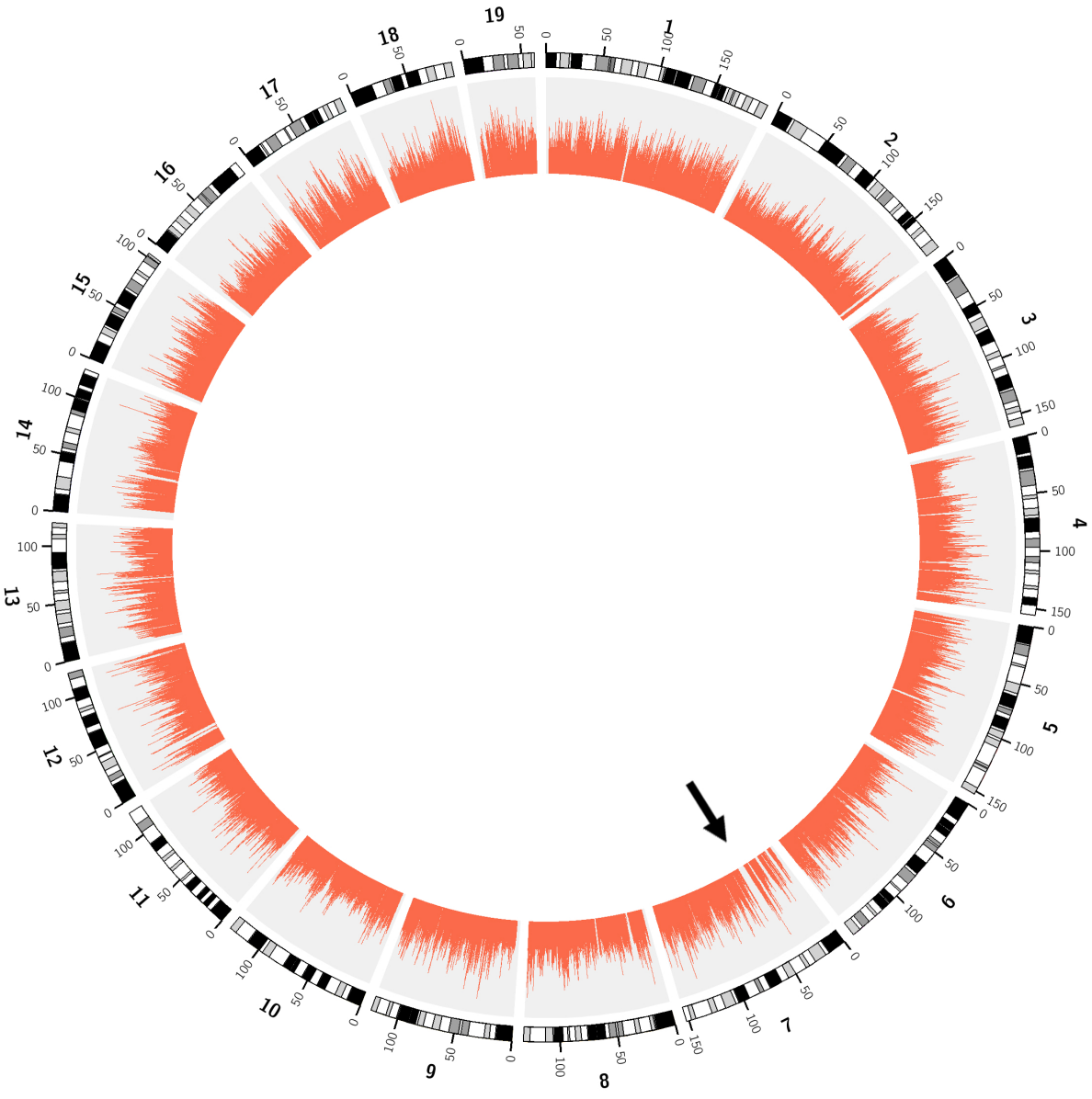
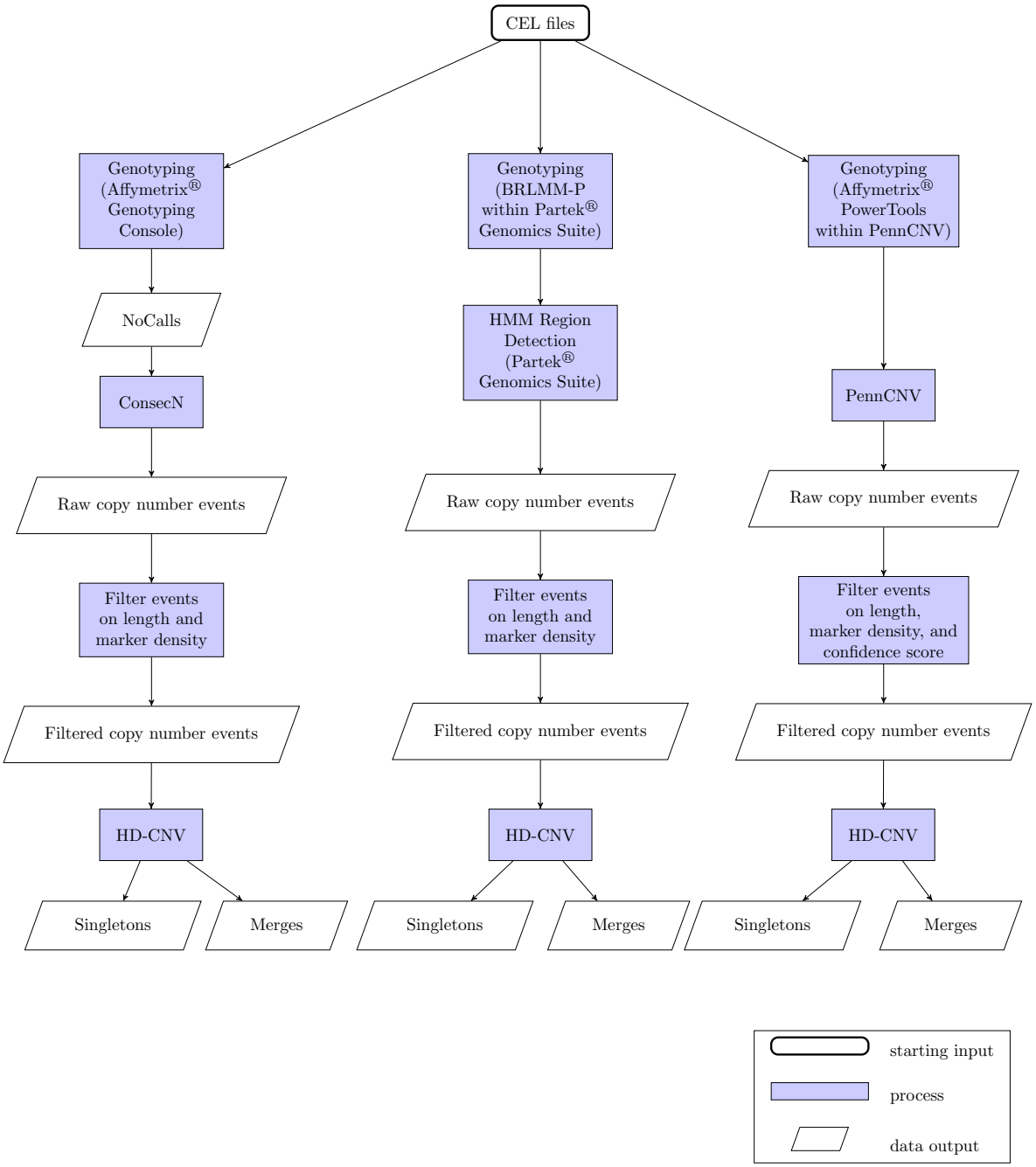

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

FIGURE 2.3: **Workflow to call copy number events with ConsecN, Partek, and PennCNV and identifying CNV events that overlap between samples using HD-CNV.** CNV events were called independently by ConsecN, Partek, and PennCNV. CNV events were then analyzed for overlap between samples with HD-CNV, which defines singletons and merges.


CEL files
Genotyping (Affymetrix® Genotyping Console)
Genotyping (BRLMM-P within Partek® Genomics Suite)
Genotyping (Affymetrix® PowerTools within PennCNV)
NoCalls
HMM Region Detection (Partek® Genomics Suite)
ConsecN
PennCNV
Raw copy number events
Raw copy number events
Raw copy number events
Filter events on length and marker density
Filter events on length and marker density
Filter events on length, marker density, and confidence score
Filtered copy number events
Filtered copy number events
Filtered copy number events
HD-CNV
HD-CNV
HD-CNV
Singletons
Merges
Singletons
Merges
Singletons
Merges
starting input
process
data output

2.4.2 Calculating the upper bound of the probability of the occurrence of two consecutive "NoCalls" by chance

The upper bound of the posterior probability of a consecutive run of an n number of consecutive SNPs returning "NoCall" can be calculated on a per-sample basis, given the known number of SNPs assayed by the MDGA and the number of "NoCalls" returned in total for that sample. This upper bound on the probability makes the assumption that the occurrence of the second "NoCall" in a run is dependent on the first "NoCall" in the run only inasmuch as it affects the remaining number of SNPs to draw from. The probability of two consecutive "NoCalls" occurring in a sample A can be expressed as

$$P(\text{NoCall } \textit{and } \text{NoCall})_A = P(\text{NoCall})_A \times P(\text{NoCall } \textit{after } \text{NoCall})_A$$

where P is probability. The probability P of the first occurrence of a "NoCall" can be expressed as

$$P(\text{NoCall})_A = \frac{N_A}{L_A}$$

where N is the number of SNPs that failed to genotype and were returned as "NoCalls", and L is the total number of SNPs assayed by the array. The probability of the second "NoCall" immediately following the first is given as

$$P(\text{NoCall } \textit{after } \text{NoCall})_A = \frac{(N_A - 1)}{(T_A - 1)}$$

where the pool of remaining "No Calls" is decreased by 1 (N_A-*1*) and the pool of all remaining SNPs is decreased by 1 (T_A-*1*). The possible number of occurrences of two consecutive "NoCalls" was determined by dividing the number of "NoCalls" found in each sample by 2 (values were rounded down to the nearest whole number). The number of two consecutive "NoCalls" expected in each sample was estimated by multiplying the probability of observing

two consecutive "NoCalls" by the possible number of occurrences. These values reflect the occurrence of exactly two consecutive "NoCalls". Because ConsecN events are defined as two *or more* consecutive "NoCalls", the occurrence of a ConsecN event containing three or more consecutive "NoCalls" will have a lower probability and be expected to occur less frequently than the estimates described here.

2.4.3 Calling putative copy number events with Partek® Genomics Suite™

All analyses were carried out with Partek® Genomics Suite™ version 6.12.1227 (St Louis, MO), henceforth referred to as Partek. Library files were downloaded from the Affymetrix® website automatically from within the Partek software. Genomic properties were based on Golden Path reference files for *Mus musculus* UCSC genome version mm9 based on the C57BL/6J strain, downloaded manually from within Partek following prompts.

In the 335 CGD CEL file set, there are seven CEL files representing seven C57BL/6J mice (4 male, 3 female). These seven C57BL/6J CEL files (herein referred to as the pure C57BL/6J reference samples) were selected to construct a pure C57BL/6J reference dataset to which Hill Laboratory samples would be compared to call copy number events within Partek.

The raw CEL files from the fifteen Hill Laboratory samples were imported simultaneously with the pure C57BL/6J reference samples. Advanced import settings were selected to employ Robust Multichip Average (RMA) background correction.[207] Sequence and cytoband files were downloaded manually from the UCSC Genome Browser sequence and annotation downloads site and manually selected in Partek.[203] After import, the following sample attributes were specified: "Sample type" (specifying sample or reference), "tissue" (spleen, cerebellum, liver, or tail), and "C57BL/6J degree" (the proportion of C57BL/6J alleles expected given the mixed genetic background of some of the mice).

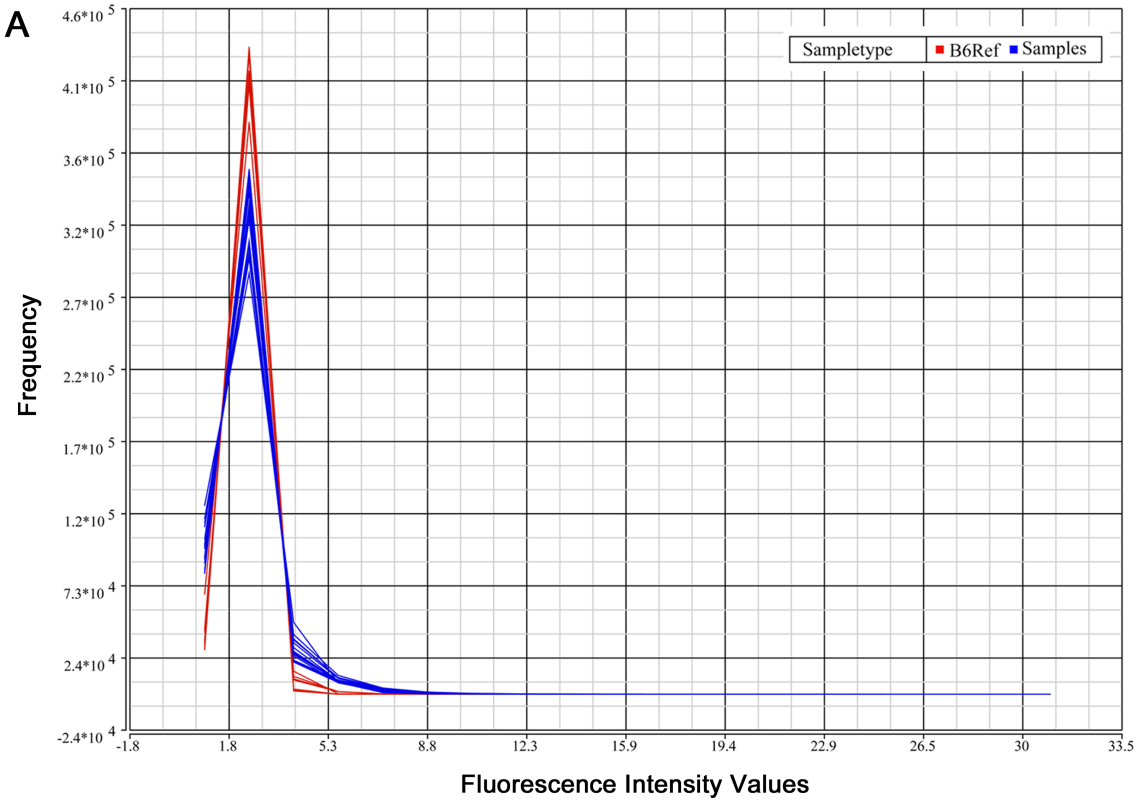
The full list of probes represented on the MDGA was filtered to include only the 451,538 autosomal SNP probes. Fluorescent intensities for these SNPs were adjusted for probe GC content and probe sequence in each CEL file.[180] The sample CEL files were normalized to the constructed baseline using the pure C57BL/6J reference samples as the control category. Quantile normalization was performed.[187] The control category was subsequently removed from the spreadsheet. The unpaired copy number operation (which allows all samples to be compared to a common reference) was selected to generate locus-level copy number from the allele-specific signal intensities documented in each CEL file. The locus-level copy number was summarized and pure C57BL/6J reference samples were selected to construct the reference.

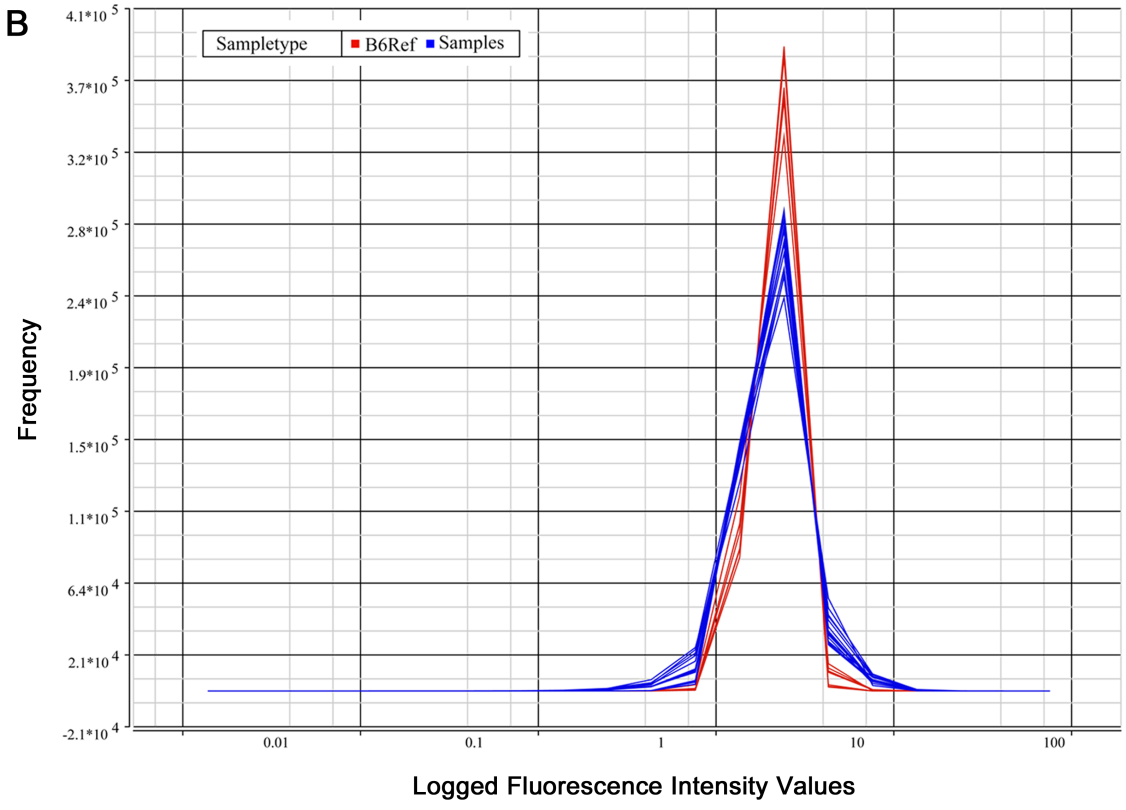
HMM (Hidden Markov Model) Region Detection was selected to identify regions of amplification and deletion across the genome. HMM Region Detection was run on the $\log_2$ ratios of the signal intensities, in accordance with Partek's in-software recommendation to use logged values if the histogram of the intensity values is more normally distributed in logged space versus unlogged space (Figure 2.4).[208] The following parameters were selected to remain at Partek default values: minimum genomic markers per region = 3;[60] maximum probability = 0.99;[209] sigma = 1;[210] genomic decay = 0 (disabled). Despite an extensive literature survey, no precedent for the optimal genomic decay parameter in calling mouse copy number seems to exist. Studies reporting the genomic decay parameters in human research are limited in number, and among those that do report genomic decay parameters, values anywhere between 10 and 10 million bp are given.[208–210]

2.4.4 Calling copy number with PennCNV

PennCNV's integrated genotyping step, which uses Affymetrix® PowerTools, requires at least 100 CEL files to be run together in order to cluster the genotypes correctly.[211] Using fewer CEL files would require a default clustering file to be used by the software, but the documentation warns that the resulting CNV calls may not be reliable.[211] For this reason, all

FIGURE 2.4: **Signal intensities of MDGA probesets are normally distributed when represented as $\log_2$ ratios.** Partek locus-level copy number data in **(A)** linear space and **(B)** logged space. "Sampletype" is a manually imputed category in which "B6Ref" denotes the seven pure C57BL/6J reference samples (downloaded from the Centre for Genome Dynamics,[201] and selected to construct the copy number reference in Partek), and "Samples" represents the fifteen experimental samples.


A
Sampletype B6Ref Samples
Frequency
4.6*10 5
4.1*10 5
3.6*10 5
3.2*10 5
2.7*10 5
2.2*10 5
1.7*10 5
1.2*10 5
7.3*10 4
2.4*10 4
-2.4*10 4
-1.8
1.8
5.3
8.8
12.3
15.9
19.4
22.9
26.5
30
33.5
Fluorescence Intensity Values


B
Sampletype B6Ref Samples
Frequency
4.1*10 5
3.7*10 5
3.2*10 5
2.8*10 5
2.4*10 5
1.9*10 5
1.5*10 5
1.1*10 5
6.4*10 4
2.1*10 4
-2.1*10 4
0.01
0.1
1
10
100
Logged Fluorescence Intensity Values

335 CGD CEL files (including the seven pure C57BL/6J reference sample CEL files used for Partek analysis) were run through PennCNV with the Hill Laboratory sample CEL files. PennCNV uses the files selected as a reference (in this case, the 335 CGD files) for reference calculations only, and does not output copy number for these files.[211]

The probe sets for the fifteen Hill Laboratory samples were summarized similarly to the Partek protocol using RMA background correction and median polish summary.[191] Unlike Partek, PennCNV does not offer full quantile normalization due to computing restraints, and sketch quantile normalization was employed instead.[212] The marker list was filtered to use only the 470,339 SNPs on autosomes and sex chromosomes identified by the SNP list provided by S. Eitutis.[202] PennCNV returns a "confidence score" for each event, which is an integer that indicates how likely it is for a CNV to occur in the region being reported, although the exact mathematical model used by PennCNV to calculate this score is not in the documentation.[213] Previous studies have used a minimum confidence score of 10 for events called by PennCNV, and the same cutoff value is employed here.[214,215] The raw list of copy number events was filtered to include autosomes only, thereby using only the 451,538 autosomal SNPs for further copy number analysis.

2.5 Analyzing putative copy number events by ConsecN, Partek, and PennCNV

The copy number events called by each of ConsecN, Partek, and PennCNV were analyzed independently but following identical workflows. Only events called on the nineteen autosomes were analyzed in this study due to a limitation in Partek regarding properly reading MDGA annotations for the sex and mitochondrial chromosomes.

2.5.1 Calculating lengths of copy number events

The lengths of ConsecN copy number events were calculated by subtracting the start position from the end position, where start position is the position of the first SNP locus calling an event and the end position is the position of the last SNP locus calling an event. Partek determines the start position of an event by taking the midpoint between the last marker (in this thesis, SNP locus) in the unchanged region before the event and the first marker in the changed region within the event. Partek calculates the end position in a similar manner, but where the end position is the midpoint between the last marker inside the event and the first marker after the event. PennCNV denotes the start and end position of a given copy number event as the first and last SNP locus in the event, respectively. The length originally returned in the PennCNV output is calculated by subtracting the coordinate of the last SNP from the coordinate of the first SNP and adding 1. In this analysis, length was recalculated for PennCNV events by only subtracting the position of the last SNP from the position of the first SNP; this was to match the length calculation for ConsecN to minimize between-method variation.

2.5.2 Filtering putative copy number events by marker density and variant length

To reduce the chance of falsely calling large variants with a relatively low number of markers, copy number events were filtered based on marker density. Marker density was calculated by dividing the number SNPs contained within the boundaries of the event by the length of the event in base pairs. Events with a marker density of fewer than 1.3 markers per 10 kb were removed from the dataset (Supplementary Table D.1). This cutoff was derived from the cutoff of 4 markers per 10 kb used by Pamphlett *et al.* for CNV analysis using the Affymetrix® Genome-Wide Human 6.0 Array.[60]

2.6 Analyzing copy number events that overlap between samples

Copy number event data from ConsecN, Partek, and PennCNV were imported independently into Microsoft® Excel™ 2010 and formatted according to the requirements for Hotspot Detector for Copy Number Variants (HD-CNV).[196] HD-CNV was run with the following parameters: minimum percent overlap for a merge = 40%, minimum percent overlap for a family = 99%, and reciprocal overlap selected. Reciprocal overlap reiterates the merging algorithm to merge together events that may not overlap themselves, but are both overlapped by a single larger event (Figure 2.5). Running HD-CNV with reciprocal overlap is beneficial when analyzing complex regions that may contain both large and small CNV events by preventing multiple merges forming in the same genomic region. This practice prevents overestimating the number of merges formed across samples by preventing double-counting of the same region.

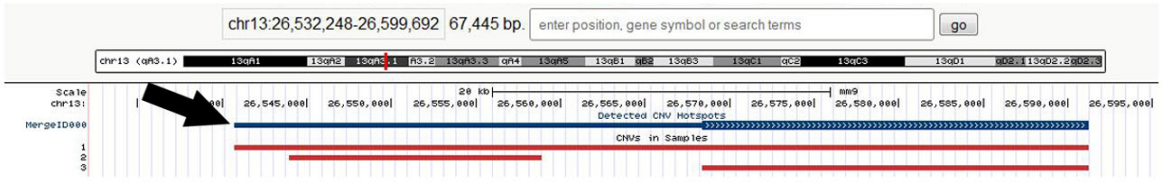
The output from HD-CNV comprises a set of two .csv files for each chromosome: one file containing data for each CNV event, and one file containing data for each merge that was formed from the data. Copy number event data were separated based on whether an event was included in a merge or not. Copy number events included in merges were referred to as "merge-associated events" (Figure 2.6). Merge-associated events, by definition, were found to overlap at least one other event in at least one other Hill Laboratory sample. Copy number events that were not included in any merge were referred to as "singletons", as they were not found to overlap with events called in any other Hill Laboratory sample.

2.6.1 Visualizing singletons and merges in UCSC Genome Browser

Optional output from HD-CNV includes a summary file, which contains genomic location information for individual CNV events as well as singletons and merges for all chromosomes. This summary file is a BEDdetail file.[198] BEDdetail files are formatted in such a way that multiple "tracks" can be documented in each file, which are UCSC Genome Browser's way

FIGURE 2.5: **Screenshot of UCSC Genome Browser with custom track display** showing a merge locus on chromosome 13 called with ConsecN method and HD-CNV **(A)** with reciprocal overlap selected, and **(B)** without reciprocal overlap selected. Thick black arrows illustrate the resulting merge or merges formed (blue lines) from the events found in the Hill Laboratory samples (red lines).

A With reciprocal overlap: 1 merge



B Without reciprocal overlap: 2 merges

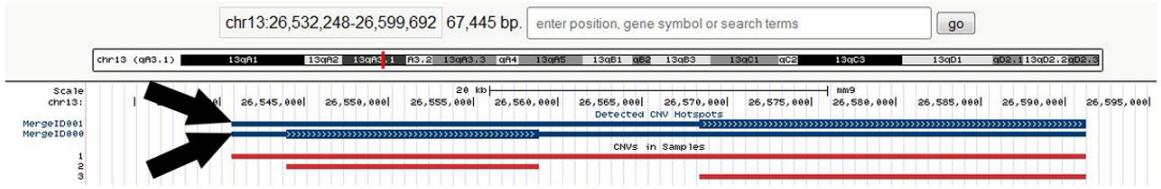
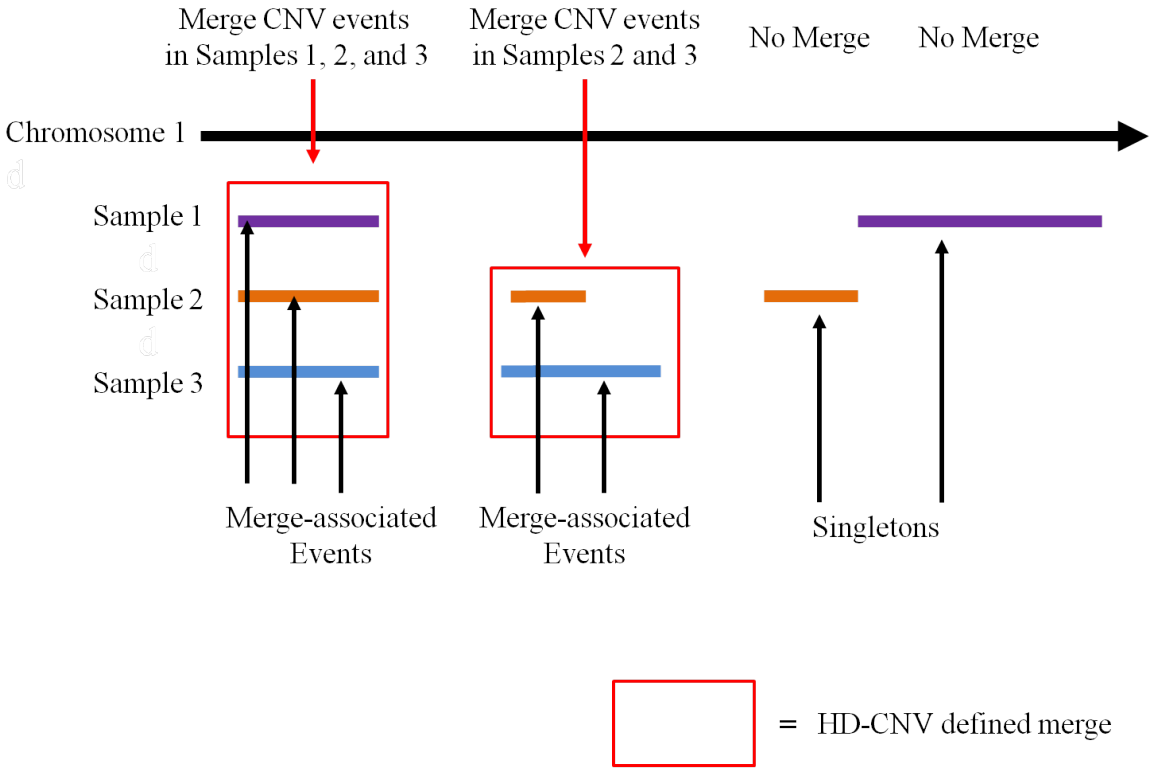


Figure 2.6: **HD-CNV groups copy number events in different samples as "merges" if they overlap in genomic location.** Copy number events that were included in a merge are defined here as merge-associated events. Copy number events that did not merge with any other events are defined here as singletons.


Merge CNV events
in Samples 1, 2, and 3
Merge CNV events
in Samples 2 and 3
No Merge
No Merge
Chromosome 1
Sample 1
Sample 2
Sample 3
Merge-associated
Events
Merge-associated
Events
Singletons
= HD-CNV defined merge

of grouping types of genomic information. For example, a user can select tracks that contain information about genes, SNPs, RNA transcripts, repeat elements, or many other datasets. Any one BEDdetail file can contain multiple defined tracks. BEDdetail files output from HD-CNV were uploaded to UCSC Genome Browser and visually inspected for correct merging.

2.6.2 Quantifying average GC content of copy number events

GC content values for singletons, merge-associated events, and merges called with each CNV calling method were calculated using the BEDTools command suite, which provides tools specifically for analyzing BEDdetail files.[216]

2.7 Evaluating concordance between events and merges defined from three CNV calling methods

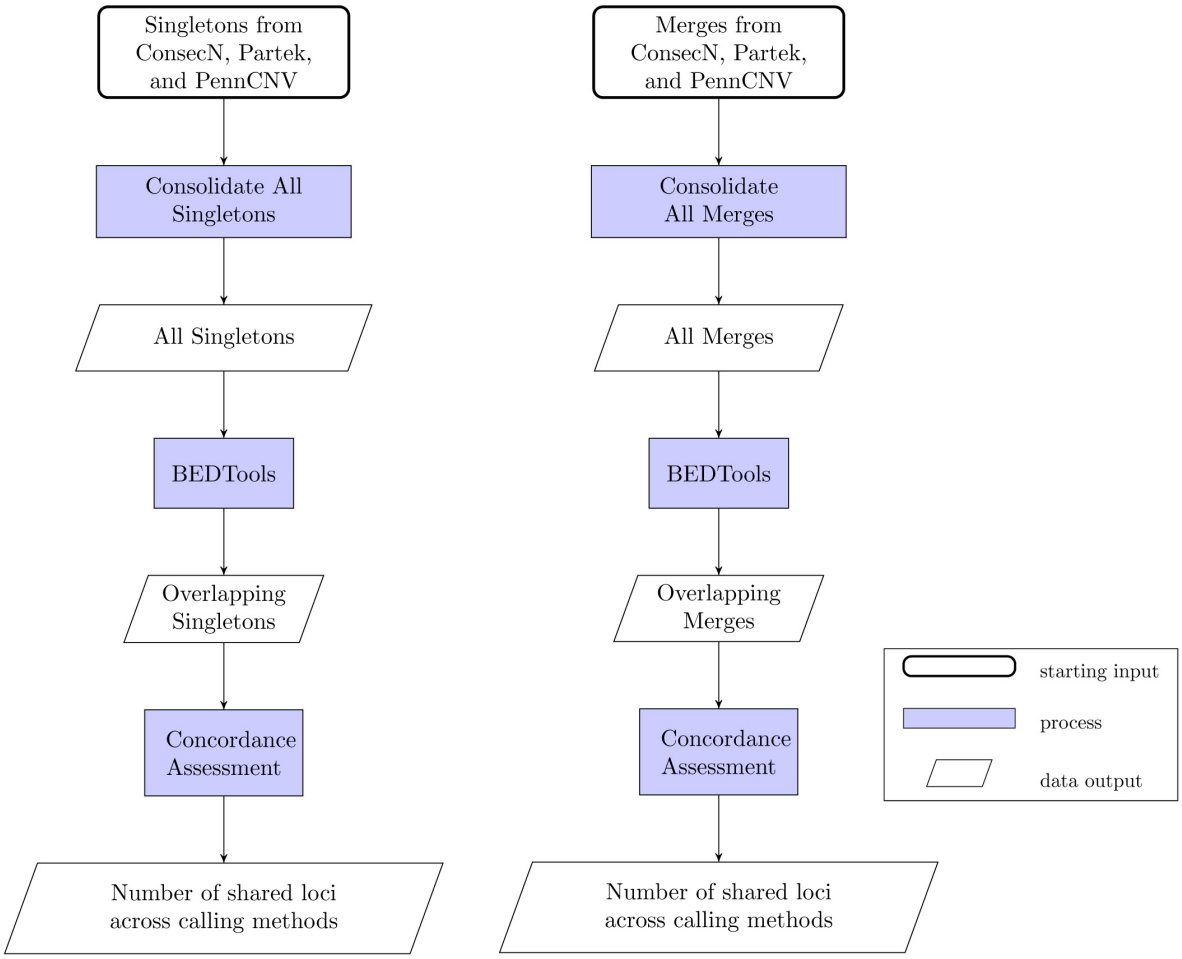
To analyze the concordance in CNV event calls between ConsecN, Partek, and PennCNV, singletons and merges defined by each method were compared with the other two calling methods for any overlap. Singletons and merges were analyzed separately (Figure 2.7).

2.7.1 Identifying singletons that overlap between CNV calling methods

Singletons defined by HD-CNV analysis for each CNV calling method were consolidated into a single file. The resulting three files were then cross-examined for any overlap between singleton loci identified by each of the three methods using the subcommand `multiIntersectBed` in BEDTools (an unpublished update to the original release; personal correspondence with A. Quinlan).[216] The minimum overlap between singletons was set to one base.[165]

Due to size discrepancies in singletons called by the different CNV event calling methods, it is possible for two or more singletons from one method to overlap with one larger singleton

FIGURE 2.7: **Workflow for assessing concordance of CNV calls between three CNV calling methods.** Singletons called by HD-CNV from ConsecN, Partek, and PennCNV were analyzed for overlap between methods using BEDtools. Merges called by HD-CNV from all three calling methods were also analyzed for overlap between methods using BEDtools.


Singletons from ConsecN, Partek, and PennCNV
Consolidate All Singletons
All Singletons
BEDTools
Overlapping Singletons
Concordance Assessment
Number of shared loci across calling methods
Merges from ConsecN, Partek, and PennCNV
Consolidate All Merges
All Merges
BEDTools
Overlapping Merges
Concordance Assessment
Number of shared loci across calling methods
starting input
process
data output

from another method. For this reason, the total genomic region containing the singletons found to overlap is referred to as an "overlapping singleton locus". If there was any overlap between singletons, the overlapping singleton locus boundaries were taken to be the outermost start and end positions from all singleton events involved in the overlap.

2.7.2 Identifying merges that overlap between CNV calling methods

Similar to overlapping singleton analysis, the merges defined by HD-CNV for each CNV calling method were consolidated into a single file for each method, which were then cross-examined for overlap using `multiIntersectBed` in BEDTools. Minimum overlap between merges was set to one base.[165] As with overlapping singleton analysis, size discrepancies in merges allowed the possibility for two or more merges from one CNV event calling method to overlap with one larger merge from another method. The genomic regions comprising merges that overlap with one another are referred to as "overlapping merge loci".

2.8 Estimating genetic distance using CNV calls

2.8.1 Calculating genetic distance based on the known pedigree of mice

The pedigree constructed from in-house breeding records was used to determine the coefficient of relatedness (r) between each sample. Tissues from the same individual were assigned an r value of one, based on the null hypothesis that tissues within an individual are isogenic (no somatic mosaicism). The r values between tissues taken from different mice were calculated based on the traditional between-individual relationships. The r values were then used to construct a distance matrix.

2.8.2 Calculating genetic distance with a modification to sharing analysis

Pairwise genetic distance was calculated using a modification to the sharing analysis approach first introduced by Lou *et al*.[139] The authors calculated the proportion of human CNVRs shared between different Chinese populations (sharing proportion, designated here as *SP*) as

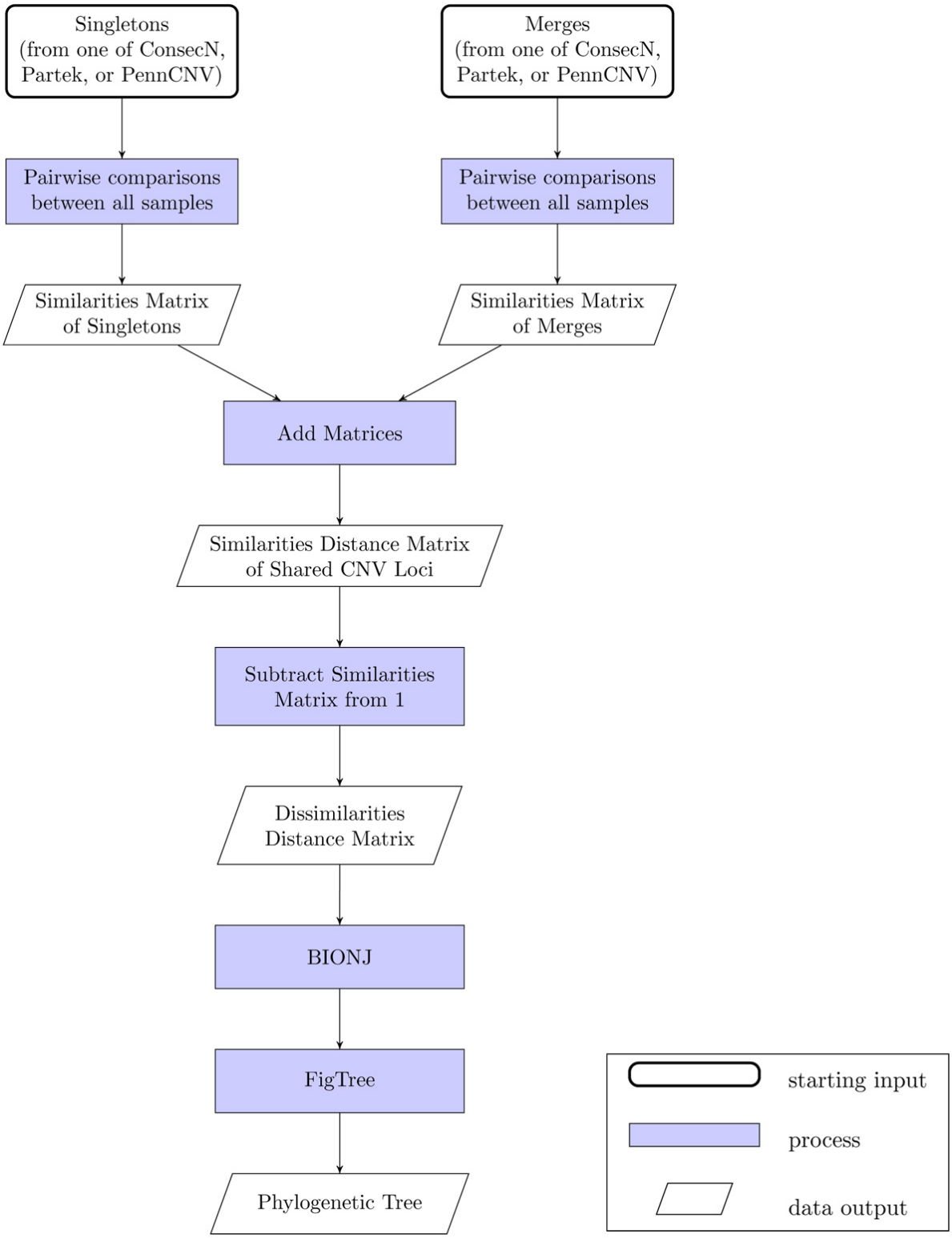
$$\mathrm{SP} = \frac{\mathrm{CNVRs}_{i,j}}{\mathrm{CNVRs}_{j}}$$

where $CNVRs_{i,j}$ is the average number of shared CNVRs between individuals from populations i and j, and $CNVRs_j$ is the total number of CNVRs detected in population j.

In the current study, genetic distance is calculated between various tissue samples taken from mice in one sample population, as opposed to genetic distance being calculated for averages across multiple populations as done by Lou *et al* (Figure 2.8). Since we are using the exact number of shared loci for each sample, we eliminate the need to average CNVR counts across populations. The denominator in the modified equation becomes the total number of CNV loci called with the respective CNV calling method (ConsecN, Partek, or PennCNV). This number is the sum of singleton loci and merge loci. Each CNV locus is treated as being analogous to a SNP locus, but instead of a single base pair position having two different alleles (A or B), the CNV locus has two different states (unchanged from the reference or changed from the reference as a gain or a loss).

Since a singleton by definition is present only in one sample from one mouse in the dataset, it is not necessary to determine how often it is shared with other samples, as the answer is always zero. Instead, a singleton is interpreted as being a CNV locus at which a single sample has the changed state, while all other samples share with each other the unchanged state. Merges, on the other hand, comprise events in different samples that overlap to some degree. We interpret a merge as occupying one locus, and so each merge is scored as such. Similar to

FIGURE 2.8: **Workflow for estimating genetic distance using shared CNV loci.** Singletons and merges identified by HD-CNV from copy number events called by one of the three calling methods (ConsecN, Partek, or PennCNV) were used to produce distance matrices based on shared CNV loci between each sample. Distance matrices were used to generate phylogenetic trees using the BIONJ algorithm. This workflow was repeated for copy number events from each of ConsecN, Partek, and PennCNV.


Singletons
(from one of ConsecN,
Partek, or PennCNV)
Merges
(from one of ConsecN,
Partek, or PennCNV)
Pairwise comparisons
between all samples
Pairwise comparisons
between all samples
Similarities Matrix
of Singletons
Similarities Matrix
of Merges
Add Matrices
Similarities Distance Matrix
of Shared CNV Loci
Subtract Similarities
Matrix from 1
Dissimilarities
Distance Matrix
BIONJ
FigTree
Phylogenetic Tree
starting input
process
data output

the scoring of singletons, a merge locus will have both a changed and an unchanged state, with the only difference from singletons being that multiple samples will share the changed state, as opposed to just one sample existing in the changed state.

For each locus called as a copy number event in at least one sample, each sample is compared pairwise with every other sample to score which samples share the same state. The sharing proportion *SP* is calculated as:

$$\mathrm{SP} = \frac{C_{i,j}}{(S + M)_n}$$

where the numerator is C number of shared CNV loci between samples i and j, and the denominator is the sum of singleton loci S and merge loci M present in the sample population n.

2.8.3 Constructing trees from genetic distances calculated from shared CNV loci

The total number of shared loci for each pairwise comparison between samples was recorded in a symmetrical matrix, in which the columns and rows represent each of the fifteen samples. Each value in this matrix was then divided by the total number of CNV loci found within the respective CNV calling method to generate a symmetrical similarity matrix. The dissimilarity matrix was then calculated by subtracting the similarity matrix from 1. Each dissimilarity matrix was then loaded into R (v. 2.15.0). Phylogeny estimation was accomplished with BIONJ, a modification to the neighbour-joining algorithm,[217] executed with the `BIONJ` command in the R package `ape` (Analysis of Phylogenetics and Evolution, v.3.0-8).[218] The output from `BIONJ` was read into the program FigTree (v.1.4.0; http://tree.bio.ed.ac.uk/software/figtree/) to produce graphical trees. All trees produced were unrooted.

2.8.4 Comparing trees to assess contribution of somatic copy number mosaicism to genetic distance

Two independent methods were used to compare the topologies of the three trees generated from genetic distance calculated by modified sharing proportion (ConsecN, Partek, and PennCNV) and the tree generated using the coefficient of relatedness *r* (Pedigree). The first comparison of tree topologies was made by counting the number of nodes separating all pairs of samples in a given tree. This node separation value was calculated pairwise between each sample. This was performed for each tree. For each tree, node separation values were placed in a matrix that followed a similar format to the matrices used to identify shared CNV loci. Tree comparisons were then made pairwise between all four trees by coupling values that occupied the same position in each matrix and performing Spearman's rank correlation. The second tree topology comparison was made using the sum of the branch lengths separating all pairs of samples in a given tree. For each tree, branch lengths were summed pairwise between all samples, and sums were placed in a matrix specific to that tree. Like node separation, between-tree comparison was made by coupling values that occupied the same position in each branch sums matrix. The correlation between paired values was analyzed by performing Pearson's product moment correlation for two trees at a time.

2.9 Constructing a local database of previously reported murine CNVs

To facilitate the comparison of copy number events called by each of the three CNV calling methods to murine CNVs previously reported in the literature, a local database was constructed. This database was constructed using supplementary data of the mouse CNV discovery studies included in the NCBI Database of Known Structural Variation (dbVar; http://www.ncbi.nlm.nih.gov/dbvar/ accessed April 2012)[197] which included genomic coordinates (chromosome, start position, and end position) for all copy number events (Supplementary Table C.1).

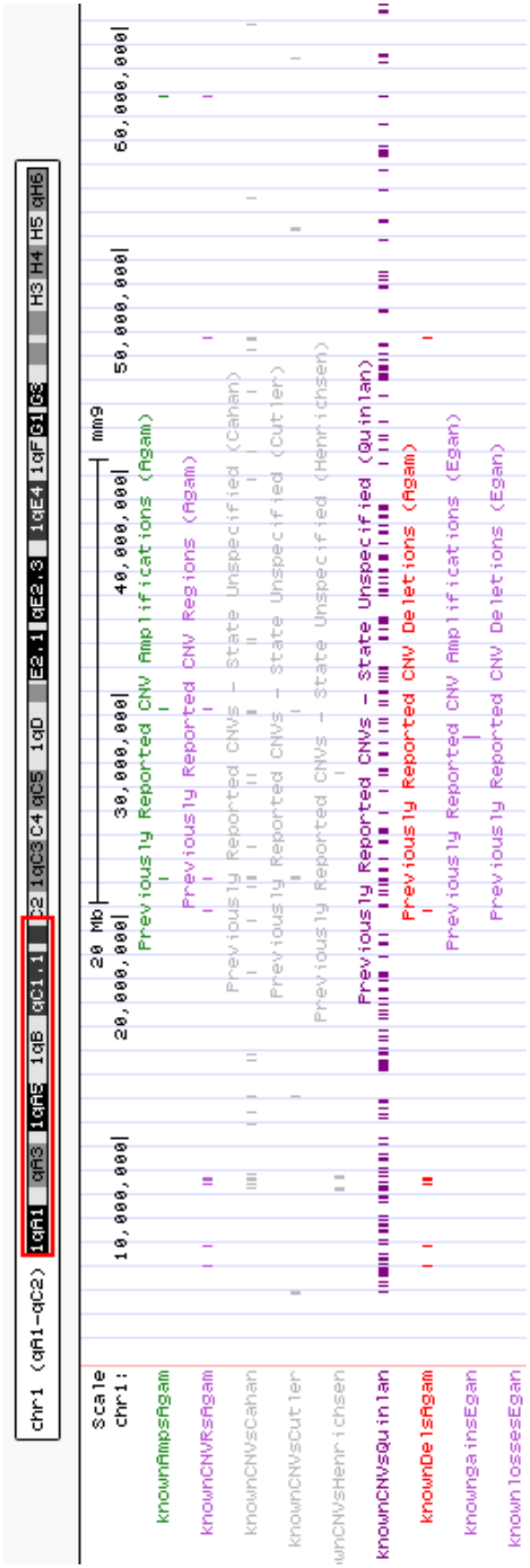
Genomic coordinates, mouse strain, and copy number state were used to create each entry in the database, with each entry representing either a CNV or a CNVR, depending on the information available in the given study. Database entries were formatted in BEDdetail format to allow allow construction of custom track files. Due to the large number of entries in the database (17,128 entries in total), each dataset was formatted as a separate track, with one track per BEDdetail file. The following custom tracks were constructed from previously published data (where *n* is number of entries in each track):

- Previously Reported CNV Amplifications (Agam *et al.*)[50] [*n*=499]
- Previously Reported CNV Deletions (Agam *et al.*)[50] [*n*=1477]
- Previously Reported CNV Regions (Agam *et al.*)[50] [*n*=783]
- Previously Reported CNV Amplifications (Egan *et al.*)[135] [*n*=38]
- Previously Reported CNV Deletions (Egan *et al.*)[135] [*n*=38]
- Previously Reported CNVs State Unspecified (Cahan *et al.*)[113] [*n*=1302]
- Previously Reported CNVs State Unspecified (Cutler *et al.*)[219] [*n*=1981]
- Previously Reported CNVs State Unspecified (Henrichsen *et al.*)[63] [*n*=1460]
- Previously Reported CNVs State Unspecified (Quinlan *et al.*)[92] [*n*=9550]

All individual tracks were uploaded to UCSC Genome Browser locally for browsing (Figure 2.9).

To investigate overlap between events called in Hill Laboratory samples and events detailed in the database, all custom tracks were incorporated into a single BEDdetail file. The `intersectBED` command in BEDTools (v.2.17.0) was used to find overlap between copy number events called in the present study and CNVs and CNVRs previously documented in the literature.[216] Database entries overlapping the copy number events from the present study were also scored for presence in C57BL/6J mice. CBA/CaJ mice were not represented in the studies included in the database and so CBA/CaJ copy number events could not be assessed in a strain-specific manner.

FIGURE 2.9: **Screen capture of custom tracks uploaded to UCSC Genome Browser showing database entries that map to chromosome 1 (1qa1-1qC2).** The custom tracks represent CNV gains, losses, unspecified copy number state, or CNVRs depending on the study from which data were summarized.


chr1 (qA1-qC2)
qA1 qA3 qA5 qB qC1.1 qC2 qC3 C4 qC5 qD E2.1 qE2.3 qE4 qF G1 G3 H3 H4 H5 qH6
Scale
chr1:
10,000,000
20 Mb
20,000,000
30,000,000
40,000,000
50,000,000
60,000,000
mm9
knownAmpsAgam
knownCNVsAgam
knownCNVsCahan
knownCNVsCutler
knownCNVsHenrichsen
knownCNVsQuinlan
knownDelsAgam
knowngainsEgan
knownlossesEgan
Previously Reported CNV Amplifications (Agam)
Previously Reported CNV Regions (Agam)
Previously Reported CNVs - State Unspecified (Cahan)
Previously Reported CNVs - State Unspecified (Cutler)
Previously Reported CNVs - State Unspecified (Henrichsen)
Previously Reported CNVs - State Unspecified (Quinlan)
Previously Reported CNV Deletions (Agam)
Previously Reported CNV Amplifications (Egan)
Previously Reported CNV Deletions (Egan)

2.10 Identifying putatively inherited CNVs and *de novo* CNCs

Only mice with two tissues available were included in the analysis of events shared between tissues (mice 904.9, 904.11, 911.49, 911.50, and 900.3). Tissue comparisons were made between tissues within individual mice (Figure 2.10). All tissue comparisons were automated using custom macro programs written in Visual Basic for Applications for Microsoft® Excel™ 2010.

As HD-CNV identified copy number events that were found to overlap between samples, only merges were analyzed to identify events shared between two tissues in the same mouse. Singletons were excluded from tissue comparisons because they were not found to overlap any other events by the minimum cutoff of 40%, and so were not considered to be shared between tissues. For each mouse, merges that contained events from both tissues were scored as containing copy number events shared between tissues and were termed "shared events" (Figure 2.11). As merges were scored on a mouse-by-mouse basis, it was possible for a given merge to be scored multiple times across mice. The start and end positions of two copy number events found to be shared between tissues were defined as the start and end positions of the merge in which both events occurred.

All singletons defined by HD-CNV were automatically scored as putative CNCs due to the fact they did not overlap with any other samples in the Hill Laboratory sample set, including the other tissue of the same mouse. Merges were analyzed on a mouse-by-mouse basis. Merges which contained an event from only one of the two tissues for a given mouse were scored as CNCs for the tissue containing the copy number event. For example, if Merge 1 contains an event in the spleen of mouse 904.9 but does not contain an event from the cerebellum of the same mouse, the sample 904.9SP (spleen) gets a score. As merges were scored on a mouse-by-mouse basis, it was possible for a given merge to be counted multiple times, once in each mouse that had a copy number event occur in the merge.

FIGURE 2.10: **Workflow for identifying putatively inherited CNVs and *de novo* CNCs. (A)** Singletons identified by HD-CNV from copy number events called by one of the three calling methods (ConsecN, Partek, or PennCNV) were analyzed to identify putative CNCs. **(B)** Merges identified by HD-CNV from copy number events called by one of the three calling methods (ConsecN, Partek, or PennCNV) were analyzed to identify putative tissue-specific CNCs and putative CNVs.

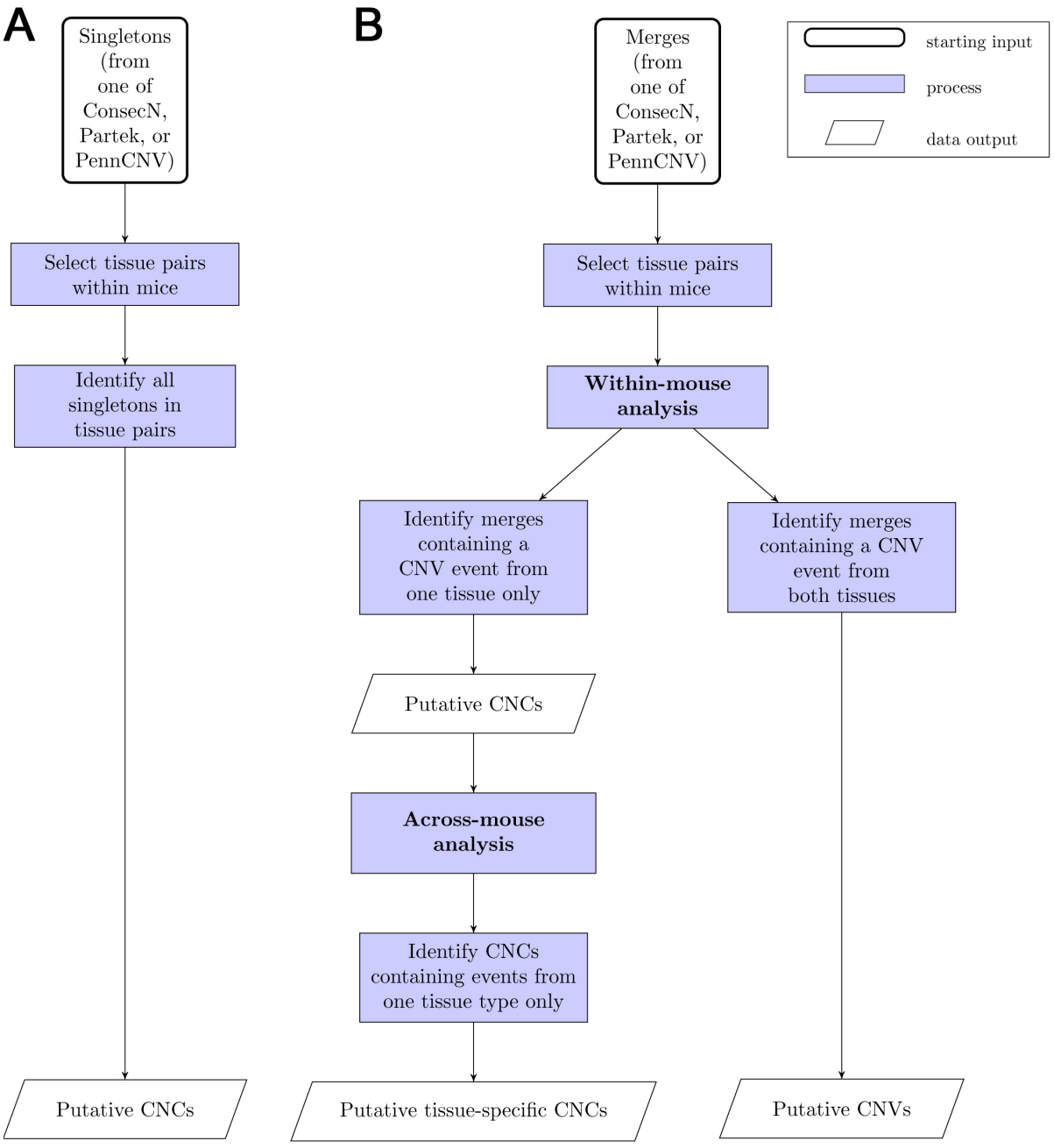
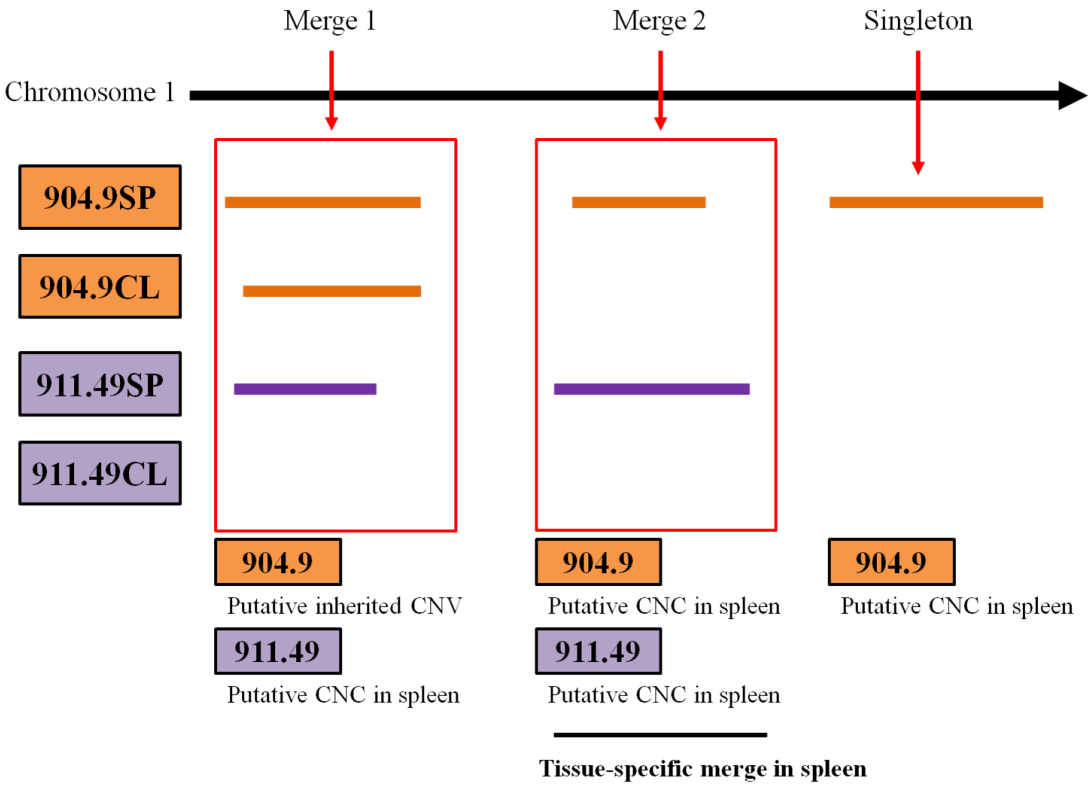

A
Singletons (from one of ConsecN, Partek, or PennCNV)
Select tissue pairs within mice
Identify all singletons in tissue pairs
Putative CNCs
B
Merges (from one of ConsecN, Partek, or PennCNV)
Select tissue pairs within mice
Within-mouse analysis
Identify merges containing a CNV event from one tissue only
Identify merges containing a CNV event from both tissues
Putative CNCs
Across-mouse analysis
Identify CNCs containing events from one tissue type only
Putative tissue-specific CNCs
Putative CNVs
starting input
process
data output

FIGURE 2.11: Illustration demonstrating the scoring of putative inherited CNVs, CNCs, and tissue-specific CNCs.


Merge 1
Merge 2
Singleton
Chromosome 1
904.9SP
904.9CL
911.49SP
911.49CL
904.9
Putative inherited CNV
911.49
Putative CNC in spleen
904.9
Putative CNC in spleen
911.49
Putative CNC in spleen
904.9
Putative CNC in spleen
Tissue-specific merge in spleen

Merges that contained events from only one tissue type were scored as putative "tissue-specific" CNCs. This scoring was done on a merge-by-merge basis. Merges that contained only events found in the spleen of the mice included in the tissue analysis (900.3, 904.9, 904.11, 911.49 and 911.50) were scored as spleen-only. Similarly, merges that contained only events found in the cerebellum among these mice were scored as cerebellum-only. Since there was only one liver sample available (mouse 900.3), sample 900.3LI was not included in tissue-specific CNC analysis (although 900.3SP was retained).

2.11 Statistical analyses

All statistics were performed in R (v.2.15.0).[220] The distributions of all one-dimensional data were tested for normality with the Lilliefors Kolmogorov-Smirnov test. Data presented in the results are normally distributed unless deviations from the normal distribution are noted. Genotyping call rates for the Hill Laboratory samples and the 335 CGD dataset were tested for significant differences between groups using the Kruskal-Wallis rank sums test. Comparisons of length and GC content were made between copy number events found with each CNV calling method (ConsecN, Partek, and PennCNV). Comparisons of length and GC content were also made between singletons and merges, as well as between gains and losses where applicable. Copy number event lengths and GC content were tested for significant differences using the Kruskal-Wallis rank sums test.

Dissimilarity distance matrices were compared using Mantel tests to determine if the matrices used to construct phylogenetic trees were different. Mantel tests were performed using the function `mantel.test` in the R package `ape` with 10,000 permutations. Topological comparisons between phylogenetic trees based on node separation were made using Spearman's rank correlation for discrete data. Comparisons between trees based on branch length sums were made using Pearson's product moment correlation for continuous data.

Chapter 3

Results

3.1 Fifteen Hill Laboratory samples passed the genotyping step

Of the sixteen Hill Laboratory samples that were genotyped from the MDGA using the filtered SNP list, fifteen samples passed the minimum genotyping call rate of 97% (Table 3.1). The sole sample that failed the minimum call rate was the spleen of mouse 300.6 (a C57BL/6J mouse, with a call rate of 96.26%). This sample was removed from all subsequent analysis, and no tissue comparisons were made for this individual. The distribution of genotyping call rates of the Hill Laboratory samples was skewed left, with higher call rates being more frequent ($D=0.1554$, $p<0.001$). The median genotyping call rate for the Hill Laboratory samples was 99.93%. The genotyping call rates of the Hill Laboratory samples were not found to be significantly different from the 335 CGD set call rates (median 99.66%; $H=108.9781$, 123 d.f., $p=0.813$).

Of the fifteen Hill Laboratory samples that passed genotyping, the lowest genotyping call rate was for sample 900.3LI (98.99%) as it had the highest number of SNPs that failed to be successfully genotyped (4741 "NoCalls"). All other samples had genotyping call rates of over

TABLE 3.1: Genotyping call rates from the final round of genotyping for each of the Hill Laboratory samples hybridized to the Mouse Diversity Genotyping Array. Genotyping was performed using Affymetrix® Genotyping Console (v. 4.1.2).

	Number of SNPs Genotyped[a]				
Sample ID	**AA**	**BB**	**AB**	**NoCall**	**Overall Genotyping Call Rate (%)[b]**
300.6SP[c]	-	-	-	-	-
300.6CL	465149	4046	828	316	99.14
900.3SP	387447	29566	49141	4185	99.11
900.3LI	387071	29406	49121	4741	98.99
904.11SP	414254	369	55451	265	99.94
904.11CL	414275	370	55344	350	99.93
904.9SP	409587	404	60019	329	99.93
904.9CL	409488	406	60155	290	99.94
911.50SP	374542	66071	29588	138	99.97
911.50CL	374186	65908	29627	618	99.87
911.50SPrpt	373829	65856	29463	1191	99.75
911.50CLrpt	373167	65697	29645	1830	99.60
911.49SP	374493	59364	36215	267	99.94
911.49CL	374514	59360	36143	322	99.93
911.143CL	376520	60361	33246	212	99.96
911.148CL	376485	59535	34082	237	99.95

[a] Successful genotype calls are as follows: AA = homozygous for the A allele, BB = homozygous for the B allele, or AB = heterozygous (both A and B allele present). Failed genotype calls are indicated by *NoCall*

[b] Call rate is the percentage of successful genotype calls (AA, BB, or AB)

[c] This sample failed the first round of genotyping with a call rate of 96.34%, and was removed from subsequent genotyping rounds and CNV analysis

99.10%. The sample with the highest genotyping call rate (and therefore lowest number of "NoCalls") was sample 911.50SP, with a call rate of 99.97% and 138 "NoCalls".

3.2 Putative copy number events were called by ConsecN, Partek, and PennCNV

Putative copy number events were detected in the fifteen Hill Laboratory samples that passed genotyping. Data were filtered at multiple points in the analysis of MDGA data based on stringent criteria to minimize false positives (Figure 3.1). The resulting copy number events were characterized by copy number state (gain or loss), length, marker density, and GC content.

3.2.1 The copy number events called by each method differed in number and copy number state

All events found with the ConsecN method were classified as deletions. The ConsecN approach found 264 copy number events called in total on the autosomes for the fifteen Hill Laboratory samples (Table 3.2). This number was reduced to 239 events after filtering out events that had a marker density below the cutoff value of 1.3 markers per 10 kb. ConsecN events were called in thirteen of the fifteen samples. Neither tissue from mouse 904.9 had any ConsecN events called.

For the seven pure C57BL/6J CGD mice used to construct the reference copy number, Partek did not call any change in copy number. Initially, Partek called 242 autosomal copy number events in total for the fifteen Hill Laboratory samples. This number was reduced to 224 after filtering events with low marker density. Partek was the only method that called copy number gains, detecting 195 gains in total for the fifteen samples. Losses accounted for 29 of the copy number events called by Partek.

FIGURE 3.1: **Applying quality control measures at multiple steps of copy number analysis.** Quality control measures were applied for probe set selection, genotyping call rate, copy number event length, marker density, and confidence score (confidence score available for PennCNV only).

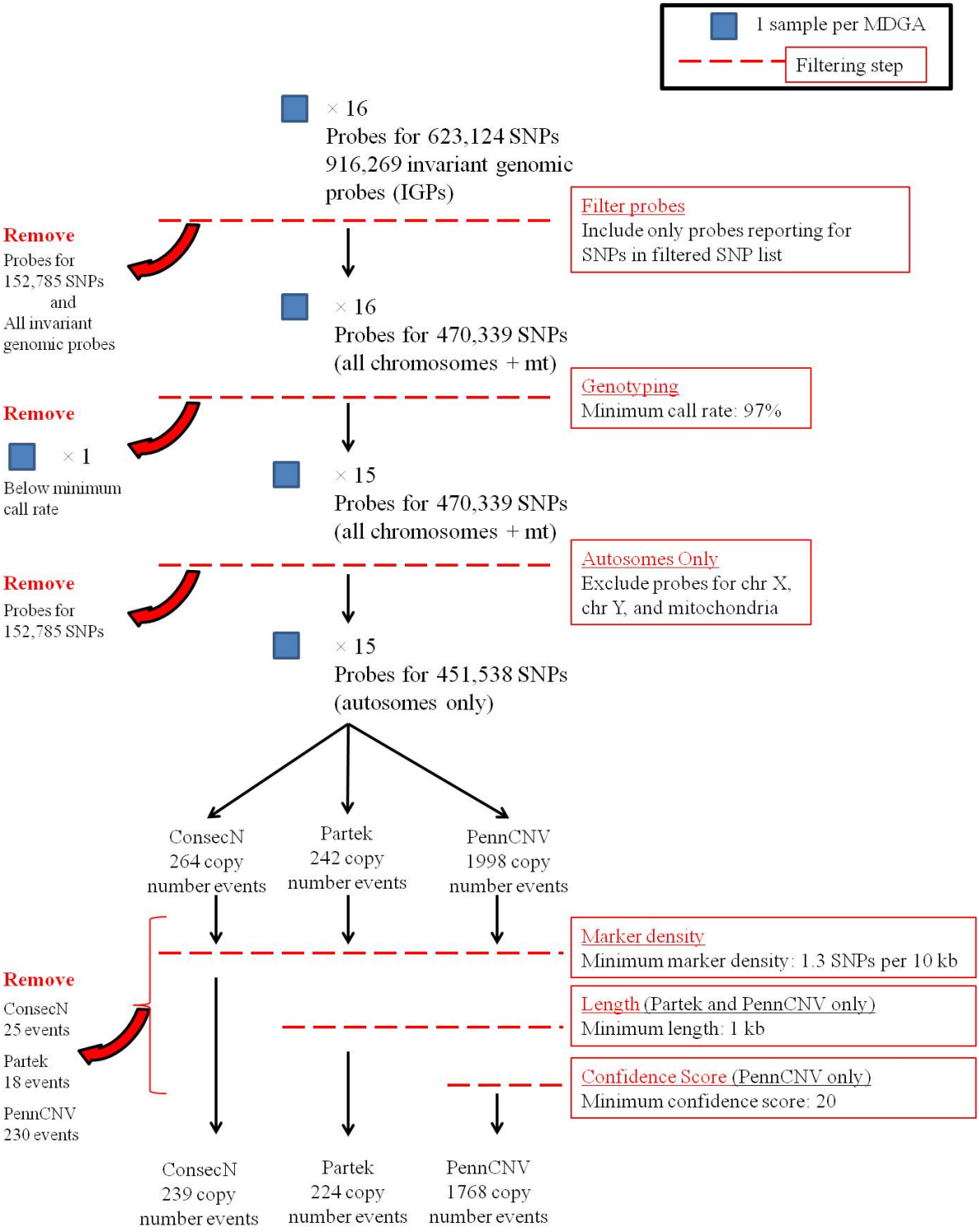

1 sample per MDGA
Filtering step
× 16
Probes for 623,124 SNPs
916,269 invariant genomic probes (IGPs)
Filter probes
Include only probes reporting for SNPs in filtered SNP list
Remove
Probes for 152,785 SNPs and All invariant genomic probes
× 16
Probes for 470,339 SNPs (all chromosomes + mt)
Genotyping
Minimum call rate: 97%
Remove
× 1
Below minimum call rate
× 15
Probes for 470,339 SNPs (all chromosomes + mt)
Autosomes Only
Exclude probes for chr X, chr Y, and mitochondria
Remove
Probes for 152,785 SNPs
× 15
Probes for 451,538 SNPs (autosomes only)
ConsecN 264 copy number events
Partek 242 copy number events
PennCNV 1998 copy number events
Marker density
Minimum marker density: 1.3 SNPs per 10 kb
Remove
ConsecN 25 events
Partek 18 events
PennCNV 230 events
Length (Partek and PennCNV only)
Minimum length: 1 kb
Confidence Score (PennCNV only)
Minimum confidence score: 20
ConsecN 239 copy number events
Partek 224 copy number events
PennCNV 1768 copy number events

TABLE 3.2: Number of copy number events in each sample as called independently by ConsecN, Partek, and PennCNV following filtering steps.

Sample ID	**ConsecN (Losses)**	**Partek (Gains)**	**Partek (Losses)**	**PennCNV (Losses)**
300.6SP	77	0	2	217
900.3SP	34	6	4	148
900.3LI	62	6	4	173
904.11SP	3	18	1	49
904.11CL	4	15	1	80
904.9SP	0	22	1	19
904.9CL	0	10	1	112
911.50SP	5	9	2	92
911.50CL	6	10	1	178
911.50SPrpt	10	10	2	59
911.50CLrpt	18	21	3	136
911.49SP	5	12	1	124
911.49CL	4	20	1	144
911.143CL	4	23	2	134
911.148CL	7	13	3	103
Total	239	195	29	1768

PennCNV called the most copy number events of the three methods, initially calling 1998 autosomal events in total for the fifteen Hill Laboratory samples. Filtering by marker density and confidence score reduced this number to 1768. PennCNV originally called both gains and losses, although only losses remained in the filtered dataset.

3.2.2 ConsecN called occurrences of two consecutive "NoCalls" more frequently than expected by chance

After genotyping SNPs, the occurrences of "NoCalls" in each sample were counted, and the probability of observing two "NoCalls" in a row was calculated for each sample. The expected occurrence of ConsecN events was less than 1 for all samples, given the number of "NoCalls". The probability of observing two consecutive "NoCalls" given the total number of "NoCalls" observed within a given sample was highest in sample 900.3LI. This sample had 4741 "NoCalls" returned, yielding an expected number of ConsecN events of 2.41×10^{-1} (Table 3.3). In this sample, 62 ConsecN events were observed (a fold difference of 2.57×10^{2}).

3.2.3 "NoCalls" do not result strictly from low fluorescence

Visual inspection of SNP clustering reveal variation in the fluorescence intensity of "NoCalls" (Figure 3.2). Some "NoCalls" result from low fluorescence intensity for a particular probeset, but not all "NoCalls" follow this pattern of low fluorescence intensity. Other "NoCalls" have fluorescence intensities comparable to successful genotype calls for the same SNP, but appear to fall between posterior clusters in such a way that they cannot be accurately called as either the heterozygous call or the homozygous call. These SNP loci are designated as "NoCalls" due to their low silhouette score, which assesses the quality of the clustering classification for each data point.

The only samples in which no ConsecN events were observed were the spleen and the cere-

Table 3.3: Upper bound estimate of the probability of obtaining two consecutive failed genotypes ("NoCalls").

Sample ID	Number of "NoCalls" observed in sample	Number of possible pairs of "NoCalls"[a]	Probability of Two Consecutive "NoCalls"[b]	Number of expected ConsecN events given observed "NoCalls"[c]	Number of observed ConsecN events	Fold difference between observed and expected
300.6SP	316	158	4.50×10^{-7}	7.11×10^{-5}	77	1.08×10^{6}
900.3SP	4185	2093	7.92×10^{-5}	1.66×10^{-1}	34	2.05×10^{2}
900.3LI	4741	2371	1.02×10^{-4}	2.41×10^{-1}	62	2.57×10^{2}
904.11SP	265	133	3.16×10^{-7}	4.19×10^{-5}	3	7.16×10^{4}
904.11CL	350	175	5.52×10^{-7}	9.66×10^{-5}	4	4.14×10^{4}
904.9SP	329	165	4.88×10^{-7}	8.02×10^{-5}	0	N/A[d]
904.9CL	290	145	3.79×10^{-7}	5.49×10^{-5}	0	N/A
911.50SP	138	69	8.55×10^{-8}	5.90×10^{-6}	5	8.48×10^{5}
911.50CL	618	309	1.72×10^{-6}	5.33×10^{-4}	6	1.13×10^{4}
911.50SPrpt	1191	596	6.41×10^{-6}	3.82×10^{-3}	10	2.62×10^{3}
911.50CLrpt	1830	915	1.51×10^{-5}	1.38×10^{-2}	18	1.30×10^{3}
911.49SP	267	134	3.21×10^{-7}	4.29×10^{-5}	5	1.17×10^{5}
911.49CL	322	161	4.67×10^{-7}	7.52×10^{-5}	4	5.32×10^{4}
911.143CL	212	106	2.02×10^{-7}	2.14×10^{-5}	4	1.87×10^{5}
911.148CL	237	119	2.53×10^{-7}	3.00×10^{-5}	7	2.34×10^{5}

[a] Given the number of NoCalls returned in the genotyping step for each sample. Rounded to the nearest whole number.

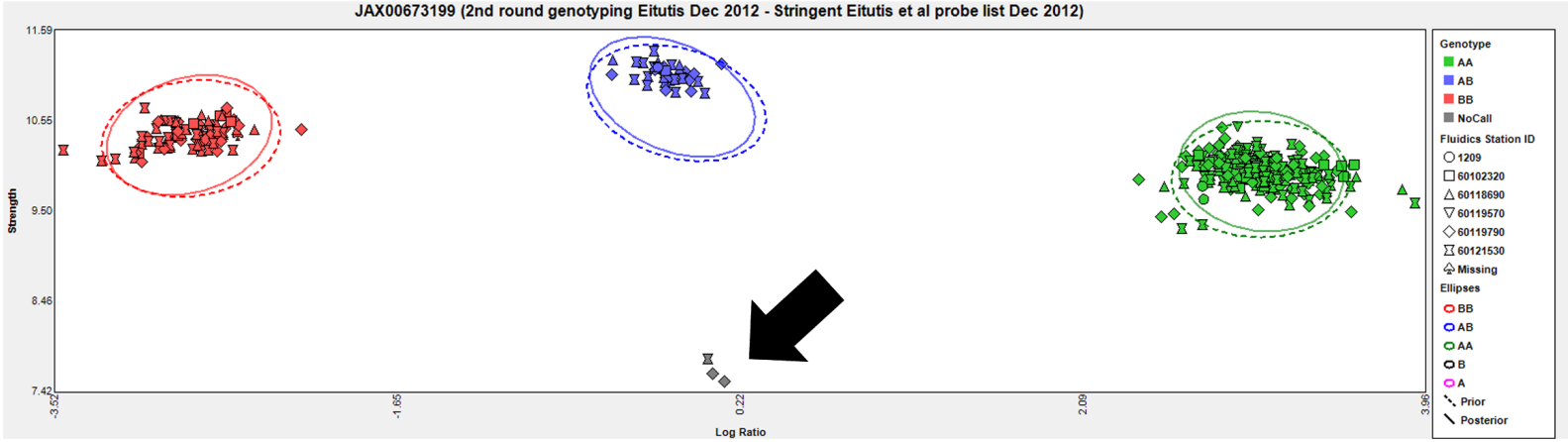
[b] $P(\text{NoCall } \textit{and } \text{NoCall})_A = P(\text{NoCall})_A \times P(\text{NoCall } \textit{after } \text{NoCall})_A$

[c] Number of possible pairs of "NoCalls" × Probability of Two Consecutive "NoCalls"

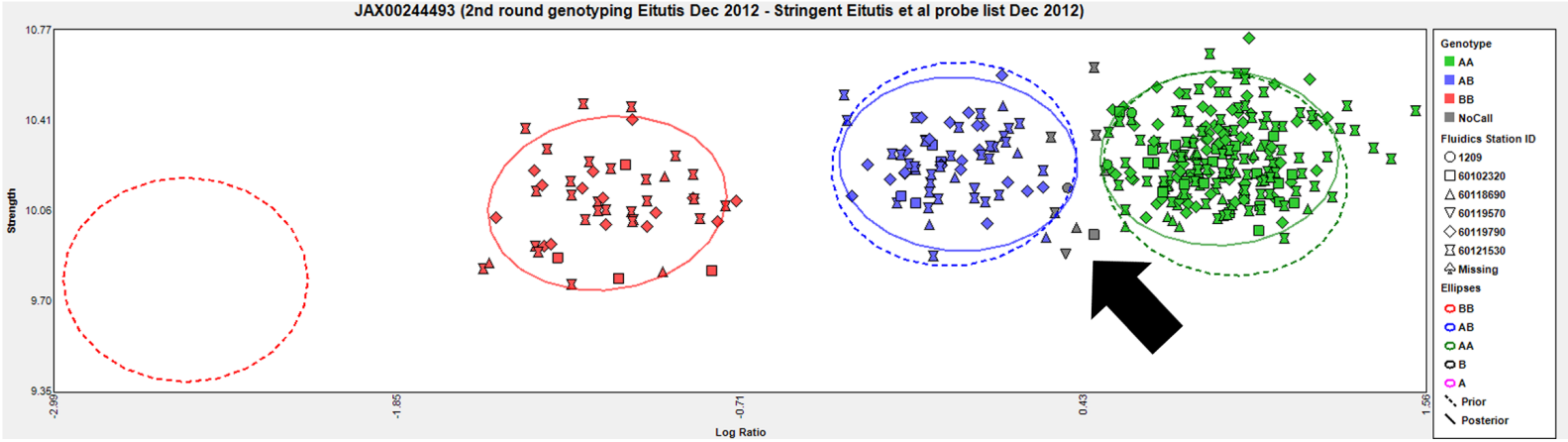
[d] Fold difference could not be determined in the absence of observed events

FIGURE 3.2: **A "NoCall" represents a SNP in a given sample that fails to cluster with the majority of other samples based on fluorescence intensity ratios for the two alleles at that locus.** Three clusters exist for the three genotype calls, homozygous AA (green), heterozygous AB (blue), and homozygous BB (red). Before genotyping, expected fluorescence intensities are clustered as prior clusters indicated by dotted lines. As the genotyping algorithm progresses, the actual fluorescence intensities from the CEL files are plotted and posterior clusters are indicated by solid lines. **(A)** "NoCalls" (grey) can result from probes that are returning fluorescence intensity but fall in between the clusters formed for heterozygous and homozygous calls. **(B)** "NoCalls" can also result from probes that do not return sufficient fluorescence intensity to be able to be clustered into a genotype call.

A


JAX00673199 (2nd round genotyping Eitutis Dec 2012 - Stringent Eitutis et al probe list Dec 2012)
Strength
11.59
10.55
9.50
8.46
7.42
-3.52
-1.65
0.22
2.09
3.96
Log Ratio
Genotype
AA
AB
BB
NoCall
Fluidics Station ID
1209
60102320
60118690
60119570
60119790
60121530
Missing
Ellipses
BB
AB
AA
B
A
Prior
Posterior

B


JAX00244493 (2nd round genotyping Eitutis Dec 2012 - Stringent Eitutis et al probe list Dec 2012)
Strength
10.77
10.41
10.06
9.70
9.35
-2.99
-1.85
-0.71
0.43
1.56
Log Ratio
Genotype
AA
AB
BB
NoCall
Fluidics Station ID
1209
60102320
60118690
60119570
60119790
60121530
Missing
Ellipses
BB
AB
AA
B
A
Prior
Posterior

bellum of mouse 904.9. The probabilities of observing two consecutive "NoCalls" in the genotyping results for the spleen and cerebellum of mouse 904.9 were 4.88×10^{-7} and 3.79×10^{-7}, respectively. Sample 300.6SP had the most ConsecN events called, with 77 events observed (316 "NoCalls", with a probability of detecting two consecutive "NoCalls" = 4.5×10^{-7}). Sample 300.6SP also had the largest fold difference between expected and observed ConsecN events (expected 7.11×10^{-5} ConsecN events, fold difference of 1.08×10^{6}).

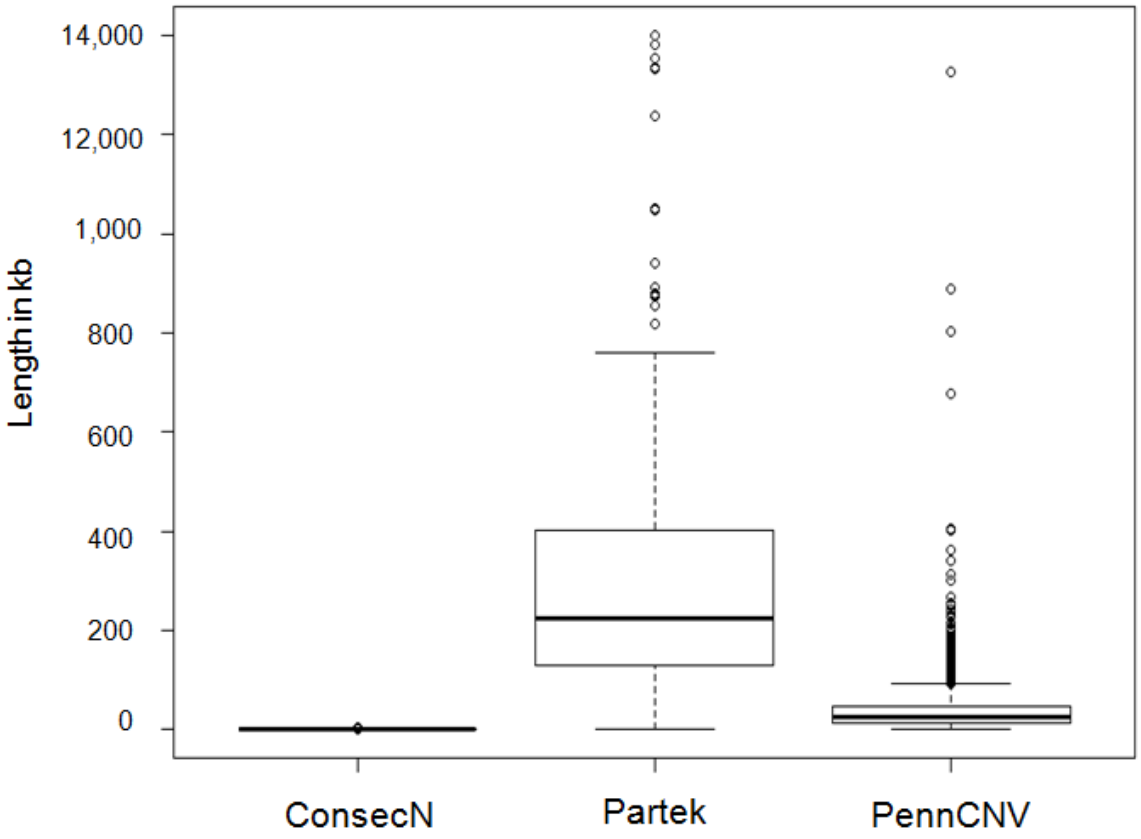
3.3 Characterization of copy number events called with ConsecN, Partek, and PennCNV

3.3.1 Each of the three CNV calling methods detected copy number events of different lengths

With gains and losses grouped together, the distributions of the lengths of copy number events for each of the three calling methods were skewed right, with smaller events more frequent ($p<0.001$). Event lengths were not similar across calling methods, with Partek calling longer events (median length of 2.2 Mb) and ConsecN calling shorter events (median length of 335 bp; Figure 3.3). The median length of PennCNV events was 2.49 kb.

The lengths of copy number events called by ConsecN ranged from 14 bp to 50,584 bp (both of these extreme values were called in sample 911.50CLrpt). The shortest event called by Partek was 5996 bp (911.148CL), while the largest was 14,018,633 bp (904.11CL). The shortest event called by PennCNV was 8,437 bp (911.50SP) and the largest was 13,280,763 bp (900.3SP). On average for each sample, the proportion of the genome affected by ConsecN events was 1.50×10^{-3}%, by Partek events 1.76%, and by PennCNV events 0.85%, based on the Golden Path length of the mouse genome.

FIGURE 3.3: **Spread of copy number event length data for events called by ConsecN, Partek, and PennCNV.** Partek called copy number events with the longest lengths of the three calling methods ($p<0.001$).


14,000
12,000
1,000
800
600
400
200
0
Length in kb
ConsecN
Partek
PennCNV

3.3.2 The marker density of copy number events is correlated with CNV length

The median marker densities for ConsecN, Partek, and PennCNV events were 59.70 markers per 10 kb, 1.87 markers per 10 kb, and 1.92 markers per 10 kb, respectively. The distributions of marker densities for each of the three calling methods were skewed right, with a few events in each method having higher numbers of SNPs for their length compared with the median ($p<0.001$ for each of ConsecN, Partek, and PennCNV). Event marker density was not equal across calling methods ($H=510.2211$, $df=2$, $p<0.001$), with the marker density of ConsecN events called higher than marker density of events called by either Partek or PennCNV ($p<0.001$; Figure 3.4). Events called by Partek and PennCNV did not differ in marker density ($W=198584$, $p=0.9332$). Marker density was highly correlated with event length in the putative CNVs detected with ConsecN ($R^2=0.9866$; Figure 3.5). The correlation between marker density and event length was weaker in the events detected with Partek ($R^2=0.2564$) and PennCNV ($R^2=0.3242$).

3.3.3 Copy number events called by ConsecN, Partek, and PennCNV differ in GC content

The GC content of copy number events was not the same across the three calling methods ($H=158.2928$, 1 d.f., $p<0.001$; Figure 3.6). ConsecN had the widest range of GC content percentages (22.50%-60.82%), while Partek events displayed the smallest range (34.56%-50.91%). ConsecN events also had the highest median GC content of the three calling methods, with a median GC content of 44.0%. PennCNV called events with the lowest GC content of the three calling methods (median=38.1%) The median GC content of the Partek events (41.15%) was the closest to the overall GC content of the mouse genome (42%).

FIGURE 3.4: **Spread of copy number event marker density for events called by ConsecN, Partek, and PennCNV.** ConsecN called copy number events with the highest marker density of the three calling methods ($p<0.001$).

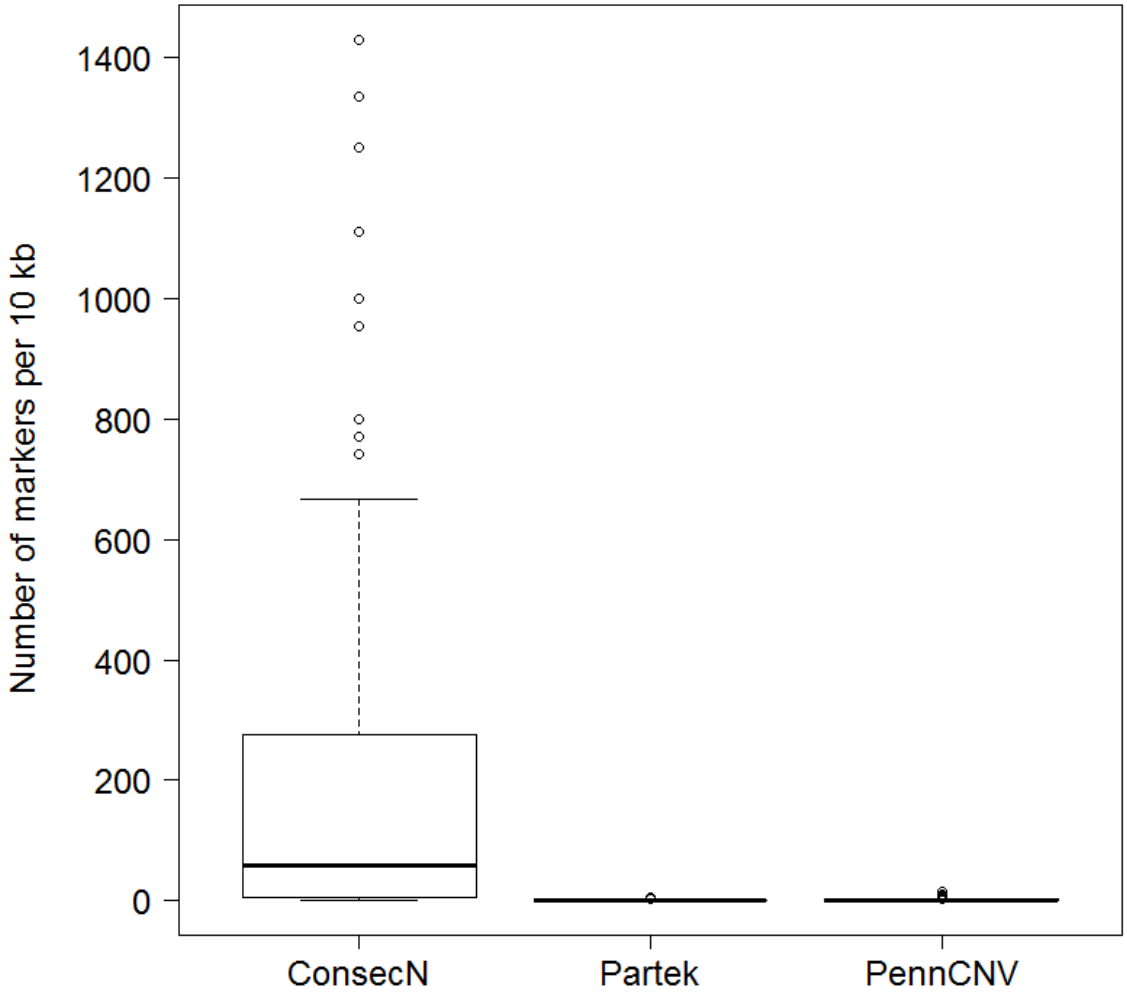

Number of markers per 10 kb
0
200
400
600
800
1000
1200
1400
ConsecN
Partek
PennCNV

FIGURE 3.5: **Marker density is correlated with copy number event length in putative CNVs detected using SNP probes on the Mouse Diversity Genotyping Array.** Copy number event marker density was calculated by dividing the number of SNP loci used to report each copy number event by the length of the event in base pairs. Logged values for event marker density are plotted against logged values of event length. **(A)** ConsecN events demonstrate a strong Pearson correlation (R^2=0.9866) between event marker density and length. **(B)** Partek events demonstrate a weak Pearson correlation (R^2=0.2564) between event marker density and length. **(C)** PennCNV events demonstrate a weak Pearson correlation (R^2=0.3242) between event marker density and length.

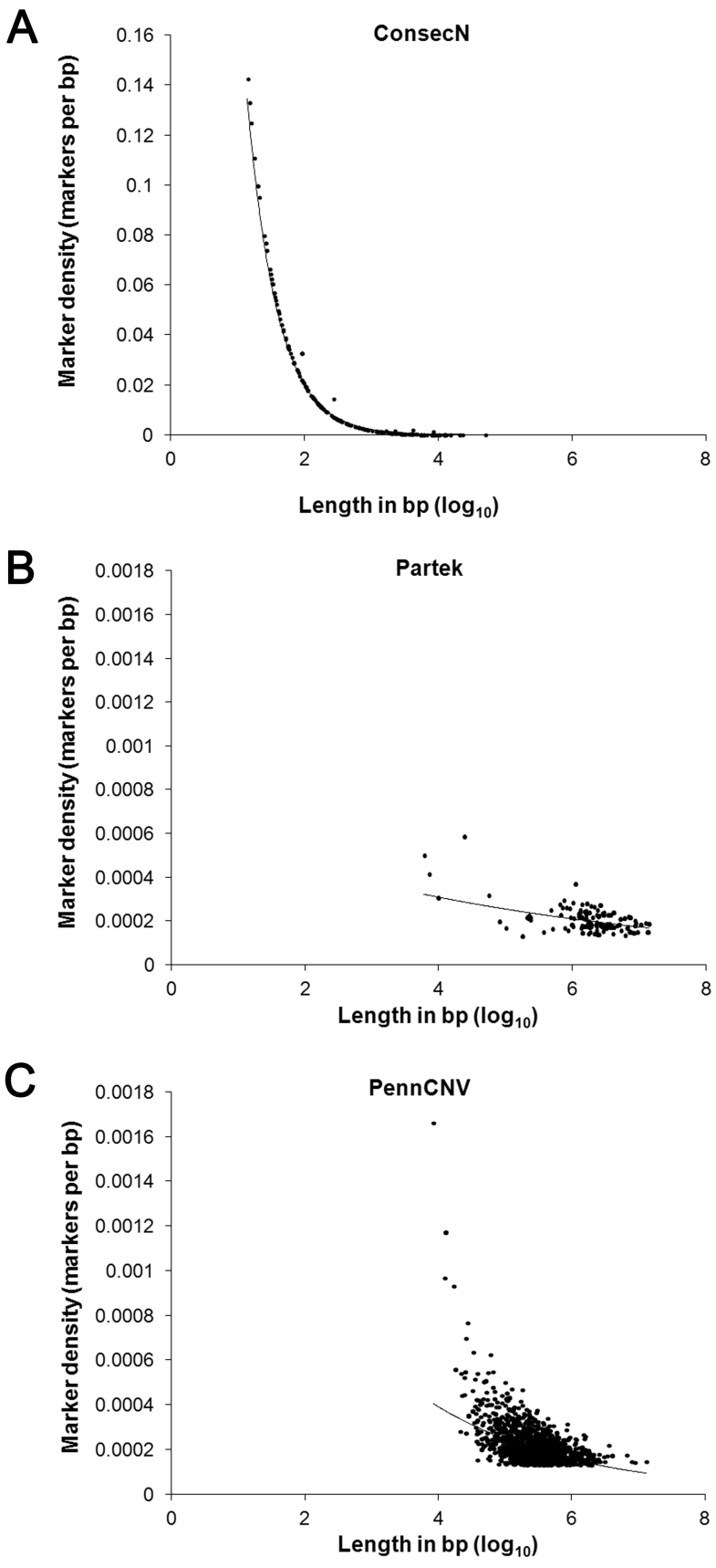
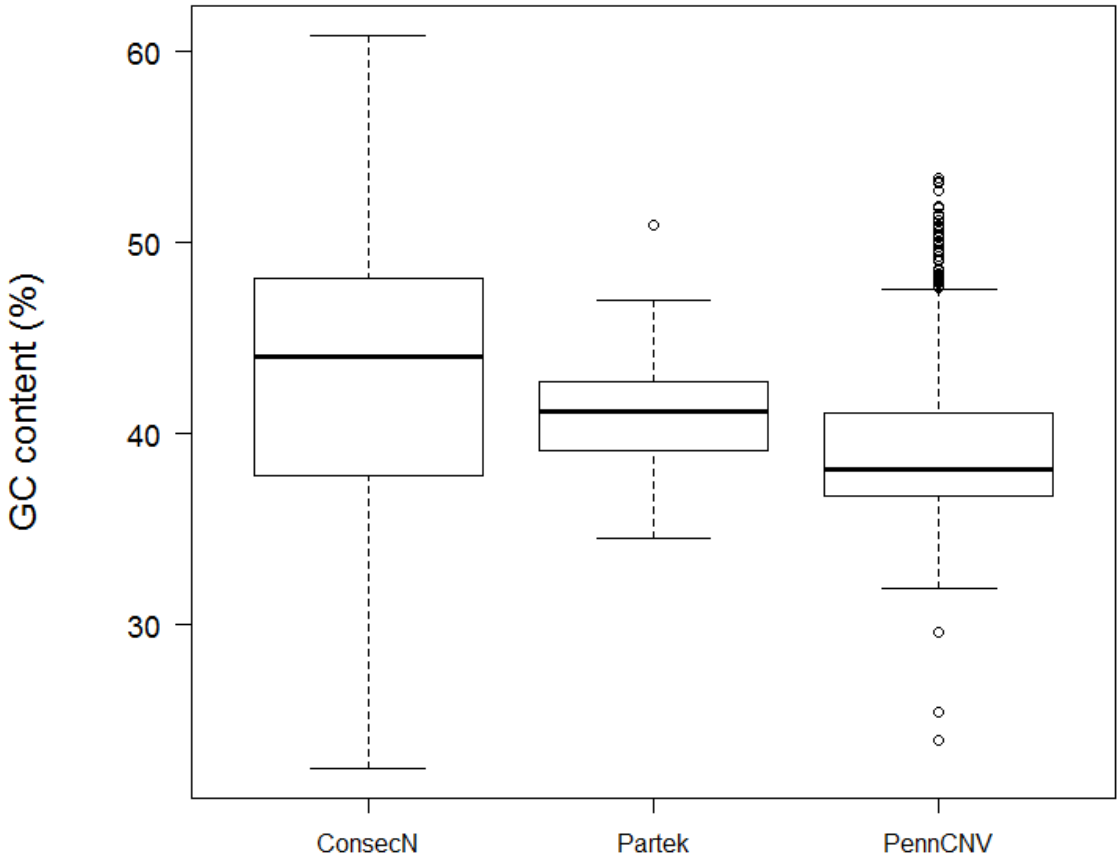

A
ConsecN
Marker density (markers per bp)
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
2
4
6
8
Length in bp (log10)
B
Partek
Marker density (markers per bp)
0.0018
0.0016
0.0014
0.0012
0.001
0.0008
0.0006
0.0004
0.0002
0
0
2
4
6
8
Length in bp (log10)
C
PennCNV
Marker density (markers per bp)
0.0018
0.0016
0.0014
0.0012
0.001
0.0008
0.0006
0.0004
0.0002
0
0
2
4
6
8
Length in bp (log10)

FIGURE 3.6: **Spread of copy number event GC content for events called by ConsecN, Partek, and PennCNV.** ConsecN called copy number events with the highest GC content of the three calling methods ($p<0.001$).


GC content (%)
60
50
40
30
ConsecN
Partek
PennCNV

3.4 Characterization of copy number events as merge-associated events or singletons

3.4.1 Each of the three calling methods detects singletons in the dataset

Of the 239 events called by ConsecN in the fifteen Hill Laboratory samples, HD-CNV found 186 events that did not overlap with any other events in other samples, and were thus classified as singletons (77.82% of events). Of the 224 Partek events, 37 (16.52%) were classified as singletons. Singletons accounted for 515 (29.13%) of the 1768 events called by PennCNV. Partek and PennCNV were similar to each other in that more events were classified as merge-associated events than were classified as singletons. Partek was the only method to call both gains and losses, and so singleton and merge-associated event analysis was further subdivided according to copy number state; however, statistics could not be performed for Partek singleton losses due to small sample size (n=4).

3.4.2 Some copy number events are present in the same genomic location across multiple samples

Of the 239 events called by ConsecN in the fifteen Hill Laboratory samples, HD-CNV found 53 events (22.18%) that were classified as merge-associated events, as they overlapped events occurring in other samples in the dataset by at least 40% of the length of the smallest event. The 53 merge-associated events called by ConsecN formed 17 merges. Although ConsecN called events on all 19 autosomes, merges were found only on chromosomes 1, 3-5, 7-9, and 13-15 (Figure 3.7). Chromosome 15 had the most ConsecN merges, with five events located between base pair positions 16116961-16295795. Of the 224 Partek events, 187 (83.48%) were classified as merge-associated events. HD-CNV called 187 merge-associated events from the Partek events which formed 41 merges (Figure 3.8). Partek merges were called

FIGURE 3.7: **The autosomal distribution of ConsecN copy number events and merges. Outer circle:** Ideogram approximates locations of bands observed on Geimsa-stained chromosomes. **Middle circle:** Each copy number event called by ConsecN is plotted against the corresponding genomic location and shown as a black line. **Inner circle:** Each merge called by HD-CNV for ConsecN copy number events is plotted against the corresponding genomic location and shown as a blue line.

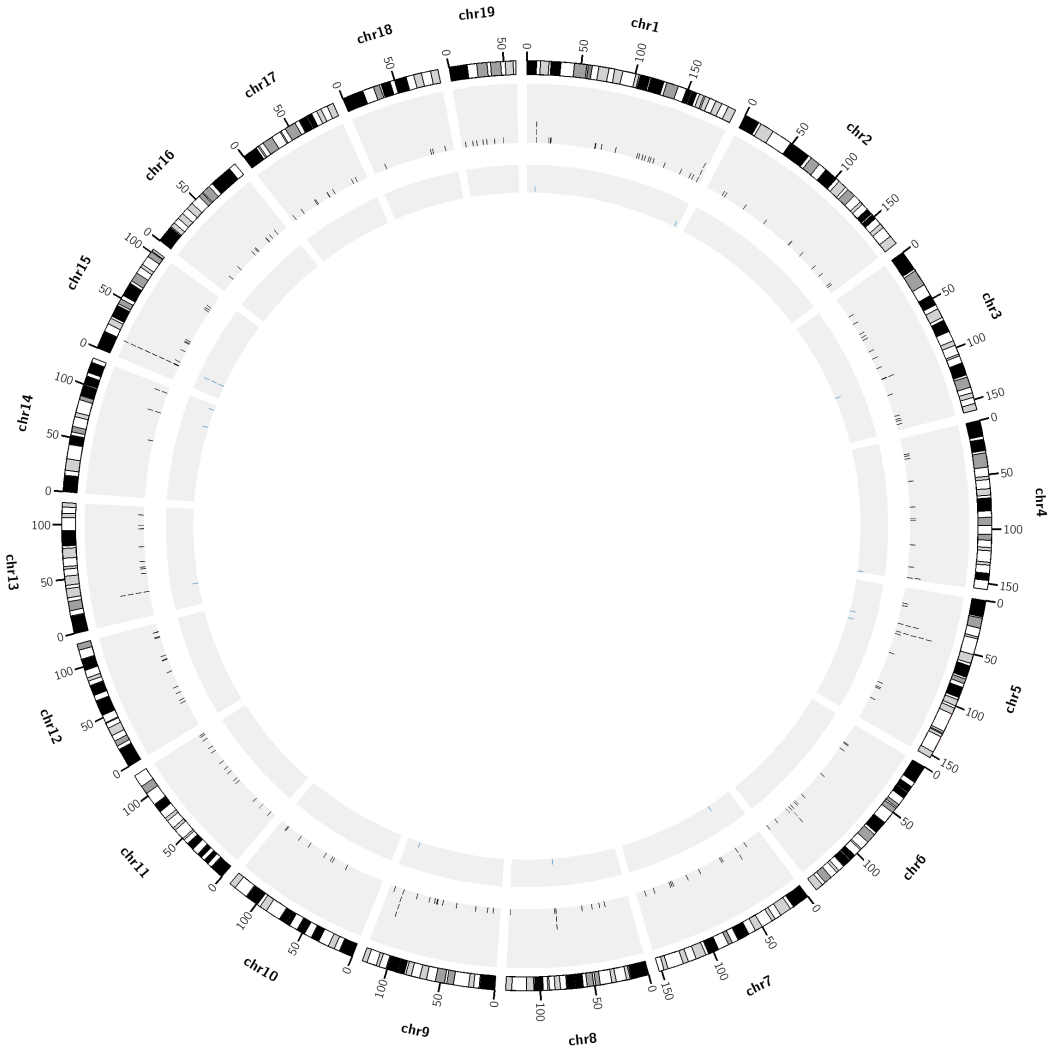
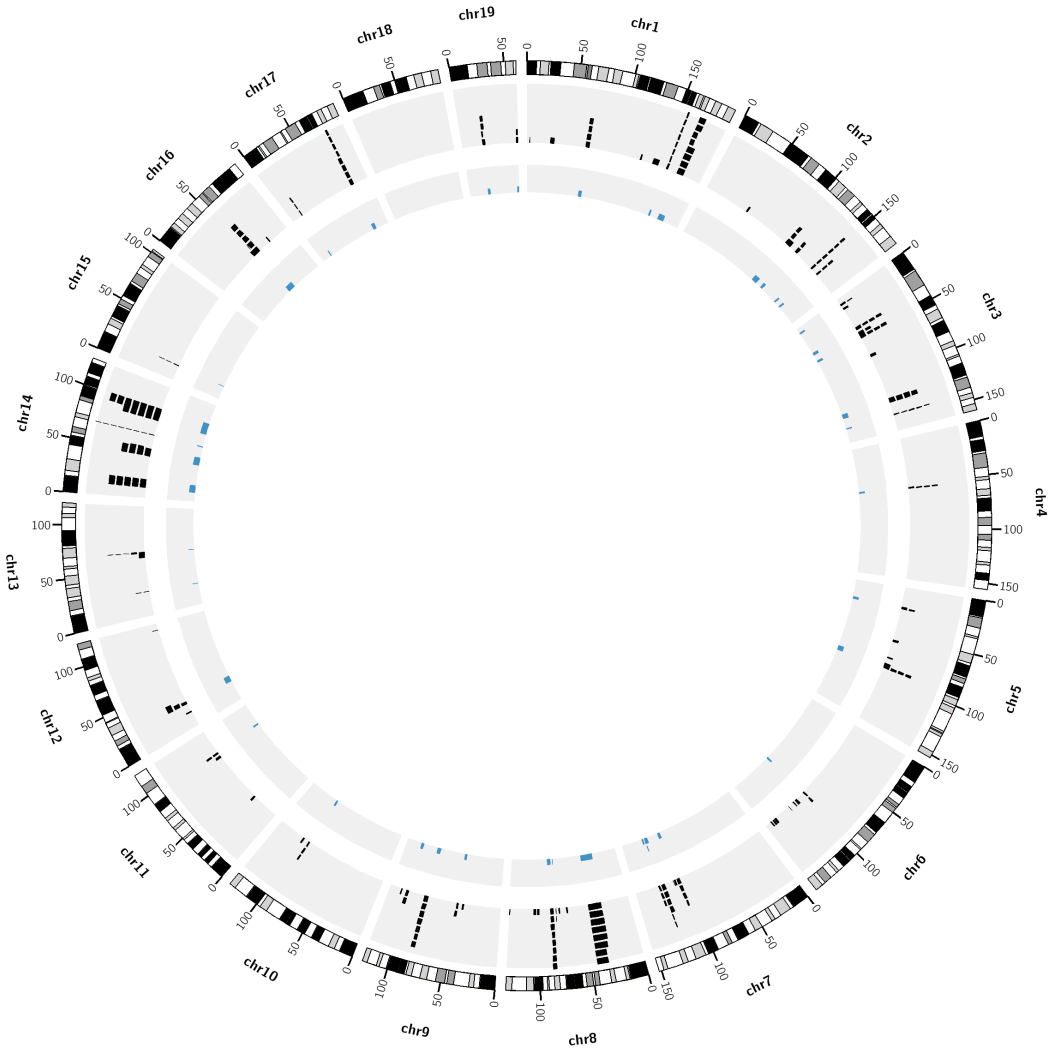

chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19

FIGURE 3.8: **The autosomal distribution of Partek copy number events and merges. Outer circle:** Ideogram approximates locations of bands observed on Geimsa-stained chromosomes. **Middle circle:** Each copy number event called by Partek is plotted against the corresponding genomic location and shown as a black line. **Inner circle:** Each merge called by HD-CNV for Partek copy number events is plotted against the corresponding genomic location and shown as a blue line.


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

for all chromosomes except chromosome 18 (Partek did not call any copy number events on this chromosome). PennCNV called 1768 events, of which 515 (29.13%) were classified as singletons and 1253 (70.87%) classified as merge-associated. From the 1253 merge-associated PennCNV events, 451 merges were found (Figure 3.9). PennCNV merges were found on all 19 autosomes.

3.4.3 The median lengths of singletons and merge-associated events differ between CNV calling methods

The lengths of ConsecN singletons (median=337 bp) were not different from the lengths of merge-associated events (median=1,622 bp; H=0.392, df=1, p=0.5312; Figure 3.10). Before accounting for copy number state, the lengths of Partek singletons (median=1,538 kb) were significantly shorter than the lengths of merge-associated events (median=2,592 kb; H=8.8835, df=1, p=0.003). Partek events were then split according to copy number state. Singleton losses (median=32 kb) were too few in number (n=4) to statistically compare lengths with singleton gains (median=1,744 kb) or merge-associated losses (median=9.8 kb). The lengths of singleton gains (median=1,862 kb) were significantly shorter than the lengths of merge-associated gains (median=3,625 kb; H= 14.4694, df=1, p<0.001). PennCNV singletons (median=250 kb) were similar in length to PennCNV merge-associated events (median=249 kb; H=2783.713, 1 d.f., p=0.2903).

The proportion of the genome affected by copy number events is a function of both the number and length of events. Based on the Golden Path length of the mouse genome, ConsecN singletons affected an average of 35,084 bp ± 15,971 in each sample, while merge-associated events affected an average of 5,731 bp ± 1,762 per sample. Partek singletons affected on average 3,806,384 bp ± 1,815,005 in each sample, while merge-associated events affected an average of 44,105,371 bp ± 4,974,297 per sample. PennCNV singletons affected an average of 13,292,831 bp ± 5311420 in each sample, while merge-associated events affected an average

FIGURE 3.9: **The autosomal distribution of PennCNV copy number events and merges. Outer circle:** Ideogram approximates locations of bands observed on Geimsa-stained chromosomes. **Middle circle:** Each copy number event called by PennCNV is plotted against the corresponding genomic location and shown as a black line. **Inner circle:** Each merge called by HD-CNV for PennCNV copy number events is plotted against the corresponding genomic location and shown as a blue line.

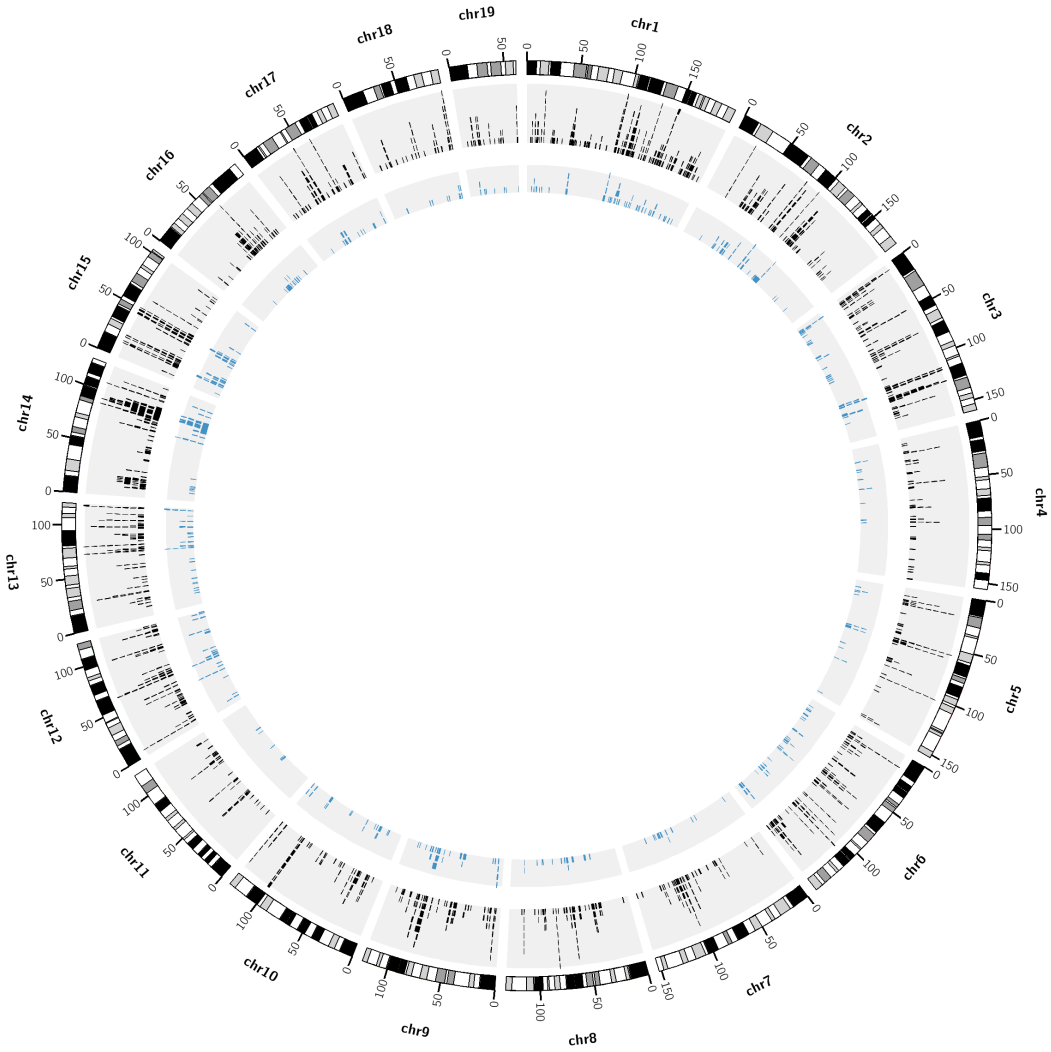
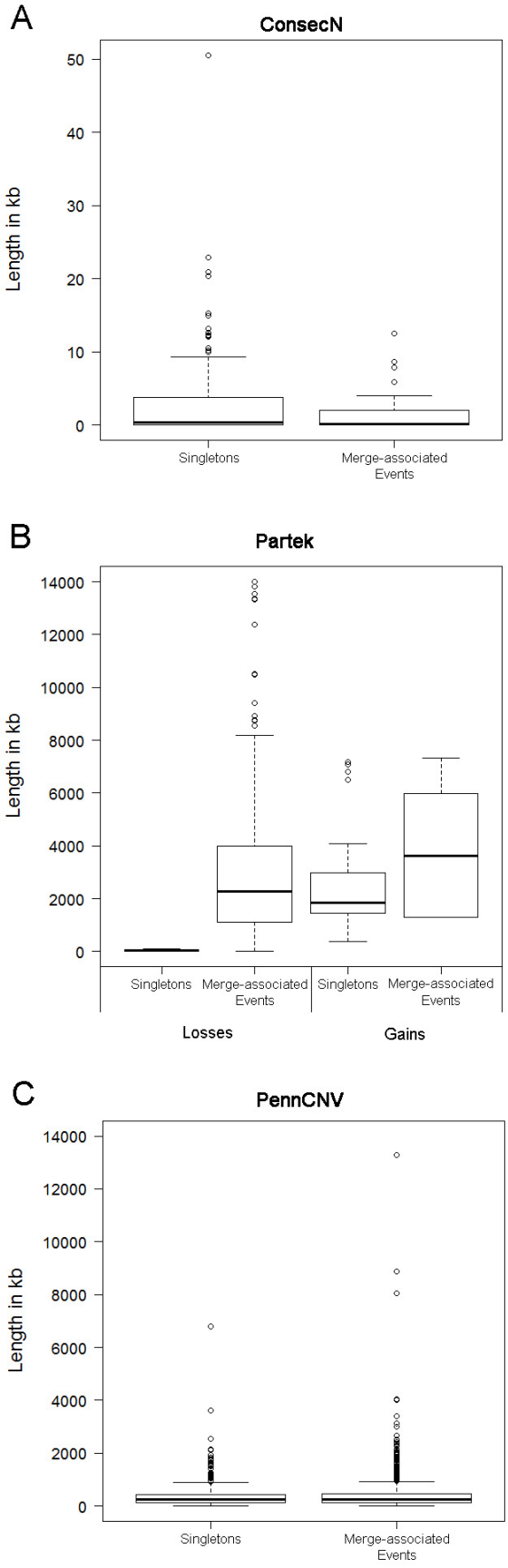

chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19

FIGURE 3.10: **Spread of the length values calculated for singletons and merges called by HD-CNV for copy number events detected with each of ConsecN, Partek, and PennCNV.** **(A)** The lengths of ConsecN singletons and merge-associated events were similar. **(B)** The lengths of Partek singleton gains were shorter than merge-associated gains ($p<0.001$). **(C)** The lengths of PennCNV singletons and merge-associated events did not differ.


A
ConsecN
Length in kb
0
10
20
30
40
50
Singletons
Merge-associated Events
B
Partek
Length in kb
0
2000
4000
6000
8000
10000
12000
14000
Singletons
Merge-associated Events
Singletons
Merge-associated Events
Losses
Gains
C
PennCNV
Length in kb
0
2000
4000
6000
8000
10000
12000
14000
Singletons
Merge-associated Events

of 9,822,690 bp ± 1,396,849 per sample.

3.4.4 Marker density differs between singletons and merge-associated events

ConsecN merge-associated events had a higher marker density (median 106.38 markers per 10 kb) than ConsecN singletons (median 59.35 markers per 10 kb; H=332.7322, 1 d.f., p<0.001; Figure 3.11). In events called by Partek, merge-associated events also had a higher marker density (median 1.88 markers per 10 kb) than singletons (median 1.77 markers per 10 kb; H=244.1718, 1 d.f., p<0.001). When Partek events were split based on gains versus losses, singleton gains had a lower median marker density (1.742 markers per 10 kb) than merge-associated gains (1.837 markers per 10 kb, H=201.4678, 1 d.f., p<0.001). Partek merge-associated losses had a higher marker density (median 2.160 markers per 10 kb) than merge-associated gains (median 1.837 markers per 10 kb; H=221.6754, 1 d.f., p<0.001). PennCNV events differed from ConsecN and Partek events in that singletons had a higher marker density (median 1.133 markers per 10 kb) than merge-associated events (median 1.83 markers per 10 kb; H=1306.132, 1 d.f., p<0.001).

3.4.5 GC content differed between singletons and merge-associated events for Partek and PennCNV events

ConsecN singletons had a median GC content of 44.02%, while merge-associated events had a median GC content of 42.45%. The GC content of ConsecN events was similar between singletons and merge-associated events (H=0.4019, 1 d.f., p=0.5261; Figure 3.12). Of the Partek events, merge-associated gains had a higher GC content (median=41.15%) than singleton gains (median=38.70%; H=314.4831, 1 d.f., p<0.001). Merge-associated losses had a higher GC content (median=46.76%) than merge-associated gains (median=41.15%;

FIGURE 3.11: **Spread of copy number event marker density values calculated for singletons and merges called by HD-CNV for copy number events detected with each of ConsecN, Partek, and PennCNV. (A)** ConsecN singletons had lower marker densities than merge-associated events ($p<0.001$). **(B)** Partek singleton gains had lower marker densities than merge-associated gains ($p<0.001$), which in turn had lower marker densities than merge-associated losses. **(C)** PennCNV singletons had higher marker densities than merge-associated events ($p<0.001$).

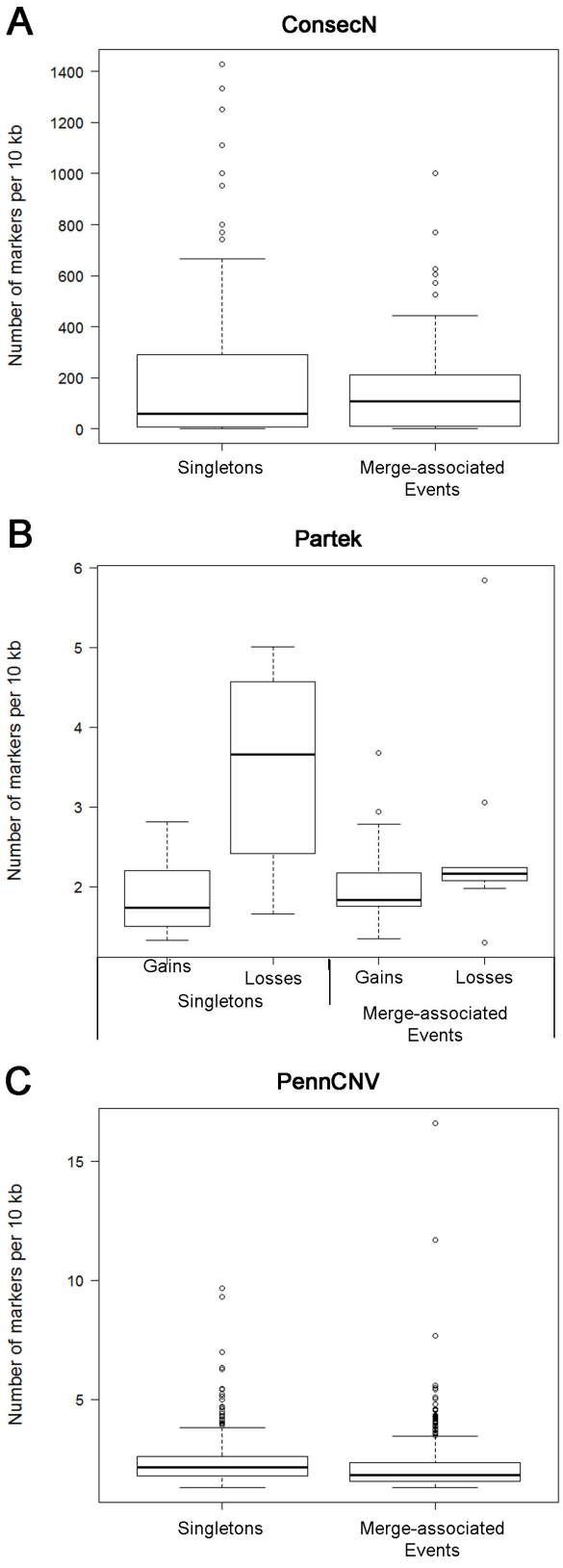
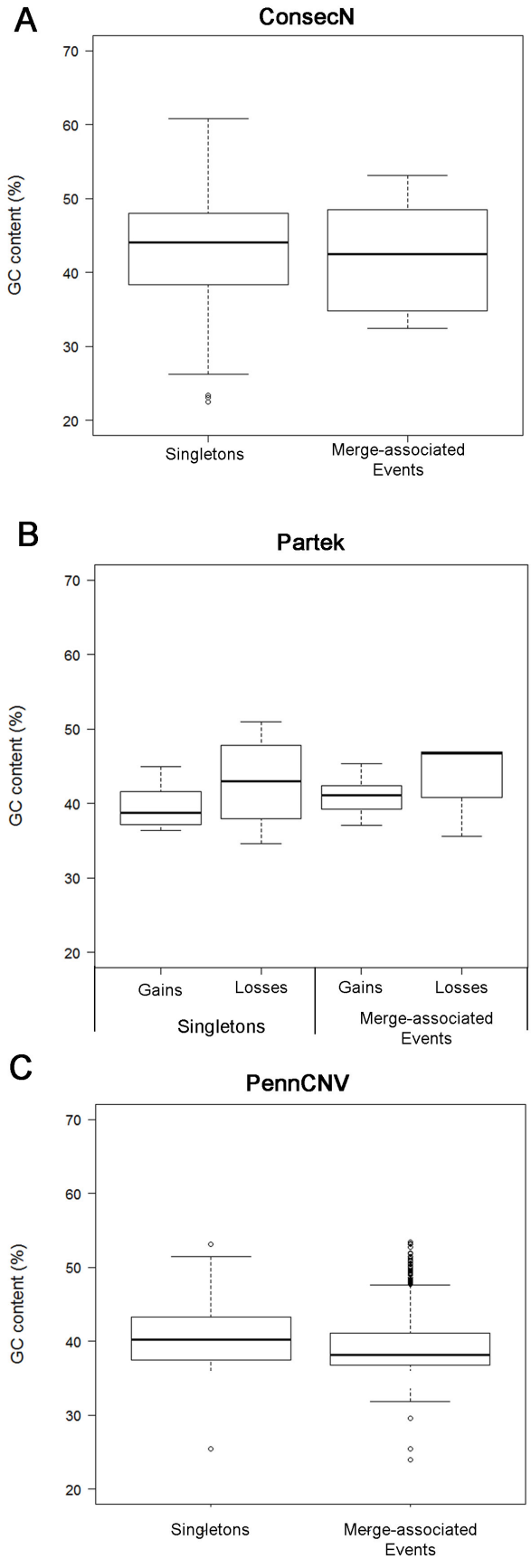

A
ConsecN
Number of markers per 10 kb
0
200
400
600
800
1000
1200
1400
Singletons
Merge-associated Events
B
Partek
Number of markers per 10 kb
2
3
4
5
6
Gains
Losses
Gains
Losses
Singletons
Merge-associated Events
C
PennCNV
Number of markers per 10 kb
5
10
15
Singletons
Merge-associated Events

FIGURE 3.12: **Spread of copy number event GC content calculated for singletons and merges called by HD-CNV for copy number events detected with each of ConsecN, Partek, and PennCNV.** **(A)** The GC content of ConsecN events did not differ between singletons and merges. **(B)** Partek singleton gains had lower GC content than merge-associated gains, which in turn had lower GC content than merge-associated losses ($p<0.001$). **(C)** PennCNV singletons had higher GC content than merge-associated events ($p<0.001$).


A
ConsecN
GC content (%)
70
60
50
40
30
20
Singletons
Merge-associated Events
B
Partek
GC content (%)
70
60
50
40
30
20
Gains
Losses
Gains
Losses
Singletons
Merge-associated Events
C
PennCNV
GC content (%)
70
60
50
40
30
20
Singletons
Merge-associated Events

H=325.7643, 1 d.f., p<0.001). PennCNV singletons had a higher GC content (40.45%) than PennCNV merge-associated events (38.10%; H=125.5902, 1 d.f., p<0.001).

3.5 Each CNV calling method detects some events that overlap with previously documented CNVs or CNVRs

Three ConsecN merges overlapped with previously detected CNVs, of which only one was previously found in a C57BL/6J mouse (Table 3.4).[219] Thirty of the 239 total events (12.55%) detected with ConsecN were found to overlap previously found CNVs. Of the 17 merges detected with ConsecN, three (17.60%) were found to overlap with previously reported events. Two hundred fifteen of the 224 total events (95.98%) detected with Partek were found to overlap previously found CNVs. Of the 41 merges detected with Partek, 38 (92.68%) were found to overlap with previously reported events. Of the 239 total events detected with PennCNV, 1123 (63.52%) were found to overlap previously found CNVs. Of the 451 merges detected with PennCNV, 346 (76.72%) were found to overlap with previously reported events.

3.6 Genetic distance between samples can be estimated using CNVs and CNCs

The Pedigree difference matrix, generated using coefficients of relatedness between samples, was similar to the Partek matrix (Z=19.48397, p=0.025); however, the Pedigree matrix was unrelated to both the ConsecN and PennCNV matrices. The ConsecN and PennCNV matrices were similar to one another (Z=4.559409, p<0.001). The Partek matrix was unrelated to both the ConsecN (Z=3.82386, p=0.750) and PennCNV (Z=6.172546, p=0.789) matrices.

Four unrooted additive phylogenetic trees were produced: the “Pedigree” tree constructed from the coefficient of relatedness *r* between samples, in addition to the three trees constructed

Table 3.4: Number of copy number events and merges called by ConsecN, Partek, and PennCNV that overlap with entries in the database of previously documented murine CNVs and CNVRs.

	Number of copy number events that overlap with database entries			**Number of copy number events that overlap with C57BL/6J entries**[a]	**Number of merges that overlap with database entries**	**Number of database entries that overlap with copy number events**
	Singletons	Merge-associated	Total			
ConsecN	27 (14.52%)	3 (5.66%)	30 (12.55%)	2 (0.84%)	1 (5.88%)	73 (0.43%)
Partek	36 (97.30%)	179 (95.72%)	215 (95.98%)	85 (37.95%)	38 (92.68%)	10 958 (63.98%)
PennCNV	334 (64.85%)	789 (62.97%)	1123 (63.52%)	110 (6.22%)	346 (76.72%)	6457 (37.70%)

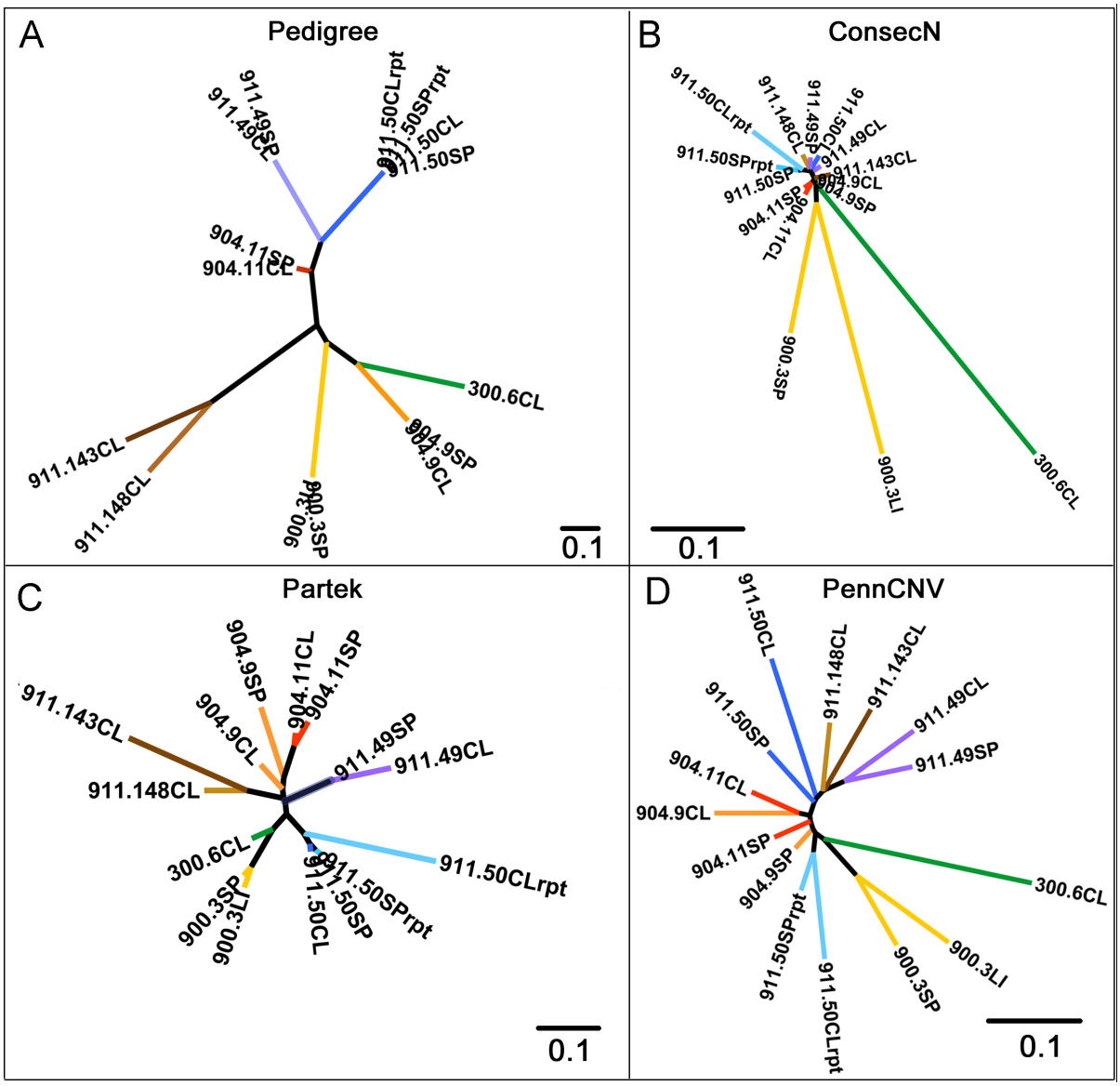
[a] C57BL/6J entries were sourced from Henrichsen *et al.* [63]

from genetic distance measures based on sharing proportion of CNV loci (ConsecN, Partek, and PennCNV; Figure 3.13). In the Pedigree tree, tissues from the same mouse were not permitted to have any genetic distance between them as imputed by the theoretical coefficient of relatedness (r=1). The distance between mice characterized by the highest number of nodes is the separation between mouse 300.6 and both 911.49 and 911.50, with five nodes. However, this relationship did not have the highest branch length sum (0.9248); instead, the separation between mouse 300.6 and mouse 911.143 tied with 300.6 and 911.148 for the greatest branch length sum (1.0446 for both relationships). All pairs of brothers were placed close to each other except for mice 904.9 and 904.11. Although the tree placed mouse 904.9 between 300.6 and 900.3, 904.9's brother 904.11 was placed at a further distance from both 300.6 and 900.3 than 904.9.

In the ConsecN tree, tissues from the same mouse clustered together for mice 900.3, 904.9, and 904.11. Both samples of mouse 911.50's spleen clustered together, which in turn were close to the replicate of 911.50's cerebellum. The first cerebellum sample from 911.50 was separated from the replicate by two nodes and a branch length sum of 0.0953. The cerebella of brothers 911.143 and 911.148 were separated by four nodes and a branch length sum of 0.466. The ConsecN tree shows a deep split between the spleen and liver of mouse 900.3 (separation of one node, branch length sum of 0.4236). The separation between mouse 300.6 and the two tissues of mouse 900.3 was even greater with a separation of two nodes and branch length sums of 0.5374 (300.6CL-900.3SP) and 0.6738 (300.6CL-900.3SP).

Overall, the Partek tree demonstrates more bifurcation than the other trees, evident by the higher number of nodes formed. All tissue pairs were separated by one node only, except for the tissues of 904.9 and the cerebellar samples from 911.50. The two 911.50 spleen samples were separated from each other by fewer nodes and shorter branch lengths than they were separated from the corresponding cerebellum samples. Though the original cerebellum sample clustered closely with the two spleen samples, the cerebellum replicate 911.50CL was more distant from this cluster. For all mice, pairs of brothers were always plotted near each other.

FIGURE 3.13: **Unrooted phylogenetic trees generated based on shared CNV loci between individual samples.** Branches of the same colour indicate tissues from the same mouse. **(A)** Pedigree **(B)** ConsecN **(C)** Partek **(D)** PennCNV.


A
Pedigree
911.49SP
911.49CL
911.50CLrpt
911.50SPrpt
911.50CL
911.50SP
904.11SP
904.11CL
300.6CL
911.143CL
911.148CL
904.9SP
904.9CL
900.3SP
900.3LI
0.1
B
ConsecN
911.50CLrpt
911.148CL
911.49SP
911.50CL
911.49CL
911.50SPrpt
911.50SP
911.143CL
904.9CL
904.11SP
904.9SP
904.11CL
900.3SP
900.3LI
300.6CL
0.1
C
Partek
904.9SP
904.11CL
904.11SP
911.143CL
904.9CL
911.49SP
911.49CL
911.148CL
300.6CL
911.50CLrpt
900.3SP
900.3LI
911.50CL
911.50SP
911.50SPrpt
0.1
D
PennCNV
911.50CL
911.148CL
911.143CL
911.50SP
911.49CL
911.49SP
904.11CL
904.9CL
904.11SP
904.9SP
300.6CL
911.50SPrpt
911.50CLrpt
900.3SP
900.3LI
0.1

Mice 300.6 and 900.3 clustered together, but with less distance between the individual samples than in the Pedigree or ConsecN trees.

Like the Pedigree, ConsecN, and Partek trees, the PennCNV tree also places the tissue samples from mice 300.6 and 900.3 together. The PennCNV tree shows both 911.50 tissue replicates closer to 300.6 and 900.3, followed by the tissues from the brothers 904.9 and 904.11. For this pair of brothers, rather than the two tissues from within each mouse clustering together, the tissues were clustered by tissue type rather than which mouse they were sampled from. The spleen samples from 904.9 and 904.11 were separated by two nodes and a branch length sum of 0.0762, while the cerebellum samples were separated from each other by one node and a branch length sum of 0.1408. The cerebellum samples from mice 904.9 and 904.11 were separated from their corresponding spleen samples by 3 and 4 nodes, respectively. This pair of brothers separated the 911.50 replicate samples previously discussed from the original 911.50 samples, the spleen and cerebellum of which were separated by two nodes and a branch length sum of 0.2268. The cerebella of the brothers 911.143 and 911.148 were clustered together, separated by one node and a branch length sum of 0.1696. Finally, both tissues from mouse 911.49 were clustered together, although they were separated from the 911.50 samples.

3.6.1 Partek and PennCNV trees, but not the ConsecN tree, are topologically similar to the Pedigree tree

The four trees were compared pairwise using two methods of topological comparisons. The first comparison was made using the number of nodes separating each pair of samples in the trees and performing a Spearman's rank correlation between each pair of trees (resulting in six topological comparisons). This comparison found that all four trees were similar to each other ($p<0.001$ for all six tree comparisons).

There are fewer significant relationships between phylogenetic trees when comparing trees based on the sums of branch lengths that separate each pair of samples. The Partek tree was

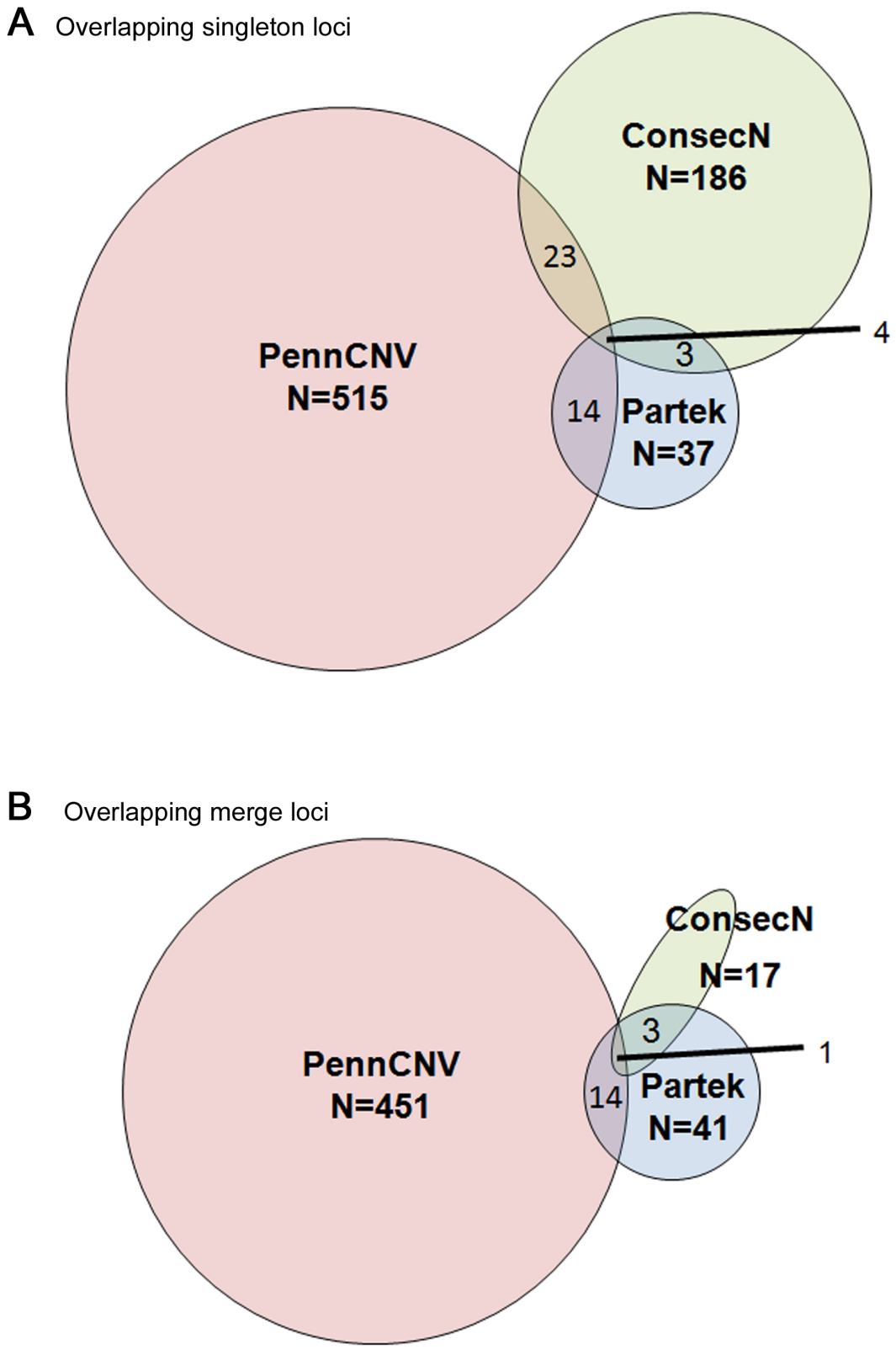
weakly correlated with the Pedigree tree ($r(103)=0.3971562$, $p<0.001$). The Pedigree tree was also weakly correlated with the PennCNV tree ($r(103)=0.3512727$, $p<0.001$). The PennCNV tree in turn was correlated with the ConsecN tree ($r(103)=0.5947334$, $p<0.001$). All trees except for the Pedigree tree were correlated with only one other tree using the branch length sum method of comparison. All other correlations were not significant.

3.7 There is a low degree of concordance in copy number events between the three CNV calling methods

Few of the singletons discovered within each CNV calling method were found to overlap with singletons found in the other two CNV detection methods. In total, 739 singletons were found using all three CNV calling methods, but only four overlapping singleton loci encompassed events found by all three CNV detection methods (Figure 3.14). ConsecN and Partek had the fewest number of overlapping singletons (three singleton loci), while ConsecN and PennCNV had the greatest number of overlapping singletons (23 singleton loci). Partek and PennCNV singletons overlapped at 14 loci. Singletons that overlapped between CNV calling methods were frequently called in different tissue samples. Two of the three ConsecN-Partek singleton loci contained singletons found in the same sample (one locus comprised singletons found in 911.143CL). One of the 23 singleton loci that overlapped between ConsecN and Partek comprised singletons found in the liver of mouse 900.3. Similarly, one of the 14 Partek-PennCNV overlapping singleton loci comprised events called in 900.3SP. The remaining overlapping singleton loci comprised singletons from different samples.

One overlapping merge locus was found that comprised merges from all three CNV calling methods. Partek and PennCNV shared 14 merge loci, while ConsecN and Partek shared three merge loci. ConsecN and PennCNV had no overlapping merges. There was a low degree of consistency in the samples represented in merges that overlapped between calling methods. Only one overlapping merge locus contained identical samples across calling methods. This

FIGURE 3.14: **Number of singleton loci and number of merge loci that overlap among the three CNV calling methods.** **(A)** The overlap of singleton loci among ConsecN, Partek, and PennCNV, where *N* indicates the total number of singletons detected by each method. **(B)** The overlap of merge loci among ConsecN, Partek, and PennCNV, where *N* indicates the total number of merge loci detected by each method.


A
Overlapping singleton loci
ConsecN
N=186
23
4
3
PennCNV
N=515
14
Partek
N=37
B
Overlapping merge loci
ConsecN
N=17
3
1
PennCNV
N=451
14
Partek
N=41

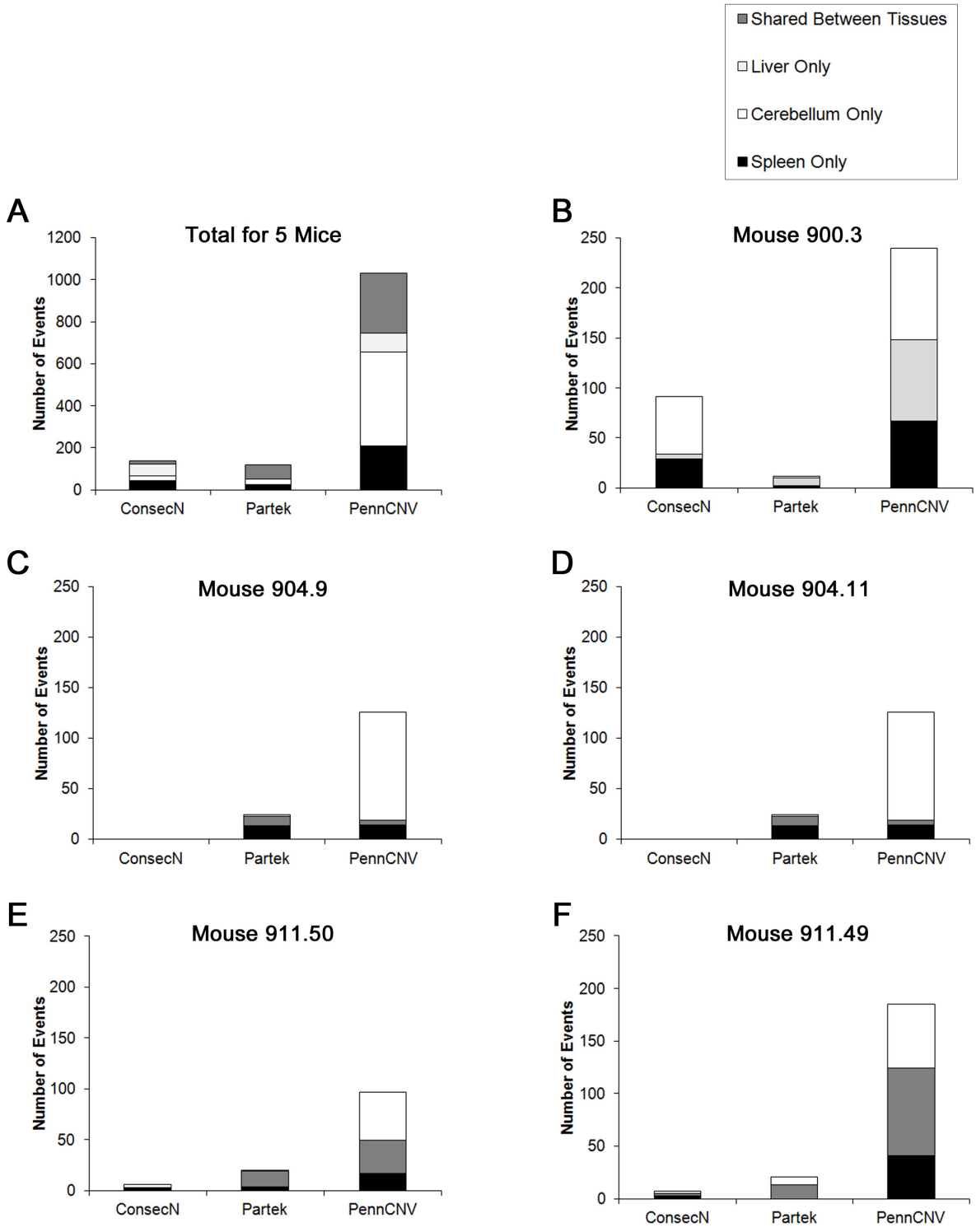
overlapping merge locus comprised events found in both spleen and liver of mouse 900.3 called by both ConsecN and Partek. The one overlapping merge locus that comprised merges from all three CNV calling methods contained events in samples 911.143CL and 911.148CL called by each of ConsecN, Partek, and PennCNV, although the PennCNV merge also contained events found in mice 911.49 and 911.50. One Partek-PennCNV overlapping merge locus contained events called by both methods in sample 911.50CLrpt, and another Partek-PennCNV overlapping merge locus contained events called by both methods in samples 911.143CL and 911.49CL. The remaining overlapping merge loci comprised events from different samples, with each sample in the overlapping merge loci represented in only one of the constituent merges.

3.8 Tissue pairs from the same mouse demonstrate both similarities and differences in copy number events

ConsecN called the fewest number of shared copy number events in the six tissue pairs (from five mice, including the 911.50 replicates), with 14 shared events in total (Figure 3.15). Partek called 64 shared events, while PennCNV called the most shared events with 284. Events called in the liver only were limited to mouse 900.3. Events called in the cerebellum only within tissue pairs were most frequent in the PennCNV-called events (447), followed by Partek (27), then by ConsecN (23). Events called in the spleen only within tissue pairs were again most frequent in the PennCNV-called events (207); however, ConsecN called more spleen-only events within tissue pairs (43) than Partek (24).

Mouse 900.3 was the only mouse with the liver substituted for the cerebellum. This mouse had five copy number events shared between the spleen and liver called by ConsecN, 8 shared events called by Partek, and 81 shared events called by PennCNV. It was the only mouse in which ConsecN called more events than Partek. ConsecN called fewer spleen-only events (29) than liver-only events (57). Partek called two tissue-specific events in both spleen and liver.

FIGURE 3.15: **Number of copy number events that occur in one or both tissues as detected by three CNV calling methods.** The number of copy number events that occur in spleen only, cerebellum only, liver only, or shared between the two tissues assayed for each individual as called by three CNV calling methods. **(A)** The total number of copy number events for five mice that each had two tissues sampled. The remaining graphs show the number of shared and tissue-specific events for individual mice: **(B)** Mouse 900.3 (spleen and liver); **(C)** Mouse 904.9 (cerebellum and spleen); **(D)** Mouse 904.11 (cerebellum and spleen); **(E)** Mouse 911.50 (cerebellum and spleen for both the original and replicate); and **(F)** Mouse 911.49 (cerebellum and spleen).


Shared Between Tissues
Liver Only
Cerebellum Only
Spleen Only
A
Total for 5 Mice
Number of Events
1200
1000
800
600
400
200
0
ConsecN
Partek
PennCNV
B
Mouse 900.3
Number of Events
250
200
150
100
50
0
ConsecN
Partek
PennCNV
C
Mouse 904.9
Number of Events
250
200
150
100
50
0
ConsecN
Partek
PennCNV
D
Mouse 904.11
Number of Events
250
200
150
100
50
0
ConsecN
Partek
PennCNV
E
Mouse 911.50
Number of Events
250
200
150
100
50
0
ConsecN
Partek
PennCNV
F
Mouse 911.49
Number of Events
250
200
150
100
50
0
ConsecN
Partek
PennCNV

Similar to ConsecN, PennCNV called fewer spleen-only events (67) than liver-only events (92).

In mouse 904.9, ConsecN did not call any copy number events. Of the events Partek called in mouse 904.9, events at 10 loci were found in both the spleen and cerebellum. Partek found more spleen-only events (13) than cerebellum-only events (1) in this mouse. PennCNV found more cerebellum-only events (107) than spleen-only events (14), while only five loci were found to have events in both spleen and cerebellum.

In mouse 904.11, the brother of 904.9, ConsecN called one shared event between the spleen and the cerebellum. All other events occurred in only one of the two tissues, with two events found only in the spleen and three events found only in the cerebellum. Partek called more shared events than ConsecN, calling 15 shared events between the tissues. Partek did not call any events that were in the spleen only, but did detect eight events found only in the cerebellum. PennCNV called the most events, with 83 events shared between the tissues, 41 found only in the spleen, and 61 found only in the cerebellum.

Mouse 911.49 had the highest ratio of shared events to tissue-specific events of the five mice analyzed. ConsecN called two shared events between the spleen and cerebellum, Partek called 13 shared events, and PennCNV called 83 shared events. ConsecN called three spleen-only events and two cerebellum-only events. All the events called by Partek in the spleen of mouse 911.49 were shared with the cerebellum; however, the cerebellum was found to have eight events not shared with the spleen. Like Partek, PennCNV also found more cerebellum-only events (61) than spleen-only events (41).

Both replicates of tissues sampled from mouse 911.50 had similar numbers of events shared between tissues as called by each of the three CNV calling methods. ConsecN called two shared events in 911.50 and four shared events in 911.50rpt. Partek called 10 shared events in 911.50 and eight in 911.50rpt. PennCNV called the most shared events in both replicates, with 43 shared events in 911.50 and 40 shared events in 911.50rpt. ConsecN found

fewer spleen-only events than cerebellum-only events in both replicates. In 911.50, ConsecN called three spleen-only events and four cerebellum-only events. The difference between tissues was greater in the 911.50rpt samples, with ConsecN calling six spleen-only events and 14 cerebellum-only events. Although Partek called the same number of tissue-specific events in both tissues in 911.50 (one tissue-specific event in the spleen and one in the cerebellum), Partek called fewer spleen-only events (4) than cerebellum-only events (16) in the replicate 911.50rpt. PennCNV called fewer spleen-only events than cerebellum-only in both replicates. In the first 911.50 tissue pair there were 49 spleen-only events called by PennCNV, compared to 135 cerebellum-only. In 911.50rpt, PennCNV called 19 spleen-only events and 96 cerebellum-only events.

In the samples in which ConsecN called events, ConsecN found more events that were cerebellum-only or liver-only than spleen-only in mice 904.11 and 911.50 (both replicates). ConsecN called more spleen-only events in mouse 911.49, with three spleen-only events compared to two cerebellum-only events. Partek called the same number of copy number events unique to one tissue for both tissues of mice 911.50 (original) and mouse 900.3. Partek called more spleen-only than cerebellum-only events in the brothers 904.9 and 904.11, and more cerebellum-only events than spleen-only events in the brothers 911.49 and 911.50 (replicate). PennCNV always found fewer events that occurred only in the spleen than in the other paired tissue (cerebellum or liver).

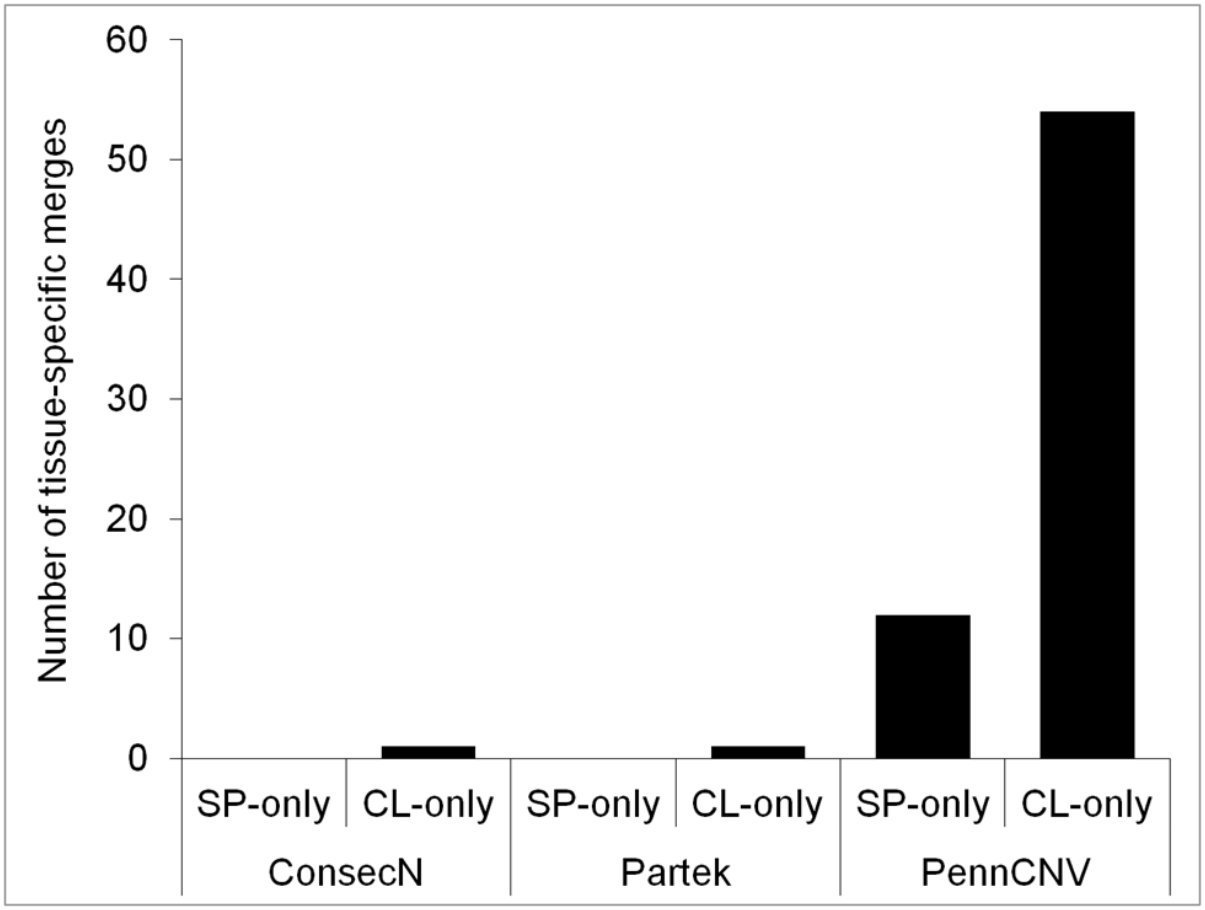
3.8.1 Merges identify genomic regions across different samples that contain overlapping copy number events

The only merge to contain events from all fifteen samples was found with copy number events called by Partek (chr.14, MergeID002). All events in this merge were called as deletions. This merge was not found to overlap with events found by either ConsecN or PennCNV; however, it did overlap several entries in the database, including entries detailing events previ-

ously found in C57BL/6J mice. Further inspection of this region using UCSC Genome Browser revealed the sequence to align with regions on human chromosome 8. This merge encompassed several whole genes in the mouse genome, including ectonucleoside triphosphate diphosphohydrolase 4 (*Entpd4*; human ortholog *ENTPD4* on chr.8), solute carrier family 25, member 37 (*Slc25a37*; human ortholog *SLC25A37* on chr.8), and envelope glycoprotein syncytin-B (*D930020E02Rik*; no human ortholog). The merge also overlapped part of lysyl oxidase-like 2 (*Loxl2*).

Some merges contained events that were detected in only one tissue type. A merge was defined as tissue-specific if all events composing the merge were detected in one tissue type (for example, spleen) and not detected in the paired tissue within the same mouse. Looking at only mice from which two tissues were sampled, all three methods detected at least one tissue-specific merge (Figure 3.16). ConsecN detected no spleen-specific merges and one cerebellum-specific merge. Partek also detected no spleen-specific merges and one cerebellum-specific merge. PennCNV detected 12 spleen-specific merges and 54 cerebellum-specific merges.

FIGURE 3.16: **Number of tissue-specific merges detected in each of the spleen (SP) and the cerebellum (CL) by three copy number calling methods.** More cerebellum-specific merges than spleen-specific merges were found by each of ConsecN, Partek, and PennCNV. Tissue pairing analysis permitted the inclusion of six spleen samples and five cerebellum samples.


Number of tissue-specific merges
60
50
40
30
20
10
0
SP-only
CL-only
SP-only
CL-only
SP-only
CL-only
ConsecN
Partek
PennCNV

Chapter 4

Discussion

4.1 General Discussion

This study reports the first genome-wide analysis of somatic copy number mosaicism using the novel Mouse Diversity Genotyping Array. Somatic copy number mosaicism was identified as differences in genomic copy number state (identified as copy number changes, or CNCs) between two somatic tissues within individual mice. All mice in this study for which two tissues were sampled exhibited CNCs as detected by at least two methods of calling copy number. Looking across mice, the spleen tended to have fewer CNCs than the cerebellum. Tissue-specific merges were identified in both spleen and cerebellum, although cerebellum-specific merges were detected across CNV calling methods, and called most frequently by PennCNV. The sample population, which includes mice from different inbred genetic backgrounds as well as their descendants with mixed inbred strain genetic backgrounds, provides a group in which individuals have different degrees of genetic relatedness that is known prior to CNV analysis. With the added benefit of a documented pedigree, this permits the analysis of variants across multiple levels of genetic relatedness, from within-individual tissue samples to

familial-level relationships. Furthermore, the application of novel software to identify putative CNVRs across tissues samples identifies genomic regions that may be susceptible to recurrent formation events, particularly in a tissue-specific manner. Applying quality control measures including background correction of array data and minimum marker density in copy number events is critical to minimize false-positive copy number calls. The construction of a locally accessible database of previously documented CNVs and CNVRs provides the opportunity to assess the genomic coordinates of putative copy number events for any overlap with events identified in published studies that represent a variety of tissues, strains, platforms, and algorithms. It is within the critical experimental framework I have presented in this thesis that we can identify putative copy number changes contributing to somatic copy number mosaicism in the mouse.

4.1.1 Tissues differ in number and location of copy number events

Under the alternate hypothesis that replication in rapidly dividing cells is responsible for the formation of *de novo* CNCs, it would be expected that of the three tissues sampled in this study, the liver and spleen would have more CNCs. The one liver sample in this study was shown to have a relatively high burden of copy number events compared with the other tissues sampled. Although the murine liver is prone to rapid accumulation of genomic rearrangements with age, that pattern was seen previously in mice over 27 months of age, while the age of the mouse in this study was 11.3 months, indicating a younger time point at which these rearrangements may be observable.[107] Despite finding a higher number of tissue-specific CNCs in the liver, the restricted sample size of only one sample in this study prevents any conclusion from being drawn regarding the accumulation of *de novo* hepatic CNCs in the mouse; however, the results indicate that the murine liver should be investigated further to see if this burden is consistently high.

The spleen shares many features of the murine liver, including high replication rate and

a role in immune function.[53,111] If replication or immune function were the main drivers of increased CNC formation, it would be expected that a similar pattern would be seen across the spleen samples in this study. This was not the case. Within individual mice, there tended to be fewer spleen-specific events than cerebellum- or liver-specific events. Additionally, if somatic hypermutation is a contributor to increased CNC formation in the genome, it would be expected that CNCs in the spleen would be frequent and overlap in immunologically-relevant gene regions.[12] Given the lack of spleen-specific merges in two detection methods and the lower number of spleen-specific merges found with PennCNV, it seems unlikely that somatic hypermutation is driving hotspots for CNC formation across the genome of the murine spleen. That being said, the spleen was not exempt from tissue-specific events occurring, though these events tended to be either singletons or merge-associated events that overlapped with events in non-splenic tissues.

The high number of cerebellum-specific events detected in this study contrasts with what would be expected if replication was a major contributor of *de novo* CNC formation, as the cerebellum was considered to be largely post-mitotic.[53] The cerebellum-specific events identified in this study support recent evidence that the genome of the cerebellum is more subject to genomic rearrangements than previously thought, and somatic copy number mosaicism exists even between neurons.[79,115,118] Despite these recent discoveries that make the occurrence of cerebellum-specific events less surprising, it is unknown as to exactly why cerebellum-specific events would not only match, but actually outnumber the events found to occur in only spleen or liver, two tissues with higher cell turnover rates as well as immunological function.

There are several possible reasons that may account for the observed differences between tissues beyond cellular replication. Array platforms, including the MDGA, require a certain amount of DNA for preparation and hybridization protocols. In addition to a sufficient signal:noise ratio, there must be enough cells in the amplified DNA sample that have a copy number change in a particular genomic region that to be reflected in the fluorescence intensity. For example, if only a select few cells contain a deletion in a region, any associated decrease

in fluorescence intensity will be drowned out by the signal produced by cells in the tissue containing two copies of the region. Additionally, what type of cells in the tissue are subject to copy number formation, and whether the cells containing a gain or loss are in sufficient quantities, are key factors in the detection of *de novo* copy number changes. Although the spleen is subject to somatic hypermutation, this may be limited to V(D)J recombination, deamination of cytosine, and base substitutions, rather than changes in copy number across the genome.[112] Compounded with a high cell turnover rate, if a copy number change in a particular region does form *de novo* in the spleen, the CNCs may not be present in a high enough number of cells and those cells may not live long enough to be in sufficient quantities to be detected by array methods. It is perhaps more likely that the spleen CNCs detected in this study are a result of developmental copy number changes in progenitor cells that give rise to clonal expansion of a particular CNC. The developmental source of *de novo* CNCs may be consistent with cerebellar CNCs as well, as the cerebellum is largely post-mitotic. The long-lived nature of neurons permits cell lineages containing CNCs that formed in development that, if not lethal, are able to persist in the tissue and thus be in sufficient quantity to be picked up by array methods. The recent identification of somatic copy number mosaicism among human neurons suggests a developmental source of genomic instability that may persist in a cell lineage.[79] Clonal expansion of neuron precursors and the very low cell turnover rate in the cerebellum are likely to contribute to the higher number of cerebellum-specific events identified in this study.

Although biases in copy number detection may stem from tissue-specific factors in DNA isolation procedures,[221] whether these biases could account for the prevalence of CNCs in the cerebella of the mice in this study is not yet known. Further work using alternative methods, such as quantitative PCR, to confirm the copy number status in the regions identified using the MDGA and subsequent analysis pipeline will give insight into the extent to which isolation biases may influence array-based copy number detection. If confirmed, the results from this study could have widespread implications in how researchers approach association studies and diagnostic procedures, particularly those that use DNA isolated from blood as a proxy for

variants occurring in the brain.

4.2 Merges may be representative of inherited variants, polymorphic regions, or hotspots for *de novo* CNC formation

Merges as described in this study are initially a catch-all for any collection of events across two or more samples that overlap reciprocally by 40% of the length of the smallest event or greater.[196] Thus, merges may represent inherited variants (which may be polymorphic at the population level), multiple *de novo* variants, or some mix of the two. CNVs are often inherited, and thus it is expected that events from different samples will overlap in their genomic regions, particularly in a group of highly related mice such as those included in this study. For example, the only merge to contain events from all fifteen samples was found with Partek events (chr.14). As it was found in all samples, it is likely an inherited variant present in the colony that differs from the C57BL/6J mice used as the reference. To contrast this merge, some merges contain events strictly from one tissue type across multiple mice, with no events called in the paired tissue for the mice represented in the merge. These tissue-specific merge events were very few and far between, and often included only two events. Despite low numbers, across all three detection methods, cerebellum-specific merges were more prevalent than spleen-specific merges (liver-specific merges were not possible as there was only one liver sample). This was most pronounced in the results from PennCNV, which called many more cerebellum-specific merges than spleen-specific merges. It is unknown whether this result is due to there being an overall higher number of cerebellum-specific events in these mice, and thus a higher likelihood of overlap by chance, or whether cerebellum-specific merges occur due to some kind of pressure resulting in localized changes in copy number.

Copy number events were detected in functionally relevant regions. Further inspection of the Partek-detected merge that contained events from all fifteen samples as described above (chr.14) can provide a case study in functional relevance. CNV studies examining the human

genome have identified copy number losses at the *LOXL2* locus on human chromosome 8.[14,222] The human gene *LOXL2* has been identified as a targeted gain leading to overexpression.[223] The increased expression of *LOXL2* was first linked to breast cancer,[224] followed by colon and esophageal cancers.[225] If confirmed, this deletion may have implications for phenotypic studies using mice from this colony. This merge serves as a preliminary example of the potential functional impact of CNVs identified in mice. It also underscores the importance of gene dosage for certain genes and the functional role changes in copy number can play in the development of disease. Although full functional characterization of all putative copy number events is beyond the scope of this study, the events detected provide the foundation for future validation and phenotypic characterization.

4.3 Generating trees based on shared copy number loci reveals different levels of genomic variation in a group of highly related samples

Visual inspection of the trees generated from genetic distance calculated from shared CNV loci reveals the samples cluster in the manner in which they would be expected based on the known relationships between the samples. When all samples are compared with each other in a pairwise fashion, different levels of relatedness are recovered, with more closely related samples having shorter between-sample branch lengths, while more distantly related samples have longer between-sample branch lengths. Tissues taken from the same mouse tend to have the shortest distances, tissues taken from mice that were littermates have longer distances, and tissues from more distantly related mice are further apart again in the trees. These expected patterns are illustrated in the Pedigree tree, and recovering these patterns in the Partek and PennCNV tree (and to a lesser extent, the ConsecN tree) gives confidence that the genetic distance calculated from the CNVs and CNCs identified by each method are reflecting the known genetic relatedness.

Tissues from the same mouse cluster together in the pedigree tree due to the coefficient of

relatedness (r) between tissues of the same mouse being imputed as one (assuming no variation between somatic tissues of the same mouse). The pedigree tree correctly places the relative distance between mice based on the estimated admixture between C57BL/6J and CBA/CaJ genetic background.[226]

The trees also support the identification of putative CNCs between tissues within an individual mouse. Although the two tissues sampled from one individual often cluster closer to each other than to any other samples (expected since these tissues derive from the same zygote), in the trees generated from copy number events detected by each method the two tissues from the same individual are actually separated by some degree of distance. This distance between tissue samples is in stark contrast to the zero-distance illustrated by the Pedigree tree, which represents the null hypothesis that tissues within an individual do not differ in copy number. The observed distances illustrate the findings that some events are not shared between tissues in an individual, and may ultimately indicate the presence of CNCs, supporting the hypothesis that somatic copy number mosaicism is present and detectable between tissues in the same mouse.

4.4 The novel pipeline for array-based data analysis provides an experimental framework for detecting somatic copy number mosaicism

Somatic copy number mosaicism can be detected using the MDGA by employing careful experimental design that includes sampling more than one tissue per individual. The one-sample-per-array design required when using array-based genotyping platforms such as the MDGA permits individual samples to be analyzed in comparison to a selected reference outside of the samples included in the experiment, as compared to array comparative genomic hybridization (aCGH) in which two samples compete on the same array, or whole genome sequencing which is costly to perform on multiple tissues. Using a genotyping array offers the opportunity to expand the sample size of a particular study while maintaining a constant refer-

ence. The CNCs identified in the present study are identified as such due to the absence of a change in copy number in the same genomic region in the corresponding paired tissue included in this study. Whether or not a given multi-tissue study will detect evidence of somatic copy number mosaicism depends on the tissues chosen for analysis.[57,135]

4.4.1 The detection of losses and gains differs between detection methods

In total, this study found more copy number losses than gains, which is consistent with the literature.[14,29,48,50,79,92,113] While ConsecN is inherently restricted to only detect putative losses, Partek and PennCNV are capable of detecting both gains and losses (although there were no PennCNV gains that remained in the dataset after filtering). The detection of more gains than losses with Partek is contrary to previous CNV detection studies using the software.[60,137] The two HMM-based methods that should both be capable of detecting gains and losses are thus at odds in this particular metric, despite calling copy number events from the same array data.

4.4.2 The lengths and marker densities of putative copy number events are highly correlated

In their genotyping array-based analysis of murine CNVs, Henrichsen *et al* used an in-house HMM-based approach and identified copy number events that ranged in size from 43 kb to over 3.3 Mb, with a size median of 61 kb.[63] The median length of Partek events fall within this range (median length of 2.2 Mb), with ConsecN and PennCNV calling shorter events (median lengths of 335 bp and 2.49 kb, respectively). ConsecN was not forced to include more than two consecutive markers, allowing it to call smaller putative events in genomic regions with lower inter-SNP spacing. Likewise, the authors of the PennCNV program state that the ability of PennCNV to call relatively shorter events provides evidence that the breakpoints are

being called at "higher resolution",[194] which could account for the lower median event length. The size distribution patterns found in this study reflect the CNV size distributions previously found in humans and rhesus macaques, suggesting that events in larger size categories are more infrequent than those closer to the lower size boundary of 1 kb in length.[66,145] As such, the frequency of smaller events detected in this study may be more indicative of the algorithms' abilities to allow for smaller events to be called, as well as the resolution of the array itself.

The application of marker density as a filtering step contributes to lowering the median event length for each method as well. All three methods called events that were subsequently removed due to low marker density. Marker density appears highly correlated with event length, with marker density decreasing as length increases. The right-skewed nature of the marker density distributions for each calling method demonstrates that most events are well represented by SNPs, and the frequency of very large events tagged by few markers is very low. Indeed, fewer than 12% of the total events called in each method are removed due to lower marker density. The strongest correlation between marker density and length was found for events called by ConsecN. Partek and PennCNV events showed less correlation between marker density and event length than ConsecN. One interpretation of this result is that the ConsecN method is a better predictor of inter-SNP distance rather than true copy number variants. Since marker density is calculated using length, it is likely that the strong correlation between marker density and length is a result of the very few SNPs present in the calculation having little modifying effect on this proportional relationship.

Applying a filtering step based on marker density seeks to limit downstream analysis to only events that are well tagged by genetic markers to increase confidence that the event is real (that is, exists in a changed copy number state in the genome *in vivo*). Although requiring a minimum number of markers used to call a copy number event provides a first step in ensuring multiple markers are reporting similar copy number states across a segment of the genome, the addition of a marker density cutoff (in this study, the cutoff value is 1.3 markers per 10 kb) increases the reliability of the calls being made.[60] If the events detected from array

data using a particular algorithm are real, increasing the array density (number of genomic regions interrogated by oligonucleotides on the array) will lead to a higher event marker density, since more markers will be used to call the copy number in a given genomic region. Events that were dropped from the present analyses due to low marker density may be recovered if future analyses interrogate more loci across the genome; for example, by including the fluorescence intensities reported from the invariant genomic probes already present on the MDGA. Additionally, interrogating more loci may permit increasing the minimum number of markers required to call an event from the three minimum markers used currently.

4.4.3 Putative copy number events show variation in GC content

The medians of GC content of the copy number events found by each method in this study (ConsecN 44.0%, Partek 41.15%, PennCNV 38.1%) do not deviate greatly from the overall GC content of the mouse genome, which is about 42%.[43] What stands out is the variation around this median as well as the variability between the three calling methods. The GC content of murine CNVs has not been explicitly reported before this study. Despite the lack of an established range of GC content in murine CNVs, it had been expected that the GC content of mouse CNVs would be near or slightly above the genome-wide GC content average based on CNV studies in humans and dogs, as well as based on the known correlation between GC content and recombination (a known factor in CNV formation) in mice.[29,30,131,136] Here, more events with higher GC content were called by ConsecN and PennCNV; however, these same methods show a large number of events with lower GC content as well. Partek events show a tighter range of GC content values, which may be indicative of a bias in the Partek algorithm towards calling fewer and longer events, eliminating regional variation in GC content.

4.4.4 The length of a copy number event may influence whether it is classified as a singleton or a merge

Due to the requirement of percent overlap between events to call a merge, the length of the events called by a given method may influence whether an event is classified as a singleton or a merge. ConsecN and PennCNV called shorter events than Partek. Subsequent analysis revealed no difference in length between singletons and merge-associated events called by ConsecN and PennCNV. Partek merge-associated events were significantly longer than singletons. This is in contrast to results found in previous studies that have compared the lengths of singletons to the lengths of CNVs found in multiple samples. These studies found that singletons were longer in length than the recurrent events, especially in cases of singleton deletions.[141,143]

After separating events based on copy number state (gains or losses), it was found that singletons called within ConsecN and PennCNV, despite all being called as deletions, were not found to differ in length from merge-associated events called by the same method. Given that events called by Partek were the longest, and that Partek merge-associated events were longer than singletons contrary to previous studies, it is possible that the overall length of the events called influences whether two or more events are found to overlap with one another.

4.5 Construction of an accessible database of previously discovered murine CNVs and CNVRs permits researchers to identify common and *de novo* variants

Compiling the genomic coordinates of previously detected murine CNVs and CNVRs into an accessible and interactive database resulted in a summary of six studies (accessed May 2012). The database now contains 17,128 entries representing several inbred strains of the laboratory mouse, including C57BL/6J. By using other bioinformatic tools, such as BEDTools

Suite and the UCSC Genome Browser, the entries in the constructed database can be explored for their genomic context as well as compared to other lists of genomic coordinates to assess overlap with previously documented CNVs, genes, or other genomic features.

It is essential that researchers do not automatically discount copy number events identified in new studies that do not overlap with entries in such a database. Although ascertainment bias is a known source of sampling error in genomics research,[169] common CNVs (despite occurring at a higher frequency in the population) are paradoxically more difficult to detect than rare variants using arrays.[137] If several arrays in a dataset contain variants in the same region, it is more difficult during normalization procedures to establish a clear reference signal, making it more difficult for algorithms to call departures from the reference.[137] It is important to be aware of the limitations such a database can provide, and understand that not all entries are equal in terms of experimental design, platform, or calling algorithm. Not all entries will have been confirmed with secondary methods. Even the database constructed in the present study contains events discovered by very different procedures including aCGH and paired-end sequencing. None of the current entries in the database were found using genotyping arrays, and so there are no MDGA-ascertained variants to which results in this study can be compared directly.

The importance of the database stands in the continued accumulation of and reference to results from many studies. Using the present study as an example, the ability to compare detected CNVs in a given inbred strain to the results from previous studies using the same strain provides an estimate of the current method's success in detecting known variants. Additionally, such a comparison provides an opportunity to identify previously undocumented variants that the new method is capable of detecting. Results can be replicated for a given strain across multiple studies using different copy number detection and platforms. Aggregating results from many discovery studies may provide the opportunity for meta-analyses to observe patterns that emerge only with a multitude of samples. Future studies that observe unexpected patterns SNP genotypes or gene expression may find the underlying cause is associated with copy number

variation if genomic coordinates are queried in a database of murine CNVs. Although constant curation of such a database requires time and resources, it is hoped that as the accuracy of CNV detection methods improves, so too can the accuracy of the information indexed in such repositories.

4.6 Advantages and limitations of microarray-based CNV discovery

The resolution of genotyping array platforms continues to increase, allowing for the inclusion of more probes, both for interrogating previously missed loci as well as increasing the accuracy with which loci are analysed. Genotyping arrays have also become more affordable, and are often the most economically feasible method for CNV detection on a per-sample basis. Despite these attractive features that make genotyping arrays among the most popular platforms for CNV detection, they do come with their own limitations. It is important to understand these limitations and take known sources of bias and error into account when performing CNV analyses from genotyping array data.

By using only probes for known SNPs in this study, it is likely that the number of CNVs detected here is lower than what may actually be present in the samples. Like the Affymetrix® Human 6.0 Array, the Mouse Diversity Genotyping Array contains an additional 900,000 IGPs that interrogate invariant regions in the genome.[168] Although excluded from the present study on the grounds that the ConsecN and Partek methods were unable to use these probes, future CNV detection work with the MDGA may be improved once the IGPs have been thoroughly vetted for performance and sequence accuracy, thereby increasing the resolution of the array and opening up previously poorly tagged regions of the mouse genome to copy number analysis.

Preprocessing steps used in the analysis of array data may lead to underestimation of copy number events. The quantile normalization step used in the RMA background correction applied here in the application of Partek and PennCNV has been shown to yield a high degree of

inter-array correlation, and this artificial correlation is particularly noticeable at small sample sizes (between 2 and 100 arrays).[173] As a smaller number of arrays was used in the normalization step in the application of Partek in this study, the Partek results presented here may have been more affected than PennCNV by this reduction in overall fluorescence variation between arrays. PennCNV was run with a high enough number of arrays that the artificial correlation effect would have been negligible.[173]

The sole sample that failed to pass the minimum genotyping call rate was the spleen of mouse 300.6. Of the samples that passed the minimum genotyping call rate, the liver and the spleen from mouse 900.3 had the lowest call rates (98.99% and 99.11%, respectively). This was to be expected as mouse 900.3 was an F1 hybrid of C57BL/6J and CBA/CaJ, and it has been established that increased heterozygosity within a mouse makes genotyping with an array platform more difficult.[169,181] Although using a subset of SNP probes on the array results in an increased average inter-SNP distance as compared to the manufacturer's advertised mean interprobe distance of 4.3 kb, the high genotyping call rates indicate that the filtered SNP list used here has successfully excluded poorly performing probes, and thus the fluorescence data being reported by the remaining probes is more reliable.[202]

4.7 Evaluation of the novel ConsecN method for calling putative copy number deletions

The variation in fluorescence intensity of SNP loci designated as "NoCalls" indicates that a "NoCall" designation does not always indicate insufficient fluorescence intensity. Without assessing the genotype clustering performance of each SNP, this method may be calling false-positives in terms of identifying deletions. The probability of two "NoCalls" occurring consecutively along a chromosome as presented here is strictly a theoretical model and does not take into account the biological phenomenon of genetic linkage, by which two genetic markers are more likely to be inherited together if they are physically closer to one another along a chromo-

some. The physical inter-SNP distance was not considered in the initial step of identifying the two or more consecutive "NoCalls", but was accounted for in the downstream step of filtering events based on marker density.

4.7.1 Characteristics of ConsecN events are not consistent with events called by other methods

Of the three methods used to detect putative CNVs in this study, ConsecN appears to be the outlier. Although genotyping calls have previously been used to infer copy number deletions,[10,183] these approaches yielded the detection of losses in the kilobase range. The putative events called by ConsecN are short, both relative to the events called by Partek and PennCNV and relative to the generally accepted minimum size of a CNV (1 kb or greater). Unlike Partek and PennCNV, these short events were permitted to be included in this study by relaxing length requirements for this method due to the limitations of seeking consecutive SNPs with failed genotyping calls as opposed to consecutive SNPs with lower fluorescence intensity. If dropping the minimum size restriction was the only reason for the inclusion of short events, and if ConsecN was indeed capable of calling real CNVs, it would be expected that the lower end of the length distribution would be more populated with events, without the loss of events in the higher end of the length distribution. This was not the case in this study. Despite calling more short events, ConsecN failed to detect events that would be of lengths that follow the generally accepted definition of CNVs being 1 kb or greater in length. For the longer events that ConsecN did call, we can look to marker density to estimate how confident we can be that these are real events. The relationship between event length and marker density was the strongest for ConsecN. This, in combination with the tendency to call shorter events, suggests that ConsecN cannot be relied upon to call large structural variants with confidence, as the marker density drops off with large events.

Not only did ConsecN call fewer large events than the other two methods, the events that

ConsecN did call did not overlap well with previously found events documented in the constructed database, with only 12.55% of ConsecN events overlapping with database entries. Of the database entries, ConsecN had the lowest rate of detection, with only 0.42% of entries in the database overlapping with ConsecN events. Given that all of the mice in this study have some degree of C57BL/6J background, and that the database included C57BL/6J events detected by Henrichsen *et al*,[63] it would be expected that more C57BL/6J events would have be ascertained by ConsecN. There is a strong possibility that ConsecN is missing events that should have been detected.

Additionally, the GC content of ConsecN events is highly variable, ranging from 22.50% to 60.82%. ConsecN picks up events with very low GC content, even lower than the average GC content of the mouse genome which is around 42%.[43] Although little is known about the GC content of murine CNVs, we would expect the GC content to be on the higher side from what we know about CNVs in other mammals. Human CNVs tend to occur in GC-rich regions,[132] as do CNVs detected in the dog.[29] The short lengths of the ConsecN events is most likely contributing to the wide range of percent GC content observed with this method.

When CNV loci shared between samples are used to calculate genetic distance, events called by ConsecN result in a difference matrix that is similar only to the PennCNV matrix. Most importantly, the ConsecN matrix is dissimilar to the Pedigree matrix which was generated using the known genetic relatedness of the closely related samples. The resulting trees further underscore the separation of ConsecN events from events called by Partek and PennCNV. As with the difference matrices, the ConsecN tree was similar in topology only to PennCNV, and dissimilar to the Pedigree and Partek trees.

Consecutive SNPs genotyped as homozygous may be indicative of small deletions below the standard minimum CNV size of 1 kb, as has been found previously in analysis of human SNPs.[184] In humans, the high degree of heterozygosity permits the inclusion of homozygous runs in deletion analysis; however, this is much more difficult to perform in laboratory mice

due to the extensive inbreeding resulting in mostly homozygous SNPs across the genome. The high degree of homozygosity in laboratory mice limits the use of homozygous runs for deletion detection. The trend towards increasing diversity in laboratory mouse populations with the Diversity Outbred project may make screening for putative deletions using only genotype calls in the future more robust with the addition of consecutive homozygous SNPs as candidate regions. At this time, using only "NoCalls" with the MDGA returns events uncharacteristic of what would be expected from CNVs (in length, GC content, and overlap with database entries), ConsecN does not appear to be an ideal method to detect putative copy number losses. Instead, a number of the smaller events (under 1 kb) called by ConsecN may be small deletions.

4.8 Evaluation of Partek and PennCNV as Hidden Markov Model-based methods of calling copy number

Of the three methods used in this study to detect copy number, Partek and PennCNV are the most similar to each other in that they both employ Hidden Markov Model-based algorithms that use fluorescence data obtained from all SNP probes on the filtered SNP list, while ConsecN only counts SNP loci for which genotyping failed. Additionally, both Partek and PennCNV were run with similar user-set parameters including normalization and background correction procedures as well as minimum number of markers used to call a copy number event. Partek and PennCNV had the highest concordance as measured by the number of overlapping loci called by the two methods. Despite these similarities, the events called by Partek and PennCNV were not equivalent, obviated by the many more events called by PennCNV than Partek. Further inspection of the copy number events called by each of the methods revealed that Partek called fewer but longer events than PennCNV. This pattern is consistent with a previous study that used Partek and PennCNV in addition to three other algorithms to call copy number from Illumina arrays, which also found Partek called the fewest but largest events. [195]

Comparing the genetic distance matrices and trees associated with each calling method to

the matrix and tree calculated from the pedigree allows the assessment of a given method's ability to detect copy number events that reflect the expected genetic distance between samples. The difference matrix calculated from Partek's shared and unshared CNVs was the only difference matrix found to be statistically similar to the Pedigree tree. Additionally, the resulting Partek tree (as well as the PennCNV tree) was found to be similar to the Pedigree tree. These results suggest that Partek may be the most likely of the three methods to be calling CNVs that reflect the known genetic relationships between samples. It is important to remember here that the Pedigree tree is forced to cluster tissues from the same mouse together, as tissue samples from the same mouse were imputed to have a coefficient of relatedness value of zero. As a result, if a calling method is less likely to pick up on CNCs between tissues, it may be more apt to appear topologically similar to the Pedigree tree. This appears be the case with Partek, which called more events shared between tissues within individual mice than events that were in only one of the two tissues. That Partek may be less likely to pick up *de novo* CNCs is strengthened by the fact that Partek called the highest percentage of copy number events that overlap previously discovered CNVs documented in the database with (95.98% of Partek events). Conversely, even though PennCNV called more events, a lower percentage of these events were found to overlap the database (63.52% of PennCNV events). Taking this result together with the finding that the PennCNV tree is topologically similar to the Pedigree tree suggests that PennCNV may be calling smaller events that are being missed by Partek.

The Partek-Pedigree and PennCNV-Pedigree tree similarities in the absence of a Partek-PennCNV tree agreement, in addition to a low degree of concordance in calls between methods, suggests that there are inherent biases in the calling methods. Pinto *et al* suggested that the different results obtained by using various methods indicate that the methods themselves may not be better or worse overall in the detection of CNVs, but instead offer "different strengths".[51] The "different strengths" of Partek and PennCNV may be evident by looking closer at the characterizations of the copy number events called by each method. Despite calling longer events overall, Partek was the only method to find a difference in lengths between singletons and

merge-associated events. Although both Partek and PennCNV events had strong correlations between marker density and event length, splitting events into singletons and merge-associated events revealed that while marker density was higher in merge-associated events than singletons for Partek, the opposite was true for PennCNV. This suggests that Partek may be less likely to call copy number events in genomic regions that are more poorly tagged with SNPs. Additionally, the narrow range of GC content in Partek events may indicate a bias in the genomic regions in which the algorithm is capable of detecting events. The smaller events called by PennCNV and the shift away from regions previously found to harbour CNVs may indicate that PennCNV is capable of detecting previously unknown or *de novo* variants.

4.8.1 The effect of reference selection on HMM-driven copy number detection

With all other factors being equal (sample data, HMM parameters) or as similar as possible (quantile normalization with Partek and sketch quantile normalization with PennCNV), the difference in the CNV calls made by these two software packages is most likely due to the choice of reference used in each approach. Since CNVs are called as departures from the constructed reference, the composition of the data used to construct the reference may influence whether an event is called in any given genomic region in an experimental sample. With mice, the reference genome is constructed from the C57BL/6J inbred strain.[43] The MDGA was designed with this in mind, with the A allele for nearly all SNPs based on the genotype carried by C57BL/6J.[168] Constructing a copy number reference based on CEL files from pure C57BL/6J mice, as was done in Partek in this study, followed this C57BL/6J-based approach in order to call CNVs that differed from the C57BL/6J reference. As noted by Marioni *et al*, the choice of reference sample influences whether a segment of DNA is regarded as a gain, loss, or "normal;"[227] and this is evident when comparing the resulting Partek calls to PennCNV calls. Although PennCNV called CNVs from the same sample set as Partek and run with similar

HMM parameters, the present application of PennCNV used a very different composition of CEL files as the reference. PennCNV was run with the 335 CGD CEL files, while Partek was run with the seven pure C57BL/6J CGD CEL files. The difference in the two subsets of CGD CEL files used lies in both the number of CEL files and the genetic diversity of the mice represented by those files.

When tackling the issue of reference diversity, it is useful to compare mouse genomic research to human genomic research. As a species, humans are genetically diverse, even within a single population, and individual genomes have a high degree of heterozygosity. This is in contrast to the controlled genetic background of inbred strains of laboratory mice, in which individual mice from one strain are nearly identical to one another, and the genomes of individuals are highly homozygous. Genomic research in humans often makes use of the HapMap population as a reference set. The first study to use the HapMap population to aid in the detection of CNVs was performed by Komura *et al* and used a novel algorithm to detect CNVs within the HapMap population.[162] The introduction of the MDGA as a copy number detection platform for the laboratory mouse has permitted the generation of CEL files from samples taken from hundreds of mice of different genetic backgrounds. These are the samples which make up the CGD set of CEL files used in this study. By using the 335 CGD CEL files that passed the genotyping step, we can best approximate the diversity seen in the HapMap population but for a diverse set of mice. The drastic differences between Partek and PennCNV results, given the same array data, underscore the necessity of selecting the appropriate reference number and composition.

4.9 Future directions

Scherer *et al.* (2007) stress that genome-wide discovery methods, including array-based methods, are best interpreted as screening assays to find regions of the genome that have an "increased probability" of being variable in copy number.[7] The independent validation of CNVs

discovered with oligonucleotide arrays presents a new set of challenges, as validation methods all have their own drawbacks in inherent biases, limitations, and cost. However, genotyping arrays continue to be valuable tools for CNV, and now, CNC, discovery experiments. The identification of putative tissue-specific regions of *de novo* CNC formation in mice may have strong implications for our understanding of how the genome may change over the lifetime of an individual, as well as for how we diagnoses genetic diseases.

The dynamic nature of the fields of genomics and bioinformatics means that methods and interpretations must constantly be re-evaluated and revised as the associated technologies are advanced and optimized. The exclusion of poorly performing probe sets, updates to the sequence and annotation of the reference genome, and better algorithms for detection and background correction not only lead to improved research methods and more accurate results, but also invalidates the accuracy of previous entries into databases to which new data are compared. Updating probe set annotations and filtering out poorly performing probe sets can have a significant impact on array performance and the interpretation of array results.[174,202]

4.10 Summary

Acknowledging that CNVs are both inherited and can form *de novo* in the germline and somatic tissues is absolutely necessary to increase accuracy in associative studies. The impact of CNVs and CNCs on variation in phenotypes, both normal and pathogenic, renders the development and optimization of experimental frameworks in which to investigate somatic copy number mosaicism a priority. The novelty of the Mouse Diversity Genotyping Array and the complexity of studying somatic copy number mosaicism presented the challenge to develop a new analysis pipeline, which was solved here by employing novel approaches and custom software to result in three different methods for identifying and analyzing CNVs and CNCs in the mouse. Despite low concordance rates between the three CNV detection methods as demonstrated by comparing the CNV loci called in each method, the abilities of Partek and PennCNV

to call CNVs that can be used to recapitulate the known genetic relationships between samples used in this study suggests that the methods are still capable of detecting true biological variants. The discovery of putative hotspots for recurrent tissue-specific CNC formation in the mouse, particularly in the cerebellum, is a significant biological finding that sheds light on genomic changes in the brain, and may ultimately lead to a better understanding of neurological disorders that have not yet been associated with underlying genetic variation. The characterization of somatic copy number mosaicism in the laboratory mouse is essential, especially when using the mouse to study mutational mechanisms and model human diseases. This study serves to provide a robust framework within which future somatic copy number mosaicism studies can be designed and performed, highlighting key experimental procedures including using filtered array probes, comparing calls to a database of known variants, and identifying steps that introduce variation between calling methods such as reference selection. As a result, patterns in somatic copy number mosaicism can be identified and further investigated for their contribution to the overall genomic diversity in *Mus musculus*.

References

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 2001; 409(6822): 860–921. doi:10.1038/35057062.

2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. The sequence of the human genome. *Science* 2001; 291(5507): 1304–1351. doi:10.1126/science.1058040.

3. The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001; 409(6822): 928–33. doi:10.1038/35057149.

4. Tanaka T. The International HapMap Project. *The International HapMap Consortium* 2003; 426: 789–796. doi:10.1038/nature02168.

5. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; 449(7164): 851–61. doi:10.1038/nature06258.

6. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. Complement factor H polymorphism in age-related macular degeneration. *Science* 2005; 308(5720): 385–9. doi: 10.1126/science.1109557.

7. Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, et al. Challenges and standards in integrating surveys of structural variation. *Nat Genet* 2007; 39(7 Suppl): S7–15. doi: 10.1038/ng2093.

8. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet* 2006; 7(2): 85–97. doi:10.1038/nrg1767.

9. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, et al. Copy number variation: new insights in genome diversity. *Genome Res* 2006; 16(8): 949–61. doi:10.1101/gr.3677206.

10. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* 2006; 38(1): 75–81. doi:10.1038/ng1697.

11. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, et al. Detection of large-scale variation in the human genome. *Nat Genet* 2004; 36(9): 949–51. doi:10.1038/ng1416.

12. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, et al. Large-scale copy number polymorphism in the human genome. *Science* 2004; 305: 525–528. doi:10.1126/science.1098918.

13. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, et al. Assessing the impact of population stratification on genetic association studies. *Nat Genet* 2004; 36(4): 388–93. doi:10.1038/ng1333.

14. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, et al. Origins and functional impact of copy number variation in the human genome. *Nature* 2010; 464(7289): 704–12. doi:10.1038/nature08516.

15. Girirajan S, Campbell CD, Eichler EE. Human copy number variation and complex genetic disease. *Annu Rev Genet* 2011; 45: 203–26. doi:10.1146/annurev-genet-102209-163544.

16. Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, et al. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* 2011; 9(4): e1001046. doi:10.1371/journal.pbio.1001046.

17. Lee C, Iafrate J, Brothman AR. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat Genet* 2007; 39(7 Suppl): S48–54. doi: 10.1038/ng2092.

18. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 2007; 315(5813): 848–53. doi:10.1126/science.1136678.

19. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. Global variation in copy number in the human genome. *Nature* 2006; 444(7118): 444–54. doi:10.1038/nature05329.

20. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* 2008; 453(7191): 56–64. doi:10.1038/nature06862.

21. Tsuang DW, Millard SP, Ely B, Chi P, Wang K, et al. The effect of algorithms on copy number variant detection. *PLoS One* 2010; 5(12): 10. doi:10.1371/journal.pone.0014456.

22. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, et al. Mapping copy number variation by population-scale genome sequencing. *Nature* 2011; 470(7332): 59–65. doi: 10.1038/nature09708.

23. Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, et al. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am J Hum Genet* 2006; 79(2): 275–90. doi:10.1086/505653.

24. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. *Nat Rev Genet* 2009; 10(8): 551–64. doi:10.1038/nrg2593.

25. Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, et al. De novo rates and selection of large copy number variation. *Genome Res* 2010; 20(11): 1469–1481. doi:10.1101/gr.107680.110.20.

26. Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. *Genetics* 2000; 156(1): 297–304.

27. Lee AS, Gutiérrez-Arcelus M, Perry GH, Vallender EJ, Johnson WE, et al. Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. *Hum Mol Genet* 2008; 17(8): 1127–36. doi:10.1093/hmg/ddn002.

28. Guryev V, Saar K, Adamovic T, Verheul M, van Heesch SA, et al. Distribution and functional impact of DNA copy number variation in the rat. *Nat Genet* 2008; 40(5): 538–45. doi:10.1038/ng.141.

29. Berglund J, Nevalainen EM, Molin AM, Perloski M, Lupa TLC, et al. Novel origins of copy number variation in the dog genome. *Genome Biol* 2012; 13(8): R73. doi:10.1186/gb-2012-13-8-r73.

30. Fadista J, Thomsen B, Holm LE, Bendixen C. Copy number variation in the bovine genome. *BMC Genomics* 2010; 11: 284. doi:10.1186/1471-2164-11-284.

31. Fontanesi L, Beretti F, Martelli PL, Colombo M, Dall'olio S, et al. A first comparative map of copy number variations in the sheep genome. *Genomics* 2011; 97(3): 158–65. doi:10.1016/j.ygeno.2010.11.005.

32. Fontanesi L, Martelli PL, Beretti F, Riggio V, Dall'Olio S, et al. An initial comparative map of copy number variations in the goat (Capra hircus) genome. *BMC Genomics* 2010; 11(1): 639. doi:10.1186/1471-2164-11-639.

33. Ramayo-Caldas Y, Castelló A, Pena RN, Alves E, Mercadé A, et al. Copy number variation in the porcine genome inferred from a 60 k SNP BeadChip. *BMC Genomics* 2010; 11(1): 593. doi:10.1186/1471-2164-11-593.

34. Wang L, Liu X, Zhang L, Yan H, Luo W, et al. Genome-wide copy number variations inferred from SNP genotyping arrays using a large white and minzhu intercross population. *PLoS One* 2013; 8(10): e74879. doi:10.1371/journal.pone.0074879.

35. Griffin DK, Robertson LB, Tempest HG, Vignal A, Fillon V, et al. Whole genome comparative studies between chicken and turkey and their implications for avian genome evolution. *BMC Genomics* 2008; 9: 168. doi:10.1186/1471-2164-9-168.

36. Wang X, Nahashon S, Feaster TK, Bohannon-Stewart A, Adefope N. An initial map of chromosomal segmental copy number variations in the chicken. *BMC Genomics* 2010; 11: 351. doi:10.1186/1471-2164-11-351.

37. Skinner BM, Robertson LBW, Tempest HG, Langley EJ, Ioannou D, et al. Comparative genomics in chicken and Pekin duck using FISH mapping and microarray analysis. *BMC Genomics* 2009; 10: 357. doi:10.1186/1471-2164-10-357.

38. Dopman EB, Hartl DL. A portrait of copy-number polymorphism in Drosophila melanogaster. *Proc Natl Acad Sci USA* 2007; 104(50): 19920–5. doi:10.1073/pnas.0709888104.

39. Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. *Science* 2008; 320(5883): 1629–31. doi:10.1126/science.1158078.

40. Muñoz Amatriaín M, Eichten SR, Wicker T, Richmond TA, Mascher M, et al. Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Genome Biol* 2013; 14(6): R58. doi:10.1186/gb-2013-14-6-r58.

41. Zhang H, Zeidler AFB, Song W, Puccia CM, Malc E, et al. Gene copy-number variation in haploid and diploid strains of the yeast Saccharomyces cerevisiae. *Genetics* 2013; 193(3): 785–801. doi:10.1534/genetics.112.146522.

42. Adams DJ, Dermitzakis ET, Cox T, Smith J, Davies R, et al. Complex haplotypes, copy number polymorphisms and coding variation in two recently divergent mouse strains. *Nat Genet* 2005; 37(5): 532–6. doi:10.1038/ng1551.

43. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002; 420(6915): 520–62. doi: 10.1038/nature01262.

44. Fenner JN. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol* 2005; 128(2): 415–23. doi: 10.1002/ajpa.20188.

45. Matsumura S, Forster P. Generation time and effective population size in Polar Eskimos. *Proc R Soc B* 2008; 275(1642): 1501–8. doi:10.1098/rspb.2007.1724.

46. Breeding Strategies for Maintaining Colonies of Laboratory Mice. Technical Report, The Jackson Laboratory, 2009.

47. Hutchins LN, Ding Y, Szatkiewicz JP, Von Smith R, Yang H, et al. CGDSNPdb: a database resource for error-checked and imputed mouse SNPs. *Database* 2010; 2010. doi:10.1093/database/baq008.

48. Graubert TA, Cahan P, Edwin D, Selzer RR, Richmond Ta, et al. A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genet* 2007; 3(1): e3. doi:10.1371/journal.pgen.0030003.

49. Cutler G, Kassner PD. Copy number variation in the mouse genome: implications for

the mouse as a model organism for human disease. *Cytogenet Genome Res* 2008; 123: 297–306. doi:10.1159/000184721.

50. Agam A, Yalcin B, Bhomra A, Cubin M, Webber C, et al. Elusive copy number variation in the mouse genome. *PLoS One* 2010; 5(9): e12839. doi:10.1371/journal.pone.0012839.

51. Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol* 2011; 29(6): 512–20. doi:10.1038/nbt.1852.

52. Hester SD, Reid L, Nowak N, Jones WD, Parker JS, et al. Comparison of comparative genomic hybridization technologies across microarray platforms. *J Biomol Tech* 2009; 20(2): 135–51.

53. Cameron IL. Cell renewal in the organs and tissues of the nongrowing adult mouse. *Texas Rep Biol Med* 1970; 28(3): 203–248.

54. Ono T, Ikehata H, Pithani VP, Uehara Y, Chen Y, et al. Spontaneous mutations in digestive tract of old mice show tissue-specific patterns of genomic instability. *Cancer Res* 2004; 64(19): 6919–23. doi:10.1158/0008-5472.CAN-04-1476.

55. Youssoufian H, Pyeritz RE. Mechanisms and consequences of somatic mosaicism in humans. *Nat Rev Genet* 2002; 3(10): 748–58. doi:10.1038/nrg906.

56. Shirley MD, Tang H, Gallione CJ, Baugher JD, Frelin LP, et al. Sturge-Weber syndrome and port-wine stains caused by somatic mutation in GNAQ. *N Engl J Med* 2013; 368(21): 1971–9. doi:10.1056/NEJMoa1213507.

57. Scavetta RJ, Tautz D. Copy number changes of CNV regions in intersubspecific crosses of the house mouse. *Mol Biol Evol* 2010; 27(8): 1845–56. doi:10.1093/molbev/msq064.

58. Jacobs KB, Yeager M, Zhou W, Wacholder S, Wang Z, et al. Detectable clonal mosaicism

and its relationship to aging and cancer. *Nat Genet* 2012; 44(6): 651–8. doi:10.1038/ng.2270.

59. Lupski JR. Genome Mosaicism - One Human, Multiple Genomes. *Science* 2013; 341(6144): 358–359. doi:10.1126/science.1239503.

60. Pamphlett R, Morahan JM, Luquin N, Yu B. Looking for differences in copy number between blood and brain in sporadic amyotrophic lateral sclerosis. *Muscle Nerve* 2011; 44(4): 492–8. doi:10.1002/mus.22095.

61. Arlt MF, Mulle JG, Schaibley VM, Ragland RL, Durkin SG, et al. Replication stress induces genome-wide copy number changes in human cells that resemble polymorphic and pathogenic variants. *Am J Hum Genet* 2009; 84(3): 339–50. doi:10.1016/j.ajhg.2009.01.024.

62. Piotrowski A, Bruder CEG, Andersson R, Diaz de Ståhl T, Menzel U, et al. Somatic mosaicism for copy number variation in differentiated human tissues. *Hum Mutat* 2008; 29(9): 1118–24. doi:10.1002/humu.20815.

63. Henrichsen CN, Vinckenbosch N, Zöllner S, Chaignat E, Pradervand S, et al. Segmental copy number variation shapes tissue transcriptomes. *Nat Genet* 2009; 41(4): 424–9. doi:10.1038/ng.345.

64. Bánlaki Z, Doleschall M, Rajczy K, Fust G, Szilágyi A. Fine-tuned characterization of RCCX copy number variants and their relationship with extended MHC haplotypes. *Genes Immun* 2012; 13(7): 530–5. doi:10.1038/gene.2012.29.

65. Nguyen DQ, Webber C, Ponting CP. Bias of selection on human copy-number variants. *PLoS Genetics* 2006; 2(2): e20. doi:10.1371/journal.pgen.0020020.

66. Omer G, Babb PL, Iskow R, Zhu Q, Shi X, et al. Refinement of primate CNV hotspots

identifies candidate genomic regions evolving under positive selection. *Genome Biol* 2011; 12(5): R52. doi:10.1186/gb-2011-12-5-r52.

67. Vissers LELM, Veltman JA, van Kessel AG, Brunner HG. Identification of disease genes by whole genome CGH arrays. *Hum Mol Genet* 2005; 14(2): R215–23. doi:10.1093/hmg/ddi268.

68. Lupski JR, Wise CA, Kuwano A, Pentao L, Parke JT, et al. Gene dosage is a mechanism for Charcot-Marie-Tooth disease type 1A. *Nat Genet* 1992; 1: 29–33.

69. Tybulewicz VLJ, Fisher EMC. New techniques to understand chromosome dosage: mouse models of aneuploidy. *Hum Mol Genet* 2006; 15(2): R103–9. doi:10.1093/hmg/ddl179.

70. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DCY, et al. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med* 2010; 362(13): 1181–91. doi:10.1056/NEJMoa0908094.

71. Stefansson H, Rujescu D, Cichon S, Pietiläinen OPH, Ingason A, et al. Large recurrent microdeletions associated with schizophrenia. *Nature* 2008; 455(7210): 232–6. doi:10.1038/nature07229.

72. Maiti S, Kumar KHBG, Castellani Ca, O'Reilly R, Singh SM. Ontogenetic de novo copy number variations (CNVs) as a source of genetic individuality: studies on two families with MZD twins for schizophrenia. *PLoS One* 2011; 6(3): e17125. doi:10.1371/journal.pone.0017125.

73. McMichael G, Girirajan S, Moreno-De-Luca A, Gecz J, Shard C, et al. Rare copy number variation in cerebral palsy. *Eur J Hum Genet* 2013; (January). doi:10.1038/ejhg.2013.93.

74. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, et al. Integrated detection

and population-genetic analysis of SNPs and copy number variation. *Nat Genet* 2008; 40(10): 1166–74. doi:10.1038/ng.238.

75. de Cid R, Riveira-Munoz E, Zeeuwen PLJM, Robarge J, Liao W, et al. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat Genet* 2009; 41(2): 211–5. doi:10.1038/ng.313.

76. Yang TL, Chen XD, Guo Y, Lei SF, Wang JT, et al. Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis. *Am J Hum Genet* 2008; 83(6): 663–74. doi:10.1016/j.ajhg.2008.10.006.

77. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, et al. Diet and the evolution of human amylase gene copy number variation. *Nat Genet* 2007; 39(10): 1256–60. doi: 10.1038/ng2123.

78. Pelak K, Need AC, Fellay J, Shianna KV, Feng S, et al. Copy number variation of KIR genes influences HIV-1 control. *PLoS Biol* 2011; 9(11): e1001208. doi:10.1371/journal.pbio.1001208.

79. McConnell MJ, Lindberg MR, Brennand KJ, Piper JC, Voet T, et al. Mosaic Copy Number Variation in Human Neurons. *Science* 2013; 342(6158): 632–637. doi: 10.1126/science.1243472.

80. Albertson DG, Collins C, McCormick F, Gray JW. Chromosome aberrations in solid tumors. *Nat Genet* 2003; 34(4): 369–76. doi:10.1038/ng1215.

81. Shlien A, Tabori U, Marshall CR, Pienkowska M, Feuk L, et al. Excessive genomic DNA copy number variation in the Li-Fraumeni cancer predisposition syndrome. *Proc Natl Acad Sci U S A* 2008; 105(32): 11264–9. doi:10.1073/pnas.0802970105.

82. Shlien A, Malkin D. Copy number variations and cancer. *Genome Med* 2009; 1(6): 62. doi:10.1186/gm62.

83. Pylkäs K, Vuorela M, Otsukka M, Kallioniemi A, Jukkola-Vuorinen A, et al. Rare copy number variants observed in hereditary breast cancer cases disrupt genes in estrogen signaling and TP53 tumor suppression network. *PLoS Genet* 2012; 8(6): e1002734. doi: 10.1371/journal.pgen.1002734.

84. Taylor BS, Barretina J, Socci ND, Decarolis P, Ladanyi M, et al. Functional copy-number alterations in cancer. *PLoS One* 2008; 3(9): e3179. doi:10.1371/journal.pone.0003179.

85. Standfuß C, Pospisil H, Klein A. SNP Microarray analyses reveal copy number alterations and progressive genome reorganization during tumor development in SVT/t driven mice breast cancer. *BMC Cancer* 2012; 12(1): 380. doi:10.1186/1471-2407-12-380.

86. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, et al. Strong association of de novo copy number mutations with autism. *Science* 2007; 316(5823): 445–9. doi: 10.1126/science.1138659.

87. Kirov G, Rees E, Walters JTR, Escott-Price V, Georgieva L, et al. The Penetrance of Copy Number Variations for Schizophrenia and Developmental Delay. *Biol Psychiatry* 2013; 1–8. doi:10.1016/j.biopsych.2013.07.022.

88. Malhotra D, Sebat J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* 2012; 148(6): 1223–41. doi:10.1016/j.cell.2012.02.039.

89. Talseth-Palmer BA, Holliday EG, Evans TJ, McEvoy M, Attia J, et al. Continuing difficulties in interpreting CNV data: lessons from a genome-wide CNV association study of Australian HNPCC/lynch syndrome patients. *BMC Med Genomics* 2013; 6(1): 10. doi:10.1186/1755-8794-6-10.

90. Jiang W, Johnson C, Jayaraman J, Simecek N, Noble J, et al. Copy number variation leads to considerable diversity for B but not A haplotypes of the human KIR genes encoding NK cell receptors. *Genome Res* 2012; 22(10): 1845–54. doi:10.1101/gr.137976.112.

91. van Deursen JM. Rb loss causes cancer by driving mitosis mad. *Genome Biol Evol* 2007; 11(1): 1–3. doi:10.1016/j.ccr.2006.12.006.

92. Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, et al. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res* 2010; 20(5): 623–35. doi:10.1101/gr.102970.109.

93. Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, et al. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat Genet* 2009; 41(7): 849–53. doi:10.1038/ng.399.

94. Traherne JA, Martin M, Ward R, Ohashi M, Pellett F, et al. Mechanisms of copy number variation and hybrid gene formation in the KIR immune gene complex. *Hum Mol Genet* 2010; 19(5): 737–51. doi:10.1093/hmg/ddp538.

95. Carvalho CMB, Pehlivan D, Ramocki MB, Fang P, Alleva B, et al. Replicative mechanisms for CNV formation are error prone. *Nat Genet* 2013; 45(11): 1319–1326. doi:10.1038/ng.2768.

96. Arlt MF, Rajendran S, Birkeland SR, Wilson TE, Glover TW. De novo CNV formation in mouse embryonic stem cells occurs in the absence of Xrcc4-dependent nonhomologous end joining. *PLoS Genet* 2012; 8(9): e1002981. doi:10.1371/journal.pgen.1002981.

97. Walentinsson A, Sjöling A, Helou K, Klinga-Levan K, Levan G. Genomewide assessment of genetic alterations in DMBA-induced rat sarcomas: cytogenetic, CGH, and allelotype analyses reveal recurrent DNA copy number changes in rat chromosomes 1, 2, 4, and 7. *Genes Chromosome Canc* 2000; 28(2): 184–95. doi:10.1002/(SICI)1098-2264(200006)28:2⟨184::AID-GCC7⟩3.0.CO;2-V.

98. Arlt MF, Cagla A, Birkeland SR, Wilson TE, Glover TW. Hydroxyurea induces de novo copy number variants in human cells. *Proc Nat Acad Sci USA* 2011; 108(42): 17360–17365. doi:10.1073/pnas.1109272108.

99. Mkrtchyan H, Gross M, Hinreiner S, Polytiko A, Manvelyan M, et al. Early embryonic chromosome instability results in stable mosaic pattern in human tissues. *PLoS One* 2010; 5(3): e9591. doi:10.1371/journal.pone.0009591.

100. Liang Q, Conte N, Skarnes WC, Bradley A. Extensive genomic copy number variation in embryonic stem cells. *Proc Nat Acad Sci USA* 2008; 105(45): 17453–6. doi:10.1073/pnas.0805638105.

101. Tam PP, Behringer RR. Mouse gastrulation: the formation of a mammalian body plan. *Mech Develop* 1997; 68(1-2): 3–25. doi:10.1016/S0925-4773(97)00123-8.

102. Arnold SJ, Robertson EJ. Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. *Nat Rev Mol Cell Biol* 2009; 10(2): 91–103. doi:10.1038/nrm2618.

103. Ono T, Miyamura Y, Ikehata H, Yamanaka H, Kurishita A, et al. Spontaneous mutant frequency of lacZ gene in spleen of transgenic mouse increases with age. *Mutat Res* 1995; 338(1-6): 183–8.

104. Zhang XB, Urlando C, Tao KS, Heddle JA. Factors affecting somatic mutation frequencies in vivo. *Mutat Res* 1995; 338(1-6): 189–201.

105. Ono T, Ikehata H, Nakamura S, Saito Y, Hosoi Y, et al. Age-associated increase of spontaneous mutant frequency and molecular nature of mutation in newborn and old lacZ-transgenic mouse. *Mutat Res* 2000; 447(2): 165–77.

106. Hill KA, Buettner VL, Halangoda A, Kunishige M, Moore SR, et al. Spontaneous mutation in Big Blue mice from fetus to old age: tissue-specific time courses of mutation frequency but similar mutation types. *Environ Mol Mutagen* 2004; 43(2): 110–20. doi:10.1002/em.20004.

107. Dollé MET, Giese H, Hopkins CL, Martus HJ, Hausdorff JM, et al. Rapid accumulation of genome arrangements in liver but not in brain of old mice. *Nat Genet* 1997; 17: 431–434.

108. Feig DI, Reid TM, Loeb LA, Feig DI, Reid TM, et al. Reactive Oxygen Species in Tumorigenesis Reactive Oxygen Species in Tumorigenesis1. *Cancer Res* 1994; 54: 1890s–1894s.

109. Burhans WC, Weinberger M. DNA replication stress, genome instability and aging. *Nucleic Acids Res* 2007; 35(22): 7545–56. doi:10.1093/nar/gkm1059.

110. Cesta MF. Normal structure, function, and histology of the spleen. *Toxicol Pathol* 2006; 34(5): 455–65. doi:10.1080/01926230600867743.

111. Apel M, Berek C. Somatic mutations in antibodies expressed by germinal centre B cells early after primary immunization. *Int Immunol* 1990; 2(9): 813–9.

112. Tonegawa S. Somatic generation of antibody diversity. *Nature* 1983; 302(14): 575–581.

113. Cahan P, Li Y, Izumi M, Graubert TA. The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. *Nat Genet* 2009; 41(4): 430–7. doi:10.1038/ng.350.

114. Zagon IS, McLaughlin PJ, Smith S. Short communications. *Brain Res* 1977; 127: 279–282. doi:10.1242/jeb.089763.

115. Muotri AR, Chu VT, Marchetto MCN, Deng W, Moran JV, et al. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* 2005; 435(7044): 903–10. doi:10.1038/nature03663.

116. Rehen SK, Yung YC, McCreight MP, Kaushal D, Yang AH, et al. Constitutional aneuploidy in the normal human brain. *J Neurosci* 2005; 25(9): 2176–80. doi: 10.1523/JNEUROSCI.4560-04.2005.

117. Westra JW, Peterson SE, Yung YC, Mutoh T, Barral S, et al. Aneuploid mosaicism in the developing and adult cerebellar cortex. *J Comp Neurol* 2008; 507(6): 1944–51. doi: 10.1002/cne.21648.

118. Yurov YB, Iourov IY, Vorsanova SG, Liehr T, Kolotii AD, et al. Aneuploidy and confined chromosomal mosaicism in the developing human brain. *PLoS One* 2007; 2(6): e558. doi:10.1371/journal.pone.0000558.

119. Hofstetter JR, Zhang A, Mayeda AR, Guscar T, Nurnberger JI, et al. Genomic DNA from mice: a comparison of recovery methods and tissue sources. *Biochem Mol Med* 1997; 62(2): 197–202. doi:10.1006/bmme.1997.2637.

120. Duncan AW, Newell AEH, Bi W, Finegold MJ, Olson SB, et al. Aneuploidy as a mechanism for stress-induced liver adaptation. *J Clin Invest* 2012; 122(9): 3307–3315. doi: 10.1172/JCI64026DS1.

121. Klaunig JE, Goldblatt PJ, Hinton DE, Lipsky MM, Trump BF. Mouse liver cell culture I. Hepatocyte isolation . *In Vitro* 1981; 17(10): 913–925.

122. Clifford RJ, Zhang J, Meerzaman DM, Lyu MS, Hu Y, et al. Genetic variations at loci involved in the immune response are risk factors for hepatocellular carcinoma. *Hepatology* 2010; 52(6): 2034–43. doi:10.1002/hep.23943.

123. Royo F, Zabala A, Paz N, Acquadro F, Echevarria JJ, et al. Genome-Wide Analysis of DNA Copy Number Changes in Liver Steatosis. *Br J Med Med Res* 2013; 3(4): 1773–1785.

124. Orozco LD, Cokus SJ, Ghazalpour A, Ingram-Drake L, Wang S, et al. Copy number variation influences gene expression and metabolic traits in mice. *Hum Mol Genet* 2009; 18(21): 4118–29. doi:10.1093/hmg/ddp360.

125. Chaignat E, Yahya-Graison EA, Henrichsen CN, Chrast J, Schütz F, et al. Copy number variation modifies expression time courses. *Genome Res* 2011; 21(1): 106–13. doi: 10.1101/gr.112748.110.

126. Eichler EE. Copy number variation and human disease. *Nature Education* 2008; 1(3).

127. Fu W, Zhang F, Wang Y, Gu X, Jin L. Identification of copy number variation hotspots in human populations. *Am J Hum Genet* 2010; 87(4): 494–504. doi:10.1016/j.ajhg.2010. 09.006.

128. Itsara A, Cooper GM, Baker C, Girirajan S, Li J, et al. Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet* 2009; 84(2): 148–61. doi:10.1016/j.ajhg.2008.12.014.

129. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, et al. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 2010; 327(5967): 836–40. doi:10.1126/science.1183439.

130. Myers S, Freeman C, Auton A, Donnelly P, McVean G. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet* 2008; 40(9): 1124–9. doi:10.1038/ng.213.

131. Khelifi A, Meunier J, Duret L, Mouchiroud D. GC content evolution of the human and mouse genomes: insights from the study of processed pseudogenes in regions of different recombination rates. *J Mol Evol* 2006; 62(6): 745–52. doi:10.1007/s00239-005-0186-0.

132. Nguyen DQ, Webber C, Hehir-Kwa J, Pfundt R, Veltman J, et al. Reduced purifying selection prevails over positive selection in human copy number variant evolution. *Genome Res* 2008; 18: 1711–23. doi:10.1101/gr.077289.108.

133. Eichler EE. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet* 2001; 17(11): 661–9.

134. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, et al. Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* 2005; 77(1): 78–88. doi:10.1086/431652.

135. Egan CM, Sridhar S, Wigler M, Hall IM. Recurrent DNA copy number variation in the laboratory mouse. *Nat Genet* 2007; 39(11): 1384–9. doi:10.1038/ng.2007.19.

136. Akan P, Deloukas P. DNA sequence and structural properties as predictors of human and mouse promoters. *Gene* 2008; 410(1): 165–76. doi:10.1016/j.gene.2007.12.011.

137. Zhang D, Qian Y, Akula N, Alliey-Rodriguez N, Tang J, et al. Accuracy of CNV Detection from GWAS Data. *PLoS One* 2011; 6(1): e14511. doi:10.1371/journal.pone.0014511.

138. Rovelet-Lecrux A, Legallic S, Wallon D, Flaman JM, Martinaud O, et al. A genome-wide study reveals rare CNVs exclusive to extreme phenotypes of Alzheimer disease. *Eur J Hum Genet* 2012; 20(6): 613–7. doi:10.1038/ejhg.2011.225.

139. Lou H, Li S, Yang Y, Kang L, Zhang X, et al. A map of copy number variations in Chinese populations. *PLoS One* 2011; 6(11): e27341. doi:10.1371/journal.pone.0027341.

140. Zhang D, Cheng L, Qian Y, Alliey-Rodriguez N, Kelsoe JR, et al. Singleton deletions throughout the genome increase risk of bipolar disorder. *Mol Psychiatr* 2009; 14(4): 376–80. doi:10.1038/mp.2008.144.

141. Priebe L, Degenhardt FA, Herms S, Haenisch B, Mattheisen M, et al. Genome-wide survey implicates the influence of copy number variants (CNVs) in the development of early-onset bipolar disorder. *Mol Psychiatry* 2012; 17(4): 421–32. doi:10.1038/mp.2011.8.

142. Girirajan S, Brkanac Z, Coe BP, Baker C, Vives L, et al. Relative Burden of Large

CNVs on a Range of Neurodevelopmental Phenotypes. *PLoS Genetics* 2011; 7(11). doi: 10.1371/journal.pgen.1002334.

143. Sykulski M, Gambin T, Bartnik M, DerwiÅska K, Wiśniowiecka-Kowalnik B, et al. Multiple samples aCGH analysis for rare CNVs detection. *J Clin Bioinforma* 2013; 3(1): 12. doi:10.1186/2043-9113-3-12.

144. Buysse K, Delle Chiaie B, Van Coster R, Loeys B, De Paepe A, et al. Challenges for CNV interpretation in clinical molecular karyotyping: lessons learned from a 1001 sample experience. *Eur J Hum Genet* 2009; 52(6): 398–403. doi:10.1016/j.ejmg.2009.09.002.

145. Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, et al. Evolution and diversity of copy number variation in the great ape lineage. *Genome Res* 2013; 23: 1373–1382. doi:10.1101/gr.158543.113.

146. Beck JA, Lloyd S, Hafezparast M, Lennon-Pierce M, Eppig JT, et al. Genealogies of mouse inbred strains. *Nat Genet* 2000; 24(1): 23–5. doi:10.1038/71641.

147. Watkins-Chow DE, Pavan WJ. Genomic copy number and expression variation within the C57BL / 6J inbred mouse strain. *Genome Res* 2008; 18: 60–66. doi:10.1101/gr.6927808.medical.

148. Ionita-Laza I, Perry GH, Raby BA, Klanderman B, Lee C, et al. On the analysis of copy-number variations in genome-wide association studies: a translation of the family-based association test. *Genet Epidemiol* 2008; 32(3): 273–84. doi:10.1002/gepi.20302.

149. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* 2011; 21(6): 974–84. doi:10.1101/gr.114876.110.

150. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, et al. Low concordance of multiple variant-

calling pipelines: practical implications for exome and genome sequencing. *Genome Med* 2013; 5(3): 28. doi:10.1186/gm432.

151. Xu B, Roos JL, Levy S, van Rensburg EJ, Gogos Ja, et al. Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet* 2008; 40(7): 880–5. doi:10.1038/ng.162.

152. Hedrich H (Ed.). The Laboratory Mouse (Elsevier Academic Press).

153. Wade CM, Daly MJ. Genetic variation in laboratory mice. *Nat Genet* 2005; 37(11): 1175–80. doi:10.1038/ng1666.

154. Chesler EJ, Miller DR, Galloway LD. The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics. *Mamm Genome* 2008; 19(6): 382–389. doi:10.1007/s00335-008-9135-8.

155. Aylor DL, Valdar W, Foulds-Mathes W, Buus RJ, Verdugo RA, et al. Genetic analysis of complex traits in the emerging Collaborative Cross. *Genome Res* 2011; 21(8): 1213–22. doi:10.1101/gr.111310.110.

156. Churchill G, Gatti DM, Munger SC, Svenson KL. The diversity outbred mouse population. *Mamm Genome* 2012; 23(9-10): 713–8. doi:10.1007/s00335-012-9414-2.

157. Xi R, Hadjipanayis AG, Luquette LJ, Kim TM, Lee E, et al. Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc Nat Acad Sci USA* 2011; 108(46): E1128–36. doi:10.1073/pnas.1110574108.

158. Krumm N, Sudmant PH, Ko A, Roak BJO, Malig M, et al. Copy number variation detection and genotyping from exome sequence data. *Genome Res* 2012; 22: 1525–1532. doi:10.1101/gr.138115.112.Mapping.

159. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, et al. High resolution analysis of DNA

copy number variation using comparative genomic hybridization to microarrays. *Nature Genet* 1998; 20(2): 207–11. doi:10.1038/2524.

160. Huang J, Wei W, Zhang J, Lui G, Bignell GR, et al. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum Genet* 2004; 1(4): 287–299. doi:10.1186/1479-7364-1-4-287.

161. Bengtsson H, Irizarry R, Carvalho B, Speed TP. Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* 2008; 24(6): 759–67. doi:10.1093/bioinformatics/btn016.

162. Komura D, Shen F, Ishikawa S, Fitch KR, Chen W, et al. Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Res* 2006; 16(12): 1575–84. doi:10.1101/gr.5629106.

163. Peiffer Da, Le JM, Steemers FJ, Chang W, Jenniges T, et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res* 2006; 16(9): 1136–48. doi:10.1101/gr.5402306.

164. Wang DG. Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science* 1998; 280(5366): 1077–1082. doi:10.1126/science.280.5366.1077.

165. Winchester L, Yau C, Ragoussis J. Comparing CNV detection methods for SNP arrays. *Brief Funct Genomic Proteomic* 2009; 8(5): 353–66. doi:10.1093/bfgp/elp017.

166. Curtis C, Lynch AG, Dunning MJ, Spiteri I, Marioni JC, et al. The pitfalls of platform comparison: DNA copy number array technologies assessed. *BMC Genomics* 2009; 10: 588. doi:10.1186/1471-2164-10-588.

167. Shifman S, Bell JT, Copley RR, Taylor MS, Williams RW, et al. A high-resolution single

nucleotide polymorphism genetic map of the mouse genome. *PLoS Biol* 2006; 4(12): e395. doi:10.1371/journal.pbio.0040395.

168. Yang H, Ding Y, Hutchins LN, Szatkiewicz J, Bell TA, et al. A customized and versatile high-density genotyping array for the mouse. *Nat Methods* 2009; 6(9): 663–6. doi: 10.1038/nmeth.1359.

169. Didion JP, Yang H, Sheppard K, Fu CP, McMillan L, et al. Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. *BMC Genomics* 2012; 13(1): 34. doi:10.1186/1471-2164-13-34.

170. Yang H, Wang JR, Didion JP, Buus RJ, Bell Ta, et al. Subspecific origin and haplotype diversity in the laboratory mouse. *Nat Genet* 2011; 43(7): 648–55. doi:10.1038/ng.847.

171. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* 2007; 39: 1–11. doi:10.1038/ng2028.Methods.

172. Affymetrix, Inc. website, 2013. [Online; accessed 12-April-2013].

173. Giorgi FM, Bolger AM, Lohse M, Usadel B. Algorithm-driven artifacts in median Polish summarization of microarray data. *BMC Bioinformatics* 2010; 11(1): 553. doi:10.1186/1471-2105-11-553.

174. Dai M, Wang P, Boyd AD, Kostov G, Athey B, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* 2005; 33(20): e175. doi:10.1093/nar/gni179.

175. Dalma-Weiszhausz DD, Warrington J, Tanimoto EY, Miyada CG. The Affymetrix GeneChip platform: an overview. *Meth Enzymol* 2006; 410(6): 3–28. doi:10.1016/S0076-6879(06)10001-4.

176. Arteaga-Salas JM, Zuzan H, Langdon WB, Upton GJG, Harrison AP. An overview of

image-processing methods for Affymetrix GeneChips. *Brief Bioinform* 2008; 9(1): 25–33. doi:10.1093/bib/bbm055.

177. Bolstad BM, Irizarry Ra, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)* 2003; 19(2): 185–93.

178. Hong H, Su Z, Ge W, Shi L, Perkins R, et al. Assessing batch effects of genotype calling algorithm BRLMM for the Affymetrix GeneChip Human Mapping 500 K array set using 270 HapMap samples. *BMC Bioinformatics* 2008; 9 Suppl 9: S17. doi: 10.1186/1471-2105-9-S9-S17.

179. Marioni JC, Thorne NP, Valsesia A, Fitzgerald T, Redon R, et al. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol* 2007; 8(10): R228. doi:10.1186/gb-2007-8-10-r228.

180. Diskin SJ, Li M, Hou C, Yang S, Glessner J, et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic acids research* 2008; 36(19): e126. doi:10.1093/nar/gkn556.

181. BRLMM-P: A Genotype Calling Method for the SNP 5.0. Technical Report, Affymetrix, 2007.

182. Lovmar L, Ahlford A, Jonsson M, Syvänen AC. Silhouette scores for assessment of SNP genotype clusters. *BMC Genomics* 2005; 6: 35. doi:10.1186/1471-2164-6-35.

183. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, et al. Common deletion polymorphisms in the human genome. *Nat Genet* 2005; 38(1): 86–92. doi:10.1038/ng1696.

184. Amos CI, Shete S, Chen J, Yu RK. Positional Identification of Microdeletions with Genetic Markers. *Hum Hered* 2003; 56: 107–118. doi:10.1159/000073738.

185. Alonso A, Julià A, Tortosa R, Canaleta C, Cañete JD, et al. CNstream: a method for the identification and genotyping of copy number polymorphisms using Illumina microarrays. *BMC Bioinformatics* 2010; 11: 264. doi:10.1186/1471-2105-11-264.

186. Rasmussen M, Sundström M, Göransson Kultima H, Botling J, Micke P, et al. Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity. *Genome Biol* 2011; 12(10): R108. doi:10.1186/gb-2011-12-10-r108.

187. Grayson BL, Aune TM. A comparison of genomic copy number calls by Partek Genomics Suite, Genotyping Console and Birdsuite algorithms to quantitative PCR. *BioData Min* 2011; 4: 8. doi:10.1186/1756-0381-4-8.

188. Eckel-Passow JE, Atkinson EJ, Maharjan S, Kardia SLR, de Andrade M. Software comparison for evaluating genomic copy number variation for Affymetrix 6.0 SNP array platform. *BMC Bioinformatics* 2011; 12(1): 220. doi:10.1186/1471-2105-12-220.

189. Baross A, Delaney AD, Li HI, Nayar T, Flibotte S, et al. Assessment of algorithms for high throughput detection of genomic copy number variation in oligonucleotide microarray data. *BMC Bioinformatics* 2007; 8: 368. doi:10.1186/1471-2105-8-368.

190. Vogler C, Gschwind L, Röthlisberger B, Huber A, Filges I, et al. Microarray-based maps of copy-number variant regions in European and sub-Saharan populations. *PloS ONE* 2010; 5(12): e15246. doi:10.1371/journal.pone.0015246.

191. Kim SY, Kim JH, Chung YJ. Effect of combining multiple CNV defining algorithms on the reliability of CNV calls from SNP genotyping data. *Genomics Inform* 2012; 10(3): 194–199. doi:10.5808/GI.2012.10.3.194.

192. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 2004; 5(4): 557–72. doi:10.1093/biostatistics/kxh008.

193. Hawthorn L, Luce J, Stein L, Rothschild J. Integration of transcript expression, copy number and LOH analysis of infiltrating ductal carcinoma of the breast. *BMC Cancer* 2010; 10: 460. doi:10.1186/1471-2407-10-460.

194. Wang K, Li M, Hadley D, Liu R, Glessner J, et al. PennCNV : An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 2007; 17: 1665–1674. doi: 10.1101/gr.6861907.1.

195. Tse KP, Su WH, Yang Ml, Cheng HY, Tsang NM, et al. A gender-specific association of CNV at 6p21.3 with NPC susceptibility. *Hum Mol Genet* 2011; 20(14): 2889–2896. doi:10.1093/hmg/ddr191.

196. Butler J, Locke MEO, Hill KA, Daley M. HD-CNV: Hotspot Detector for Copy Number Variants. *Bioinformatics* 2013; 29(2): 262–263. doi:10.1093/bioinformatics/bts650.

197. Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, et al. DbVar and DGVa: public archives for genomic structural variation. *Nucl Acids Res* 2013; 41(Database issue): D936–41. doi:10.1093/nar/gks1213.

198. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. The Human Genome Browser at UCSC. *Genome Res* 2002; 12(6): 996–1006. doi:10.1101/gr.229102.

199. Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, et al. The UCSC Genome Browser database: extensions and updates 2013. *Nucl Acids Res* 2013; 41(Database issue): D64–9. doi:10.1093/nar/gks1048.

200. Veytsman B, Akhmadeeva L. Drawing Medical Pedigree Trees with TEX and PSTricks. Technical Report 4, 2006.

201. The Center for Genome Dynamics at The Jackson Laboratory, 2013. [Online; accessed 12-April-2013].

202. Eitutis ST. 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. Master's thesis, The University of Western Ontario, 2013. doi:10.1016/S0925-8388(12)01848-8.

203. UCSC Genome Bioinformatics Sequence and Annotation Downloads, 2013. [Online; accessed 9-January-2013].

204. Affymetrix Genotyping Console 2.1 User Manual. Technical Report, 2009.

205. Affymetrix. Analysis Guide Axiom™ Genotyping Solution Data Analysis Guide. Technical Report, 2011.

206. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, et al. Circos: an information aesthetic for comparative genomics. *Genome Res* 2009; 19(9): 1639–45. doi:10.1101/gr.092759.109.

207. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003; 4(2): 249–64. doi:10.1093/biostatistics/4.2.249.

208. Jasmine F, Ahsan H, Andrulis IL, John EM, Chang J, et al. Whole genome amplification enables accurate genotyping for micrarray-based high density SNP array. *Cancer Epidemiol Biomarkers Prev* 2008; 17(12): 3499–3508. doi:10.1158/1055-9965.EPI-08-0482.

209. Li A, Omura N, Hong SM, Goggins M. Pancreatic cancer DNMT1 expression and sensitivity to DNMT1 inhibitors. *Cancer Biol Ther* 2010; 9(4): 321–329. doi:10.4161/cbt.9.4.10750.

210. Walter MJ, Payton JE, Ries RE, Shannon WD, Deshmukh H, et al. Acquired copy number

alterations in adult acute myeloid leukemia genomes. *Proc Natl Acad Sci U S A* 2009; 106(31): 12950–5. doi:10.1073/pnas.0903091106.

211. PennCNV: copy number variation detection, 2011. [Online; accessed 12-April-2013].

212. Valsesia A, Stevenson BJ, Waterworth D, Mooser V, Vollenweider P, et al. Identification and validation of copy number variants using SNP genotyping arrays from a large clinical cohort. *BMC Genomics* 2012; 13(1): 241. doi:10.1186/1471-2164-13-241.

213. Lin P, Hartz SM, Wang JC, Krueger RF, Foroud TM, et al. Copy number variation accuracy in genome-wide association studies. *Hum Hered* 2011; 71(3): 141–7. doi: 10.1159/000324683.

214. Pankratz N, Dumitriu A, Hetrick KN, Sun M, Latourelle JC, et al. Copy number variation in familial Parkinson disease. *PLoS One* 2011; 6(8): e20988. doi:10.1371/journal.pone. 0020988.

215. Collins AL, Kim Y, Szatkiewicz JP, Bloom RJ, Hilliard CE, et al. Identifying bipolar disorder susceptibility loci in a densely affected pedigree. *Mol Psychiatry* 2012; 1–2. doi:10.1038/mp.2012.176.

216. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010; 26(6): 841–2. doi:10.1093/bioinformatics/btq033.

217. Gascuel O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* 1997; 14(7): 685–95.

218. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 2004; 20(2): 289–290. doi:10.1093/bioinformatics/btg412.

219. Cutler G, Marshall LA, Chin N, Baribault H, Kassner PD. Significant gene content variation characterizes the genomes of inbred mouse strains. *Genome Res* 2007; 17(12): 1743–54. doi:10.1101/gr.6754607.

220. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2013.

221. van Heesch S, Mokry M, Boskova V, Junker W, Mehon R, et al. Systematic biases in DNA copy number originate from isolation procedures. *Genome Biol* 2013; 14(4): R33. doi:10.1186/gb-2013-14-4-r33.

222. Xu H, Poh WT, Sim X, Ong RTH, Suo C, et al. SgD-CNV, a database for common and rare copy number variants in three Asian populations. *Hum Mutat* 2011; 32(12): 1341–9. doi:10.1002/humu.21601.

223. Kresse, Stine H and Skårn, Magne and Ohnstad, Hege O and Namløs, Heidi M and Bjerkehagen, Bodil and Myklebost, Ola and Meza-Zepeda, Leonardo A. DNA copy number changes in high-grade malignant peripheral nerve sheath tumors by array CGH. *Mol Cancer* 2008; 7: 48. doi:10.1186/1476-4598-7-48.

224. Akiri G, Sabo E, Dafni H, Akiri G, Sabo E, et al. Lysyl oxidase-related protein-1 promotes tumor fibrosis and tumor progression in vivo. *Cancer Res* 2003; 63: 1657–1666.

225. Fong SFT, Dietzsch E, Fong KSK, Hollosi P, Asuncion L, et al. Lysyl oxidase-like 2 expression is increased in colon and esophageal tumors and associated with less differentiated colon tumors. *Genes Chromosome Canc* 2007; 655: 644–655. doi:10.1002/gcc.

226. Kopelman NM, Stone L, Gascuel O, Rosenberg NA. The behavior of admixed populations in neighbor-joining inference of population trees. *Pac Symp Biocomput* 2013; 273–284. doi:10.1142/9789814447973_0027.

227. Marioni JC, White M, Tavaré S, Lynch AG. Hidden copy number variation in the HapMap population. *Proc Nat Acad Sci USA* 2008; 105(29): 10067–72. doi:10.1073/pnas.0711252105.

228. Skopek T, Kort K, Marino D. Relative sensitivity of the endogenous hprt gene and lacI transgene in ENU-treated Big Blue B6C3F1 mice. *Environ Mol Mutagen* 1995; 26: 9–15.

229. Sommer SS, Ketterling RP. How precisely can data from transgenic mouse mutation-detection systems be extrapolated to humans?: lessons from the human factor IX gene. *Mutat Res* 1994; 307: 517–531. doi:10.1016/0027-5107(94)90263-1.

230. Prtenjaca A, Hill KA. Mutation frequency is not elevated in the cerebellum of *harlequin*/Big Blue® mice but Class II deletions occur preferentially in young *harlequin* cerebellum. *Mutat Res* 2011; 707(1-2): 53–60.

231. Price TS, Regan R, Mott R, Hedman A, Honey B, et al. SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. *Nucleic Acids Res* 2005; 33(11): 3455–64. doi:10.1093/nar/gki643.

232. Cahan P, Godfrey LE, Eis PS, Richmond TA, Selzer RR, et al. wuHMM: a robust algorithm to detect DNA copy number variation using long oligonucleotide microarray data. *Nucleic Acids Res* 2008; 36(7): e41. doi:10.1093/nar/gkn110.

233. Hupé P, Stransky N, Thiery JP, Radvanyi F, Barillot E. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* 2004; 20(18): 3413–22. doi:10.1093/bioinformatics/bth418.

Appendices

Appendix A

Animal Use Protocol Approval

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

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

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

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

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

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

Submitted by: [redacted]
on behalf of the Animal Use Subcommittee

[redacted]

The University *of* Western Ontario
Animal Use Subcommittee / University Council on Animal Care

[redacted]

Appendix B

Supplementary Methods

B.1 Filtering the list of SNP probesets for optimal genotyping

Of the 623,324 SNP loci assayed by the MDGA,[168] a filtered list containing 526,162 SNPs was used for the first genotyping run using Affymetrix® Genotyping Console™ as outlined by Eitutis (2013).[202] After this first genotyping round, CEL files that did not reach a SNP minimum calling rate of 97% were removed from the analysis. After CEL file removal, a second round of genotyping was performed on the passing CEL files using the subset of 526,162 SNPs. After the second genotyping round, all SNPs that did not meet the minimum criterion of 97% genotyping call rate were removed. This removal step resulted in 470,339 remaining SNPs that were genotyped with an overall call rate of 97%.

B.2 Animal care and housing

All protocols were approved by the University of Western Ontario Animal Use Subcommittee (Appendix A). The male founder mouse in this study (300.6) was an inbred C57Bl/6J Big Blue® mouse homozygous for the Big Blue® Lambda LIZ (LacI/Z) bacteriophage shuttle vector (Taconic Farms, Germantown, NY; Table B.1). The Big Blue® transgenic mutation

Table B.1: Quality control measures for high molecular weight genomic DNA extractions. Source adapted from Eitutis (2013).[202]

Sample ID	Big Blue® Genotype[a]	*Aif* Genotype[b]	Tissue Type	Tissue Mass (mg)	DNA Extraction Protocol[c]	Sample Processing Location[d]	[DNA] (μg/μL)	260/280 Ratio	260/230 Ratio
300.6SP	+/+	XY	SP	9.1	Gentra	LRGC	0.26	1.9	2.44
300.6CL	+/+	XY	CL	42	Gentra	LRGC	0.1	1.81	1.43
900.3SP	-/-	$X^{hq}X$	SP	9.4	Gentra	LRGC	0.48	1.87	2.42
900.3LI	-/-	$X^{hq}X$	LI	10	Gentra	LRGC	0.22	1.82	1.44
904.11SP	+/-	$X^{hq}Y$	SP	30	Wizard	JAX	1.69	1.82	2.23
904.11CL	+/-	$X^{hq}Y$	CL	40	Wizard	JAX	0.37	1.76	1.34
904.9SP	+/-	$X^{hq}Y$	SP	50	Wizard	JAX	1.35	1.85	2.16
904.9CL	+/-	$X^{hq}Y$	CL	60	Wizard	JAX	0.27	1.74	1.31
911.50SP	ND	XY	SP	30	Wizard	JAX	1.53	1.83	2.28
911.50CL	ND	XY	CL	70	Wizard	JAX	0.3	1.76	1.33
911.50SPrpt[e]	ND	XY	SP	8.6	Wizard	JAX	0.25	1.88	2.33
911.50CLrpt	ND	XY	CL	11	Wizard	JAX	0.1	1.88	2.34
911.49SP	ND	XY	SP	30	Wizard	JAX	1.84	1.84	2.23
911.49CL	ND	XY	CL	60	Wizard	JAX	0.28	1.73	1.27
911.143CL	-/-	XY	CL	40	Wizard	JAX	0.19	1.73	1.16
911.148CL	-/-	XY	CL	40	Wizard	JAX	0.19	1.72	1.13

[a] Presence or absence of the Big Blue® construct
[b] Presence or absence of the sex-linked *harlequin* (*hq*) mutation in *Aif*
[c] Gentra = Gentra® Puregene® Kit, Wizard = Wizard® Genomic DNA Purification Kit
[d] JAX = The Jackson Laboratory (Bar Harbor, Maine), LRGC = London Regional Genomics Centre (London, Ontario)
[e] rpt = replicate (for tissues sampled from mouse 911.50)

detection construct was used in this mouse strain in separate studies of spontaneous mutations and baseline endogenous mutation frequencies and patterns in individual tissues were observed.[106,228,229] The female founder mouse in this study (900.3) is a carrier of a retroviral insertion downregulating the *Apoptosis-inducing factor* (*Aif*) gene referred to as the *harlequin* (*hq*) mutation (*B6CBA Ca A$^{w\text{-}J}$/APdcd8^{Hq}/J*; The Jackson Laboratory, Bar Harbor, ME). Analysis of spontaneous mutations in a gene target using individual tissues of mice with deficiencies in AIF revealed near baseline endogenous mutation frequencies and patterns.[230] The procedures for genotyping mice for the presence of the Big Blue® construct and the *Aif* mutation have been published previously.[230]

Mice were housed in a Canadian Council on Animal Care approved facility with a 14/10 hour light/dark cycle at a relative humidity of 44 to 66% with a constant temperature of 21 ± 1°C. Mice were fed a standard diet (PMI Foods, St. Louis, MO) and given water *ad libitum*. Mice were not intentionally exposed to any known mutagen, so any observed departures from the reference sets were inferred to be either tissue-specific or spontaneous biological phenomena.

B.3 Tissue harvest and storage

Mice were euthanized by carbon dioxide (CO_2) inhalation at the following middle-adulthood ages: 300.6 at 10.4 months; 900.3 at 11.3 months; 904.9 and 904.11 at 7.7 months; 911.50 and 911.49 at 8.7 months; and 911.143 and 911.148 at 15.2 months (Table B.1). The spleen, cerebellum, and liver were harvested and flash frozen in liquid nitrogen prior to long-term storage at -80°C.

B.4 DNA extraction and preparation

Tissues were removed from freezer storage and a portion of each tissue was taken for DNA extraction, with the remaining portion returned to the freezer for long-term storage (Table B.1). DNA was isolated, prepared, and hybridized to arrays in two separate batches: the first being the initial 10-sample batch in which the tissues were sent to The Jackson Laboratory (Bar Harbor, Maine), and the second batch being additional samples (including one repeat each of mouse 911.50's spleen and cerebellum) sent to The London Regional Genomics Centre (LRGC; London, Ontario). Table B.1 indicates which samples were prepared, hybridized, and scanned at The Jackson Laboratory and which samples were prepared, hybridized, and scanned at The London Regional Genomics Centre (henceforth referred to as "JAX samples" and "LRGC samples", respectively).

The Jackson Laboratory received spleen and cerebellum samples, from which their technicians extracted high molecular weight DNA using the phenol-chloroform extraction protocol specified by the Affymetrix® Genome-Wide Human SNP Nsp/Sty 6.0 Assay Kit 5.0/6.0 (Affymetrix® Genome-Wide Human SNP Nsp/Sty 6.0 user guide, 2008). The London Regional Genomics Centre received spleen, liver, and cerebellum samples, from which their technicians prepared high molecular weight DNA using the Gentra® Puregene® (Qiagen, Mississauga, ON) extraction protocol. Both centres prepared high molecular weight DNA for hybridization to the Mouse Diversity Genotyping Array and carried out hybridization steps in accordance with the Affymetrix® Genome-Wide Human SNP Nsp/Sty 6.0 user guide, with one DNA sample per array. Fluorescence intensities for each array were measured with Affymetrix® GeneChip scanner as relative fluorescence units (RFUs) on a scale from 21 to 216 (private correspondence with D. Carter, London Regional Genomics Centre). Fluorescence data for each array was quantified in CEL files (one CEL file per array) to be used for bioinformatic analysis.

Appendix C

Supplementary Tables

This appendix contains supplementary tables.

Table C.1: Details of previous CNV discovery studies used to construct the custom database. C57BL/6J strain appears as red text.

Study	Platform[a]	CNV Calling Software	Mouse Tissues	Mouse Strains	CNV or CNVR State(s)
Agam *et al* 2010[50]	aCGH (NimbleGen Mouse 2.1M array)	SW-ARRAY[231]	unspecified[b]	A/J, AKR/J, BALB/cJ, C3H/HeJ, CBA/J, DBA/2J and LP/J	CNV gains, CNV losses, CNVRs
Cahan *et al* 2009[113]	aCGH (custom tiling array)	wuHMM[232]	spleen, liver, kidney, tail	SWR/J, SM/J, SJL/J, PL/J, NZB/BlNJ, NOD/Ltj, LG/J, KK/HlJ, FVB/NJ, DBA/2J, C58/J, C57L/J, C3H/HeJ, BTBR T+ tf/J, BALB/cByJ, AKR/J, A/J, 129X1/SvJ, 129S1/SvImJ	unspecified state CNVs

Continues on next page

Table C.1 – continued from previous page

Study	Platform[a]	CNV Calling Software	Mouse Tissues	Mouse Strains	CNV or CNVR State(s)
Cutler *et al* 2007[219]	aCGH (Agilent 244K Mouse Genome Array)	GLAD[233]	tail or unspecified[c]	129S1/SvImJ, 129X1/SvJ, A/J, AKR/J, BALB/cJ, BTBR T+ tf/J, BUB/BnJ, C3H/HeJ, C57BL/10J, C57BLKS/J, C57BR/cdJ, C57L/J, C58/J, CAST/EiJ, CBA/J, CE/J, CZECHII/EiJ, DBA/1J, DBA/2J, FVB/Ntac, I/LnJ, JF1/Ms, KK/HIJ, LP/J, MA/MyJ, MOLF/EiJ, MSM/Ms, NOD/LtJ, NON/LtJ, NZB/BINJ, NZW/LacJ, PERA/EiJ, PL/J, PWK/PhJ, RIIIS/J, SEA/GnJ, SJL/J, SM/J, SPRET/EiJ, SWR/J, WSB/EiJ	unspecified state CNVs
Egan *et al* 2007[135]	ROMA (custom array)	custom HMM methods	liver, tail	C57BL/6J substrains	CNV gains, CNV losses
Henrichsen *et al* 2009[63]	aCGH	custom time-dependent HMM[63]	liver	C57BL/6J, 129S2, A/J, AKR/J, BALB/cByJ, C3HeJ, C3HeB/FeJ, CAST/Ei, DBA/2J, LP/J, PL/J, Spret/Ei, SJL/J	unspecified state CNVs

Continues on next page

Table C.1 – continued from previous page

Study	Platform[a]	CNV Calling Software	Mouse Tissues	Mouse Strains	CNV or CNVR State(s)
Quinlan *et al* 2010[92]	paired-end sequencing, whole-genome shotgun sequencing	HYDRA[92]	unspecified[d]	DBA/2	unspecified state CNVs

[a] aCGH = array comparative genomic hybridization; ROMA = representative oligonucleotide array
[b] authors acquired DNA from the Jackson Laboratory
[c] authors acquired additional DNA from the Jackson Laboratory
[d] authors acquired sequence reads from the NCBI Trace Archive (http://www.ncbi.nlm.nih.gov/Traces)

Appendix D

Calculations

D.1 Marker density cutoff calculations

TABLE D.1: Marker density calculations.

Number of features represented	**Affymetrix® Human 6.0**	**MDGA (Filtered SNPs only)**
# SNP probes + # CN probes	906,600 + 946,000	470,339 + 0
Total # of array probes	1,852,600	470,339

	Affymetrix® Human 6.0	**MDGA (Filtered SNPs only)**
Total # of array probes	1,852,600	470,339
	÷	÷
Golden Path Length (bp)	3,093,120,360	2,716,965,481
Features per genomic bp	6.0^{-4}	1.7^{-4}

	Affymetrix® Human 6.0
Features per genomic base pair	6.0^{-4}
	÷
Marker density cutoff[60]	0.0004
Ratio	1.5

	MDGA (Filtered SNPs only)
Probes per genomic base pair	1.7^{-4}
	÷
Ratio	1.5
Marker density cutoff	0.00013

Curriculum Vitae

Name:	Andrea E. Wishart	
Education:	**Bachelor of Science,** The University of Western Ontario Honors Specialization in Biology Minor in Art History and Criticism	2010
Honours and Awards:	**Student and New Investigator Travel Award** Environmental Mutagen and Genomics Society	2013
	Department of Biology Graduate Student Teaching Award The University of Western Ontario	2012
	Graduate Thesis Research Award The University of Western Ontario	2012
	Teaching Support Centre Great Ideas in Teaching Award The University of Western Ontario	2011
Related Work Experience:	**Teaching Assistant,** The University of Western Ontario Biology 2290G Scientific Methods in Biology	2014

Biology 3594a Genome Organization, DNA Repair, and Mutagenesis	2010-2013
Biology 4243G Political Biology	2011-2013
Summer Research Assistant, The University of Western Ontario	2010

Publications:

2. Castellani CA, Melka MG, **Wishart AE**, Locke MEO, Awamleh Z, O'Reilly RL, Singh S. (2014) Biological relevance of CNV calling methods using familial relatedness including monozygotic twins. *BMC Bioinformatics 15:114.*

1. Luloff TW, **Wishart AE**, Addison SMF, MacDougall-Shackleton SA, Hill KA. (2011) Radiation exposure differentially affects songbird 8-hydroxy-2'-deoxyguanosine plasma profiles: ionizing radiation damage response in songbirds DNA damage response in songbirds. *Environmental Molecular Mutagenesis 52(8):658-63.*

Select Presentations:

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

5. **Wishart AE**, Locke MEO, Eitutis ST, Hill KA. Patterns of recurrent and tissue-specific copy number variants in the mouse genome. *Mouse Molecular Genetics Meeting, Pacific Grove, CA*. October 2012. Poster presentation.

4. **Wishart AE**, Eitutis ST, Butler J, Locke MEO, Daley M, Hill KA. Patterns of copy number changes between spleen and cerebellum differ in the *harlequin* mouse model of mitochondrial dysfunction. *Environmental Mutagen Society Annual Meeting, Bellevue, WA.* September 2012. Poster presentation.

3. **Wishart AE**, Eitutis ST, Hill KA. Copy number changes across the mouse genome discovered using the Mouse Diversity Genotyping Array show tissue and genotype specificity. *Environmental Mutagen Society Annual Meeting, Montreal, QC.* October 2011. Poster presentation.

2. **Wishart AE**, Eitutis ST, Hill KA. Evidence for an altered profile of copy number changes in the cerebellum of the *harlequin* mouse mimic of human aging-associated neurodegeneration. *International Congress of Human Genetics, Montreal, QC.* October 2011. Poster presentation.

1. **Wishart AE**, Bernard L, Prtenjaca A, Hill KA. Gene-environment interactions for nocturnal home cage behaviour in a mouse model of aging and neurodegeneration. *Genetics Society of Canada Annual Meeting, Hamilton ON.* June 2010. Poster presentation **[Best Poster Presentation Award]**