

PHENOTYPIC IMPORTANCE AND POPULATION DYNAMICS OF GENOMIC
STRUCTURAL VARIATION

A Dissertation

Presented to the Faculty of the Graduate School
of Cornell University

In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by

Jeremiah Daniel Degenhardt

May 2010

© 2010 Jeremiah Daniel Degenhardt

PHENOTYPIC IMPORTANCE AND POPULATION DYNAMICS OF GENOMIC
STRUCTURAL VARIATION

Jeremiah Daniel Degenhardt, Ph. D.

Cornell University 2010

The fundamental goals of molecular genetic studies are 1) understanding how variation within an organism's genome translates into phenotypic differences and 2) how this variation segregates and becomes fixed within populations. An enormous amount of work has been invested in looking at point mutations, or single nucleotide polymorphisms (SNPs). However, far less work has been devoted to studying larger structural variation (SV). Advances in molecular techniques (i.e. PCR, Sanger sequencing, genotyping arrays and comparative hybridization arrays) have greatly expanded our ability to detect these larger mutations. Here, I focus on developing methods for the analysis of data generated using these techniques and apply these methods to several data sets to broaden our understanding of the role of SV in genome evolution as well as the phenotypic consequences of this variation. Specifically, I first develop a method for the detection of a particular type of SV, known as copy number variation (CNV). This method is applied to genotyping array data collected from domestic dogs. I analyze these data to detect the extent of CNV in domestic dogs and their close relatives and to better understand population genetics and evolution of CNV. The analysis reveals nearly 10,000 CNVs segregating in domestic dogs and covering nearly 400 Mb of the dog genome. Next, using a retrospective study design, I investigate the role of *CCL3L* CNV

in the progression rate to simian AIDS in rhesus macaque. I find strong evidence that reduced copy number of *CCL3L* increases progression rates in rhesus macaques. This is a similar finding to that seen in humans in an earlier study. Therefore, rhesus macaque is a promising model organism for understanding how *CCL3L* CNV is affecting HIV progression in humans. Characterizing the role of *CCL3L* CNV will allow researchers to increase power in vaccine trials by controlling for this natural variation. Finally, I introduce a novel method for mapping the pseudoautosomal region in mammalian genomes. I apply this method to data collected from domestic and wild canids as well as rhesus macaque and use the results of this study to further the understanding of PAR evolution across the mammalian tree.

BIOGRAPHICAL SKETCH

Jeremiah was born to Robert and Muree Degenhardt in Spokane Washington in February 1979. His interest in nature and biology started early. Behind the family home was a large field where he spent many hours collecting insects and trying to identify them with his *Golden Guide to Insects*. From the 2nd grade through high school, Jeremiah was home-schooled by his mother. They took frequent trips to the library, where nature and biology books drew him in. By the time he reached nine years of age, he had read every book in the local library relating to big cats. Two of these books in particular impacted the decisions he would make later in life. These books were *The Ghost Walker* by R.D. Lawrence and *Cougar: Ghost of the Rockies* by Karen McCall, the second of which documented the work of Maurice Hornocker. Jeremiah began volunteering at a zoological rescue center for big cats in Mead, Washington when he was 14. Within two years he was employed full time as zookeeper, and entrusted with the care of 27 big cats, including four cougars. While working as a zookeeper, he gave public education tours and outreach presentations. This work, along with the books he had read as a child, led to his interest in pursuing a career where he could support the conservation of big cats. However, because he was home schooled, he had limited experience in a classroom setting; he was nervous about college.

Despite this nervousness, in September of 1999 he began taking courses at Spokane Falls Community College. He was surprised and relieved to find that he excelled in biology and science courses. While at Community College, he met his future wife, Angie, and her two lovely daughters, Kelsey and Gwyn who were then five and three.

In the winter of 2001 he decided to transfer to the University of Idaho

where he could pursue Angie and a bachelors degree in Zoology. His goal was to join the group of Maurice Hornocker, to study conservation biology of cougars. Maurice had shifted his focus to Siberian tigers, and his field schedule kept him away from Idaho most of the time, so Jeremiah began inquiring around for lab to join. By a fortuitous coincidence involving *Plethodon idahoensis*, and Bryan Carstens, he and Angie both found a home for their undergraduate research in the laboratory of Jack Sullivan. While learning and working in the Sullivan lab, Jeremiah studied phylogenetics and phylogeography. It was here that he began to delve into molecular genetics. This research led to his change of focus from conservation biology to molecular genetics and evolution.

He completed his Bachelors of Science in May 2005, the same month he was married, and was offered a position in the laboratory of Carlos Bustamante. So, in July of 2005 he and his new family (which by this time included 8 snakes and 2 dogs) packed a moving van, and headed across the country. In the Bustamante lab Jeremiah studied population genetics with a focus on genome evolution. He developed a strong foundation in population genetics, statistics and programming. During this time, his interests again shifted and he began to focus more on genetics and particularly in understanding genome evolution and gene duplication/structural variation.

While living in Ithaca there was another addition to the family. With a midwife at his side, Jeremiah delivered his beautiful daughter, Oonagh, in October of 2007.

Upon completing his Ph.D. at Cornell University in May 2010, he is taking a position as a computational biologist at Genentech in South San Francisco and the family is again moving across the country.

To Angie and the kids

ACKNOWLEDGMENTS

Special thanks goes to my thesis advisor, Carlos Bustamante for your continued support of my projects. During my time at Cornell you have not only allowed me the freedom to follow my interests but have excitedly pushed me to pursue these interests even though they have sometimes fallen outside your area of expertise. Our interactions during these projects have been some of the greatest learning experiences of my educational career. In addition to Carlos, thanks also goes to the other members of my committee, Andy Clark and Adam Siepel. It has truly been a life changing and exciting experience to work with some of the most amazing minds in population genetics and genomics.

I would not have progressed to this point without the help and support of my good friends and undergraduate mentors Bryan Carstens, Darin Rokyta and Chris Smith, as well as the members of the Bustamante lab, past and present. Ryan, Hong, Kasia, Kirsten, Abra, Koni, Kirk, Amit, Andy, the Adams, Scott, Shaila, Keyon and Danni you have all provided invaluable support and I hold you all among my good friends. Three of you in particular have had a major impact on both my academic and non-academic life. Amit Indap and Andy Reynolds, you were not only essential in helping me to gain a better understanding of scripting and programming but also I feel we have become great friends. I hope to have many more great rides with both of you in the future. Amit, without your patience and help in particular, my first years at Cornell would have been much more difficult. You taught me to program.

My office mate, Kirk Lohmuller, has potentially had the largest impact on me of any of my cohort. Kirk, our hours of conversation over the last two

years of my degree have greatly influenced both my understanding and my thinking about population genetics and evolution. Through this time, Kirk, you have also become one of my very best friends.

Finally I would like to thank my family. Thank you Mom and Dad for your help and support over the years. I would not have been able to make it without the support that you have provided. Thank you also to my wife Angie. You have been a continuous source of help, support, guidance and patience through my academic career. The past nine years have not always been easy and we have had to make sacrifices and hard decisions, but we have had a lot of fun and exciting and interesting experiences along the way. I would also like to thank my stepdaughters, Gwyn and Kelsey for being a source of fun and distraction from the sometimes-monotonous work of research. I think this road has probably been hardest on the two of you. Your mother and I have not always had the time to do the things you wanted or the money to do them but you have both been pretty patient. I hope you have had some fun along the way and I hope you enjoy the years to come. And lastly, thanks to my beautiful daughter Oonagh, you have brought much additional joy to our lives. You are too young to realize it now, but you will likely benefit most from our long road.

TABLE OF CONTENTS

Biographical Sketch	iii
Dedication	v
Acknowledgements	vi
Table of Contents	viii
List of Figures	ix
List of Tables	xi
List of Abbreviations	xii
Chapter 1: Population Genomics of Copy Number Variation in the Domestic Dog	1
References	23
Chapter 2: Copy number variation of <i>CCL3</i> -like genes affects rate of progression to simian-AIDS in rhesus macaques (<i>Macaca mulatta</i>)	26
References	70
Chapter 3: A novel method for mapping the pseudoautosomal boundary and application to 5 canid species and Rhesus Macaque	74
References	101
Appendix: Supplementary tables for Chapter 1	104

LIST OF FIGURES

Figure 1.1: Copy number HMM	4
Figure 1.2: Genome and Length distribution of discovered CNVs	7
Figure 1.3: Frequency Spectrum of CNVRs	9
Figure 1.4: Population Distribution of CNVs Across Dog Breeds	10
Figure 1.5: Copy Number Variation Hot-spots	12
Figure 1.6: Example of Average LD Decay Between SNPs and Between SNPs and CNVs in Two Breeds	15
Figure 1.7: SNP/SNP and CNV/SNP Pair-wise LD values within 1 Mb windows	16
Figure 1.8: Definition of CNVRs	21
Figure 2.1. Calibration curve for rtPCR assay using A431 cell line as a standard	30
Figure 2.2: Calibration and Verification of rtPCR copy number	32
Figure 2.3: Histogram of copy number estimates	34
Figure 2.4: Rhesus macaque survival analysis	36
Figure 2.5. Structure results of the retrospective individuals from the 53 microsatellite loci sorted by assumed population	38
Figure 2.6. PCA results for the retrospective sample	40
Figure 2.7. Heat plots summarizing genetic relatedness in the sample based on 53 unlinked microsatellite loci	42
Figure 2.8: Test of genome-wide significance of CCL3L CNV	44
Figure 2.9. Bootstrap simulations to assess power of Cox proportional hazard regression of survivorship on <i>CCL3L</i> copy number applied to each population separately	47

Figure 2.10: Population and species level copy number variation	50
Figure 2.11: Predicted Kaplan-Meier survival curves based on Cox-Proportional hazard model of post-SIV survivorship including <i>CCL3L</i> copy number and population-of-origin as covariates	53
Figure 3.1: Validation of novel method with human HapMap data	79
Figure 3.2: Example of novel method run on a single male dog chip	81
Figure 3.3: Heat-map representation of the transformed posterior probabilities from spot_PAR for all dogs and wild canids	84
Figure 3.4: Extended PAR attrition in domestic dogs	86
Figure 3.5: Evolution of PAR genic content across the Mammalian groups	88
Figure 3.6: GC content of human and canine PAR	91

LIST OF TABLES

Table 2.1: Results of necropsy for 57 animals used in the retrospective study	28
Table 2.2. Likelihood ratio test statistics for analysis of multiple variables contributing to survivorship based on Cox proportional hazard model. The test statistics are asymptotically χ^2 distributed	35
Table 2.3: Results of survival analysis	46
Table 2.4: Summary statistics for CCL3L copy number distribution among primate species and populations	55
Table 2.5: Total number of polymorphic sites found per primer/probe/individual for <i>CCL3L</i> rtPCR assay	64
Table 2.6: Summary of microsatellite data	66

LIST OF ABBREVIATIONS

aCGH	Array comparative genome hybridization
BAC	Bacterial artificial chromosome
CNV	Copy number variation
CNVR	Copy number variable region
FISH	Fluorescent <i>in situ</i> hybridization
HMM	Hidden Markov model
kb	kilobase
LRT	Likelihood ratio test
Mb	Mega base
PAR	Pseudoautosomal region
pdg	Per diploid genome
rtPCR	Real-time PCR
SIV	Simian immunodeficiency virus
SNP	Single nucleotide polymorphism
SV	Structural variation

CHAPTER 1

Population Genomics of Copy Number Variation in the Domestic Dog¹

1.1 Abstract

Recent research has greatly expanded our understanding of the extent of copy number variation (CNV) in humans. However, our understanding of the mutational properties and population genetics of CNV is still in the early stages, due to limited sampling of both individuals and taxa. Here we begin to address this issue by completing the largest analysis of CNVs in a non-human mammal done to date. We discover and analyze CNVs in combined set of over 781 domestic dogs and 111 wild canids using a genome-wide 125K Affymetrix SNP chip. Our analysis reveals nearly 10,000 CNVs segregating in domestic dogs and covering nearly 400 Mb of the dog genome. More than 90% of these regions are novel to this study. Our large data set revealed 15 regions of CNV hotspots in the dog genome. Additionally, we evaluate linkage disequilibrium (LD) between CNV and surrounding SNPs and find a much faster decay of LD when we look across all breed groups in CNVs than in SNPs. We also find evidence that the majority of CNVs are deleterious, with deletions being more strongly affected by negative selection than duplications. Taken together, these analyses provide a detailed picture of the population genetic forces impacting the distribution of CNVs within and among dog breeds as well as among dogs, coyotes, and wolves. The CNV map we have created here will further refine regions of the dog genome that are unique from the wolf and may, therefore, contain domestication mutations and/or species-specific changes.

¹ Degenhardt et al. in preparation.

1.2 Introduction

Copy number variants (CNVs), sizeable duplication or deletion mutations segregating within populations, are a major component of genetic variation in mammals^[1-4]. Despite the important role CNVs play in disease^[5, 6], an understanding of the population genomics of CNVs across multiple species, and the role of CNVs in normal phenotypic variation within and/or between populations is limited^[7]. Purebred domestic dogs, due to their extreme phenotypic variation, closed breeding populations, and history of selective breeding, provide an ideal system for addressing these questions.

The domestic dog (*Canis lupus familiaris*) is likely the oldest domesticated animal,^[8] with intensive human-controlled breeding practices driving diversification of functional form across hundreds of purebred lines. The complex demographic history of the domestic dog likely included several severe bottlenecks/founding events with a broad range of occurrence and severity, breed crosses to capture novel traits into distantly related breed stocks and breeding of closely related individuals to form pure lines^[9]. As a consequence, dog breeds are characterized by low genetic diversity and high linkage disequilibrium (LD) within breeds and high levels of diversity and low LD across breeds^[10-12]. Here, we use the unique population structure of the domestic dog to investigate the population genetics and mutational dynamics of CNVs on a genome-wide scale. The primary focus of this study was to characterize the extent to which CNVs are private to an individual, segregating within a breed, fixed within breeds, or segregating across multiple breeds. Likewise, we sought to understand the extent to which CNVs are consistent with, and informative about, the demographic history of domestic dogs, and to compare patterns of CNV among dogs and their closest relatives (wolves,

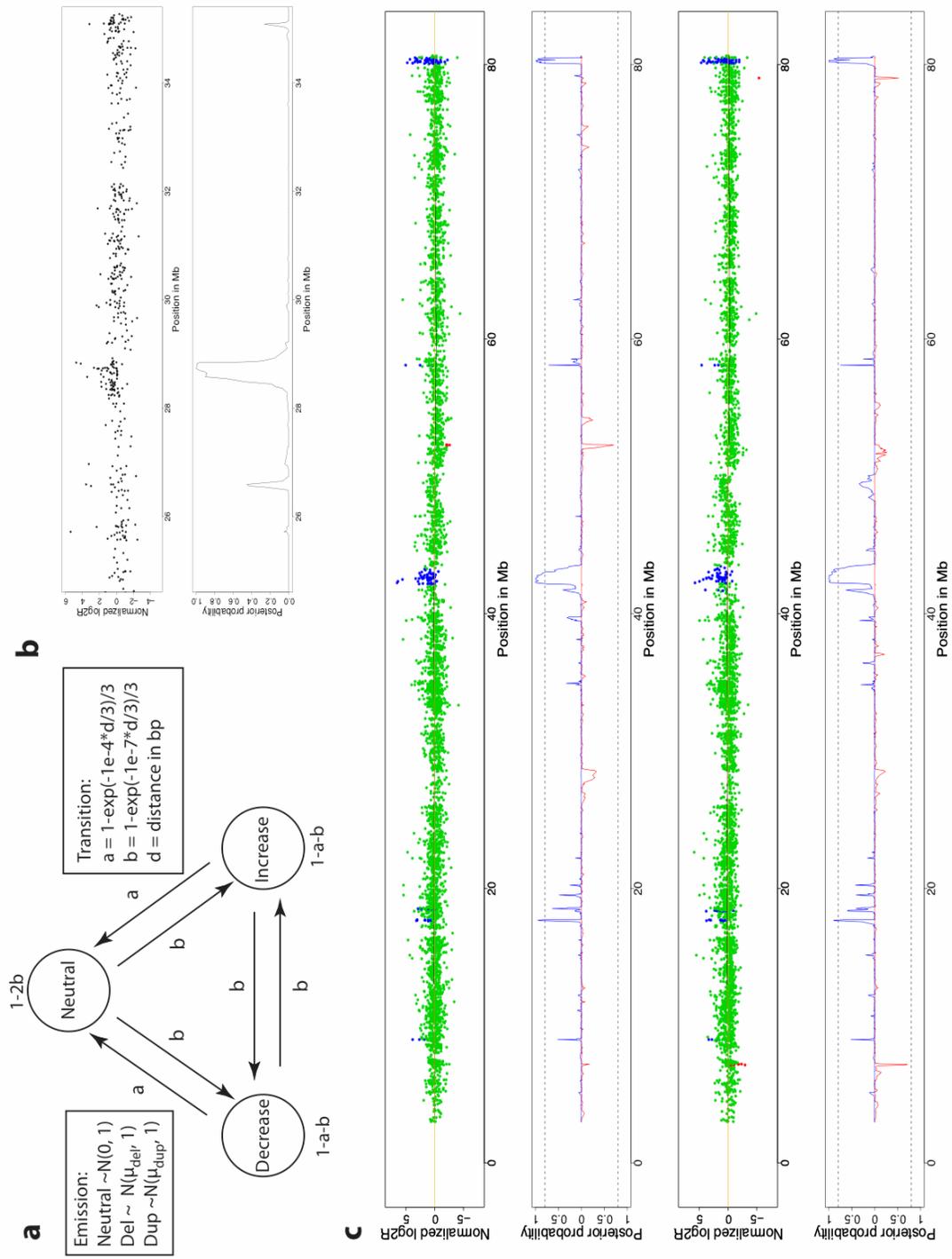
coyotes).

1.3 Results

We estimated genome-wide copy number by analyzing intensity data from the Affymetrix Canine V2 custom SNP chip used to genotype 781 domestic dogs and 111 wild canids. The unique nature of the Canine V2 chip with nearly equal numbers of invariant (63,706) and variable (61,468) probes necessitated the development of a new CNV calling algorithm. The new algorithm, termed **SPOT_CNV**, uses a Hidden Markov Model (HMM) framework^[13] (see Figure 1.1A) to model the \log_2 intensities for the 25-mer probes on the array. Most HMM approaches for CNV detection use the Viterbi algorithm to provide a single most probable path through the data^[13, 14]. However, we rely on the Forward/Backward algorithm to estimate the posterior probability of a given state (i.e., duplicated, unchanged, or deleted relative to reference; see Methods). The key benefit of quantifying uncertainty in the modeled copy number states for each SNP/probe position (Figure 1.1B) is that it allows us to focus our analyses on calls with high statistical evidence of support.

As described in the Methods section, we empirically defined two subsets of the data for analysis. The first is a “discovery panel” comprised of samples genotyped using the optimized Affymetrix protocol with at least 48 dogs per batch and standard deviation of the \log_2 intensities < 0.35 , which we used to discover CNV regions in the dog genome. The second dataset is a “population genetic analysis panel” consisting of all arrays run with standard deviation of the \log_2 intensities < 0.35 , which we use to investigate the frequencies of CNVs across different breeds.

Figure 1.1: Copy number HMM. A) Schematic drawing of the **Spot_CNV** HMM showing transition and emission distributions. B) Example of **Spot_CNV** applied to data showing evidence of a duplication region. This also shows the posterior probability of the duplication event. C) Shows the normalized $\log_2 R$ colored by calls (green = copy neutral, red = deletion, blue = duplication) for an entire chromosome for a replicate chip.



Using **SPOT_CNV** with the discovery panel, we mapped 9789 high confidence CNVs, defined as segments with posterior probability for that state > 0.95 , found in ≥ 2 individuals, and composed of ≥ 8 consecutive SNP positions agreeing in state. The CNVs were clustered into 1220, non-overlapping, copy number variable regions (CNVR) that cover nearly 400 Mb (15%) of the 38 canine autosomes (Figure 1.2A; see Methods for description of clustering). We find that 122 (10%) of the discovered CNVRs overlap with segmental duplications and that 929 (76%) overlap with genes listed in Ensembl (see Methods). Our results are qualitatively similar to those obtained in a human study using the Affymetrix 500K SNP chip^[1], in terms of the number of CNVRs detected, fraction of the genome that is copy number variable, as well as genic content.

We observe that many of the high frequency CNVs discovered here are consistent with those discovered in two previous studies of canine CNV, which examined $n < 3$ dogs per breed for 9 and 17 dog breeds, respectively^[15, 16]. Overall, we replicate $\sim 60\%$ of the autosomal CNVs discovered in Nicholas et al.^[16] and 79% described by Chen et al.^[15]. However, previously discovered CNVs only account for 94 of the 1220, or $\sim 8\%$ of the CNVRs discovered in our study. Therefore, more than 90% of the CNVRs discovered in this study are novel, highlighting both an increase in the proportion of the genome covered by the Affymetrix SNP chip compared to tiling array designed for segmental duplications^[16], as well as the substantially larger population of both dogs and breeds examined here^[15, 16].

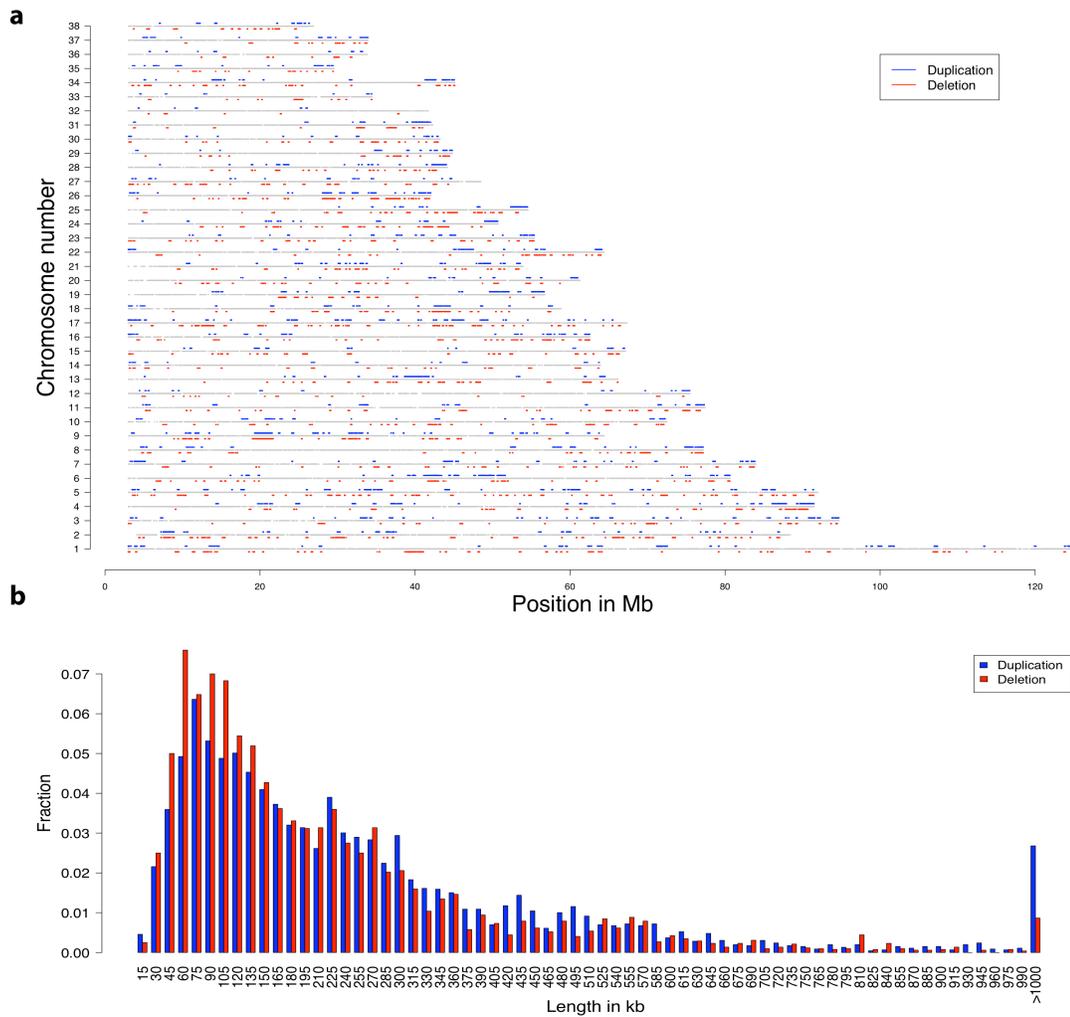


Figure 1.2: Genome and Length distribution of discovered CNVs. A) Map of the autosomal CNVRs found in the 551 dogs in the discovery set. Blue regions highlight duplications and red indicate deletions. Grey bars represent chromosomes and gaps in the bars show gaps in SNP coverage. B) Length distribution for deletions (red) and duplications (blue).

We find that regions of CNV are distributed heterogeneously across canine autosomes, with some regions enriched for CNVs while others are relatively barren (Figure 1.2A). We also observe that many regions segregate for both duplications and deletions. Furthermore, the majority of CNVs detected are less than 300 kb in length, consistent with CNV studies in

humans that find smaller CNVs occur with greater frequency than larger ones^[7].

To examine the differential patterns of selection acting on deletion versus duplication CNVs, we evaluated the length and frequency distribution for CNVs seen in at least two individuals (i.e., non-singletons). We find that deletions are, on average, shorter than duplications (p-val = $<2.2 \times 10^{-16}$; one sided t-test; Figure 1.2B). Likewise, we find that deletions tend to be found at lower population frequencies than duplications (p-value = 2.826×10^{-4} ; Mann-Whitney U test; Figure 1.3A), and that large CNVs tend to occur at a lower frequency than small CNVs (Figure 1.3B) with an overall dearth of long deletions relative to duplications. These results, taken together, suggest that deletions, and specifically large deletions, are more strongly selected against than duplications.

Initial analysis using the population genetic analysis panel (i.e. 781 individuals from 75 breeds and 111 wild canids, see Methods and Figure 1.4), revealed 15 large, complex (multi-state) CNV regions of the dog genome (Figure 1.5). Some regions are fixed for a particular state in some breeds, while other regions are segregating for both duplications and deletions within a single breed. Additionally, several of these regions were comprised of adjacent smaller copy number events, as some dogs were alternately segregating for duplications and deletions (Figure 1.4 and Figure 1.5).

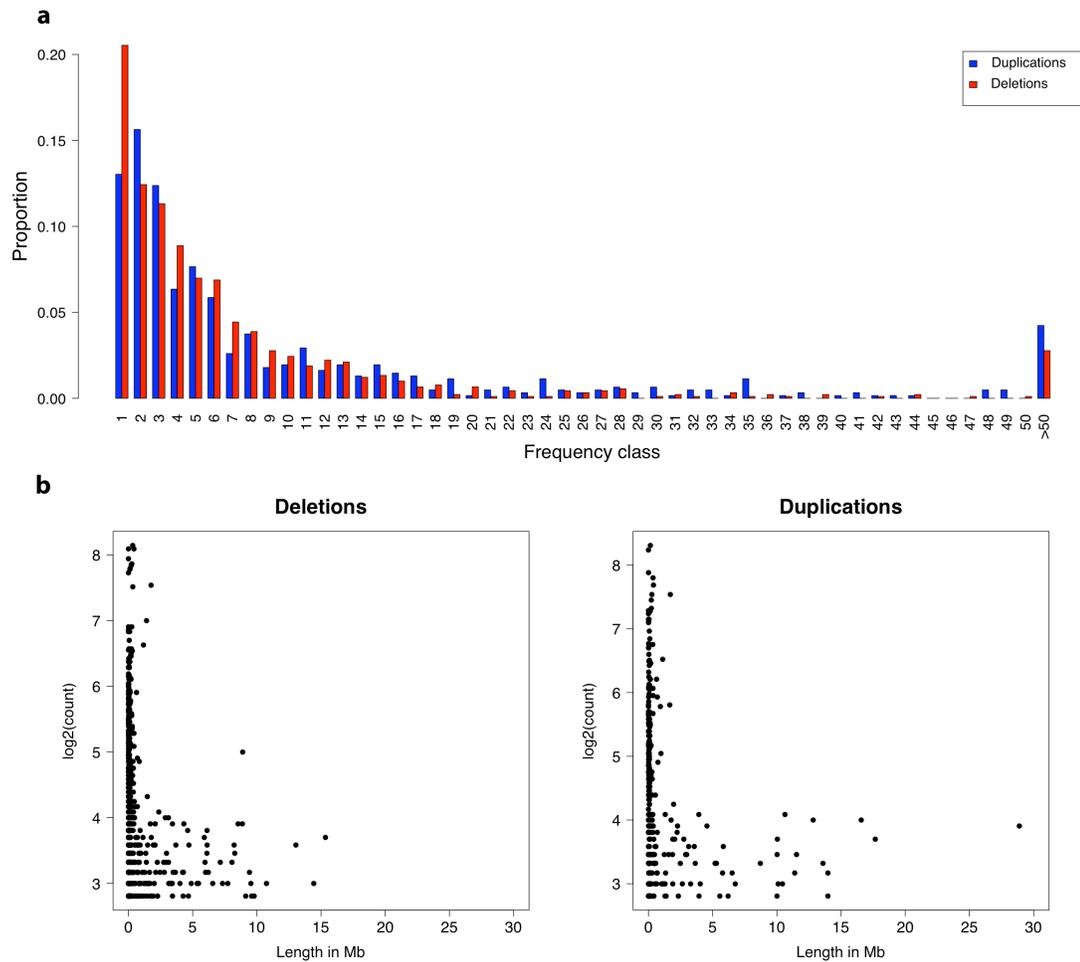


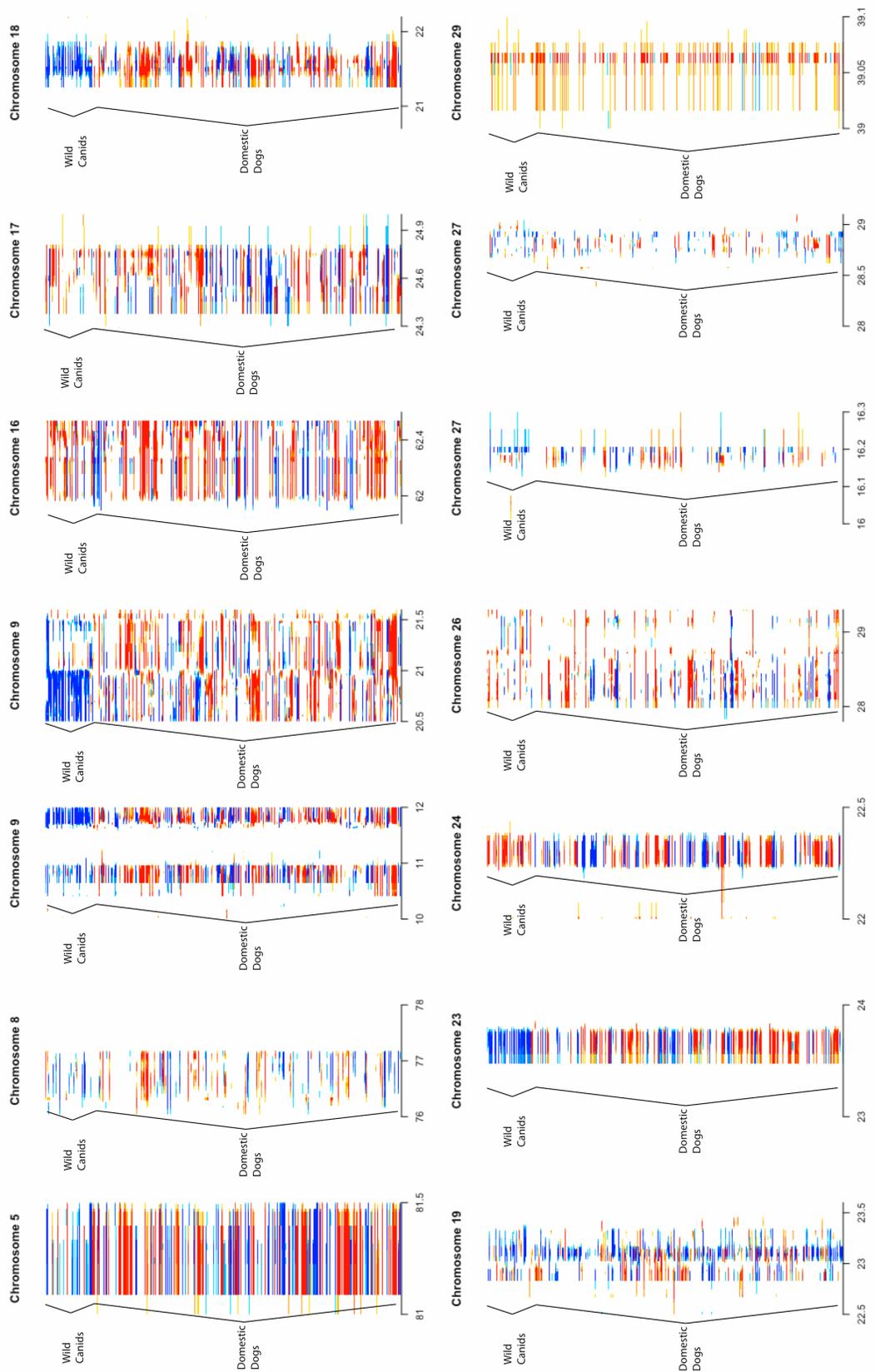
Figure 1.3: Frequency Spectrum of CNVRs. A) Frequency spectrum of detected CNVR for deletions (red) and duplications (blue). All dogs with any CNV overlapping the CNVR were counted. The resulting count is the maximum count spectrum. The actual frequency of individual events will likely be skewed to more rare events. B) Scatter plots showing the $\log_2(\text{count})$ of number of individuals carrying CNV plotted by the length of CNV regions.

external and independent validation of this observation.

Analysis of the population genetics panel also demonstrates that CNVs differ widely in their distribution across breed groups and wild canids. We find, on average, that an individual domestic dog carries 18 regions that are deleted or duplicated relative to the population median, with an average length spanning a total of 4.2 Mb of the genome. The number of CNVs and area of the genome covered is similar in wild canid groups with 15 CNVs and 4.3 Mb on average in wolves and 16 CNVs and 5.1 Mb in coyotes. When considering CNVs defined by the discovery panel, we find that 728 (60%) are segregating at a frequency of >40% in at least one breed. Of these, 158 (22%) are segregating at a high frequency in only a single breed, 570 (78%) are observed at high frequency across multiple breeds, and 47 (6.5%) are fixed in one or more breeds. In wild canids, we find 168 CNVs segregating at frequency >40% with 61 occurring at high frequency only in wild canids. Of these, 17 CNVs are fixed in coyote with three of these only observed in coyote and red wolf at high frequency. These coyote/red wolf specific regions, which are all deletions, may represent either genomic regions lost in coyote and red wolf since the divergence from a common ancestor or regions of novel insertions in grey wolves and dogs.

Interestingly, many closely related breeds do not appear to share CNV states. For example, we found a deletion at position 20.5 Mb on CFA15 which is found at high frequency in whippets and West Highland white terriers (Westies), but at low frequency in greyhounds and Scottish terriers (Figure 1.4C), the closest “sister breeds” to whippets and Westies respectively.

Figure 1.5: Copy Number Variation Hot-spots. Heat-map representation of CNVs for 15 hotspots of copy number variation detected in the dog genome.



That is not to say that all regions are devoid of historical information. For example, there is a CNV region on CFA6 at 43.8 Mb which is found at high frequency only in Akitas, Chinese Shar Pei and Chow Chow, three breeds from the “ancient/ancestral group” defined by Parker et al.^[12] (Figure 1.4B). A potential reason CNVRs may mask historical relationships is that high levels of recurrent mutation lead to homoplasy, that is, breeds exhibit convergence in CNV state rather than identity by descent.

To further assess the mutational dynamics of high frequency CNVs, we quantified LD between common CNVs (frequency $\geq 5\%$ across all dogs) and SNPs, and compared this to LD observed between frequency and region matched SNPs (see Methods). When assessed within breed groups, we find that many high frequency CNVs and SNPs are taggable by a SNP within 1 Mb window at an r^2 value >0.80 (38% (range across breeds 18-82%) for CNVs as compared to 83% (range 59-99%) for SNPs (Figure 1.6. When the correlation is calculated across all breed groups we observe a much faster decay of the correlation between SNPs and CNVs than between SNPs alone (Figure 1.7). In addition, the percent of CNVs that are tagged drops to 0.2% for CNVs as compared to 13% for SNPs.

1.4 Discussion

Our survey of copy number variation in domestic dogs and wild canids has important implications for evolutionary and domestication genomics. First, we have identified regions of differentially fixed copy number among coyote, dogs and wolves; these may well harbor species defining mutations, including key domestication loci. Therefore, cataloging the specific genes and mutations within and among the CNV regions is an important next step towards understanding the genetic basis of dog domestication.

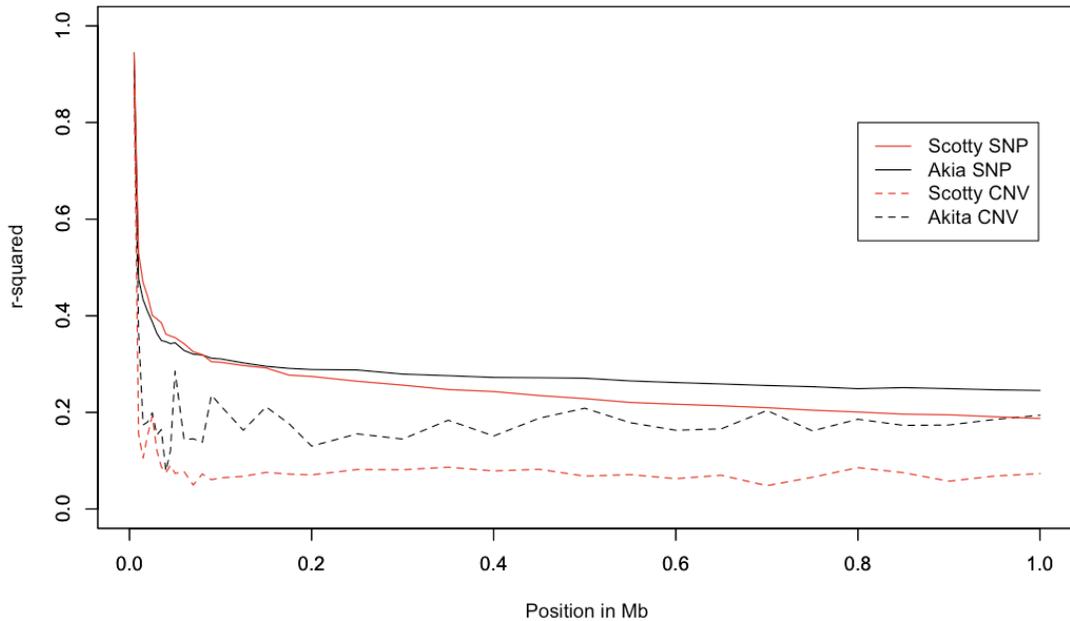


Figure 1.6: Example of Average LD Decay Between SNPs and Between SNPs and CNVs in Two Breeds. Within breed LD calculated for two representative breeds (Akita in black and Scottish Terrier in red). The dotted lines show the curve for CNVs while the solid lines show the same calculation for SNPs.

Likewise, we have demonstrated that there are many fewer CNVs at high frequency within the wild canid groups than among domestic dog breeds. High frequency and breed fixed CNVs are likely the result of strong selection, breed bottlenecks, and/or “popular sire effects” differentially fixing CNVs among breeds. The presence of extensive, and relatively rare, CNVs in all breeds reflects the high mutation-drift-selection equilibrium maintained in domestic dogs due to small effective population size within breeds. These patterns are in sharp contrast to the distribution of CNVs among human populations where few, if any, CNVs reach appreciable frequency differences among human populations^[7, 17].

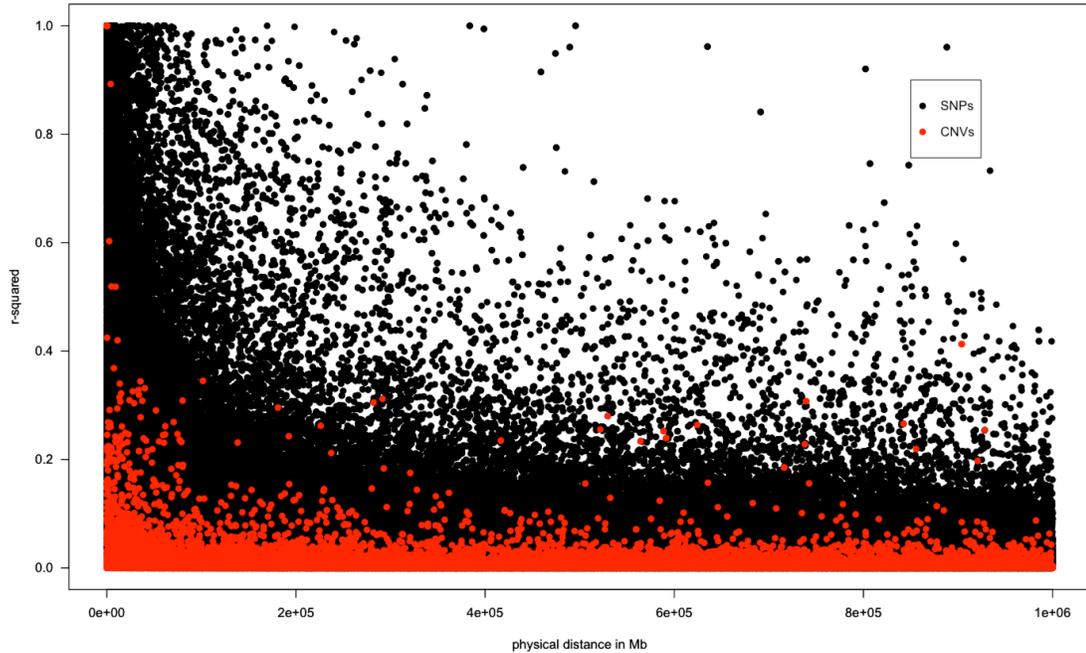


Figure 1.7: SNP/SNP and CNV/SNP Pair-wise LD values within 1 Mb windows. Linkage disequilibrium among SNPs and CNVs. Red dots highlight all pair-wise genotype r^2 values as calculated in `PLINK` version 1.5 between CNVs with a frequency between 5% and 45% and SNPs. Black dots show the same calculation for all SNPs with frequency between 5% and 45%.

Additionally, our results show that CNVs within the domestic dog genome are less taggable than SNPs, especially when evaluated across dog breeds. Based on these results we suggest that CNV mutations have occurred on several haplotype backgrounds and are therefore recurrent when viewed across all dog breeds.

These findings are inconsistent with the hypothesis of Chen et al. that many CNVs segregating in dog breeds could be informative regarding breed relationships and in high LD with surrounding SNPs^[15]. Chen and colleagues based their results on a sampling scheme of generally one individual per breed^[15]. We demonstrate here that the use of a more complete panel, with

an average of 12 dogs per breed and 79 breeds provides a more accurate picture of the CNV landscape in the canine genome.

Taken together, these results highlight the importance of characterizing CNVs across multiple model systems in order to understand how mutation, natural selection, and genetic drift shape the evolution of copy number variation across time and space.

1.5 Methods

Sample Collection

See vonHoldt et al.^[18] and Boyko et al. (PLoS Biology; in submission) for details of sample collection. During the data generation phase of this project, there was modification in the Affymetrix protocol used for running the CanFam V2 chips. The Affymetrix protocol was originally written for Human SNP chip, and the volume to be hybridized was too large to be load on the canine SNP chip. Therefore, in the first experiments we only hybridized half of the volume, and in the later experiments we evaporated after the labeling step so the 90 ng could be hybridized. This change resulted in higher concordance with less noise (and, therefore, more CNV calls) between dogs run in duplicate for the second half of our sample. Therefore, we split our data into two sets. The first consisted of the CNV discovery set and was made up of batches run after the change in protocol and with greater than 48 individuals per batch. Individuals were in these batches were retained for analysis if the standard deviation of the \log_2 intensity ratio was less than 0.35. This set of dogs was then used for characterizing regions of the dog genome that are copy number variable and would be used in later analyses. The second data set was the “population genomic analysis set” and consists of all dogs where the standard deviation of the \log_2 intensity ratio was less than 0.35.

Array normalization and CNV Calling

Here we describe a set of C++ programs for processing array data for CNV analysis as well as a novel flexible HMM for detection of copy number variable regions from array data. This method is general enough to work on any platform with data that can be transformed into \log_2 intensities, including Affymetrix genotyping chips and NimbleGen tiling path arrays. The data processing can be broken up into three categories: pre-processing, HMM analysis, and post-processing.

Pre-processing

To pre-process the probe intensity values, we first normalized the values using the standard quantile normalization as implemented in the Affymetrix Power Tools (APT). These normalized probe intensities were then extracted from the CEL files using `apt_cel_extract`. Next, the probe values were summarized using a C++ program we developed for the Affymetrix Canine V2 chip, to provide a single intensity value per SNP position assayed. We here refer to this as the total position intensity (TPI). The program works by first calculating the median of each group of 5 A and B-allele, sense and anti-sense probes. Next the mean of the sense and anti-sense probes is calculated for the A and B-alleles separately to get the average A and average B allele intensities. These two intensities were then summed to get the TPI. This procedure was run for all ~125000 SNP positions assayed on the CanFamV2 SNP chip. We found that this simple procedure gave an efficient and robust summary of the probe intensity data.

To calculate copy number variation for each individual we next needed to define our reference set. For the autosomal dataset, we opted to use the median value for each TPI as the “reference” value. Under the assumptions

that 1) most of the genome is diploid and 2) that most CNVs are at low frequency, the median value should be a reasonable proxy for the diploid state. Previously studies have shown that this is a reasonable solution^[19]. The final step of the summary process was to calculate $\log_2(\text{TPI}/\text{reference value})$ to get the \log_2 ratio (\log_2R) value for each individual and position. These values were then fed into our HMM method to determine the posterior probability of CNV states for each position.

Copy number HMM

Our novel hidden-Markov-model, **SPOT_CNV**, is applied to each chromosome for each individual separately to reduce memory requirements. We begin by standardizing the \log_2R values for each chromosome by subtracting off the mean \log_2R value and dividing by the standard deviation of the \log_2R . Similar to what has been noticed in previous CNV studies^[20], we found that the normalized \log_2R values retain a low-frequency oscillation characteristic when viewed along the each chromosome. Early analyses showed that this oscillation occurred at a much larger scale than the average size of CNVs of interest in our dataset, and while the cause of this pattern remains unknown, it was correlated across individuals run using the same Affymetrix protocol, therefore, we chose to use an additional simple smoothing step for each chromosome. We employed a sliding-window method where the mean normalized \log_2R value was calculated for a 12 Mb window centered over the SNP position of interest. The \log_2R value for that position was then shifted by the mean value and the window was moved by one SNP position. We found this method worked adequately to remove the oscillation without substantially reducing our power to detect CNVs.

We utilize the full Forward/Backward algorithm in **SPOT_CNV** to

calculate the posterior probability of each of the three states in our model (deletion, neutral and duplication). **SPOT_CNV** uses normal emission probabilities with the mean optimized for each state. A heterogeneous transition matrix was used where the transition probabilities were determined by the distance between probes using the following equation:

$$\frac{1 - e^{(-1 \times 10^{-4} * d)/3}}{3}$$

d is the physical distance between SNPs in base-pairs. This model is similar to that used by Li and Stephens for recombination-rate estimation^[21] and the pennCNV model used for CNV detection^[13], however, the map-size of the dog genome is unknown, therefore we use the approximation of 1 Mb per centimorgan.

Next we apply a simple set of posterior decoding rules (described in the text) to determine the position of the breakpoint for each CNV position.

Clustering CNVRs

To define regions of the dog genome that are copy number variable for analyses we employed two different definitions. First for frequency spectrum we joined all CNVs that had any overlap with the region (i.e. we extended CNVs to cover the maximal length in a given region). The frequency of a given CNVR was calculated by counting all dogs with an overlapping CNV of the same state called as the highest posterior probability state in that region regardless of whether the breakpoint in all dogs was identical (Figure 1.8).

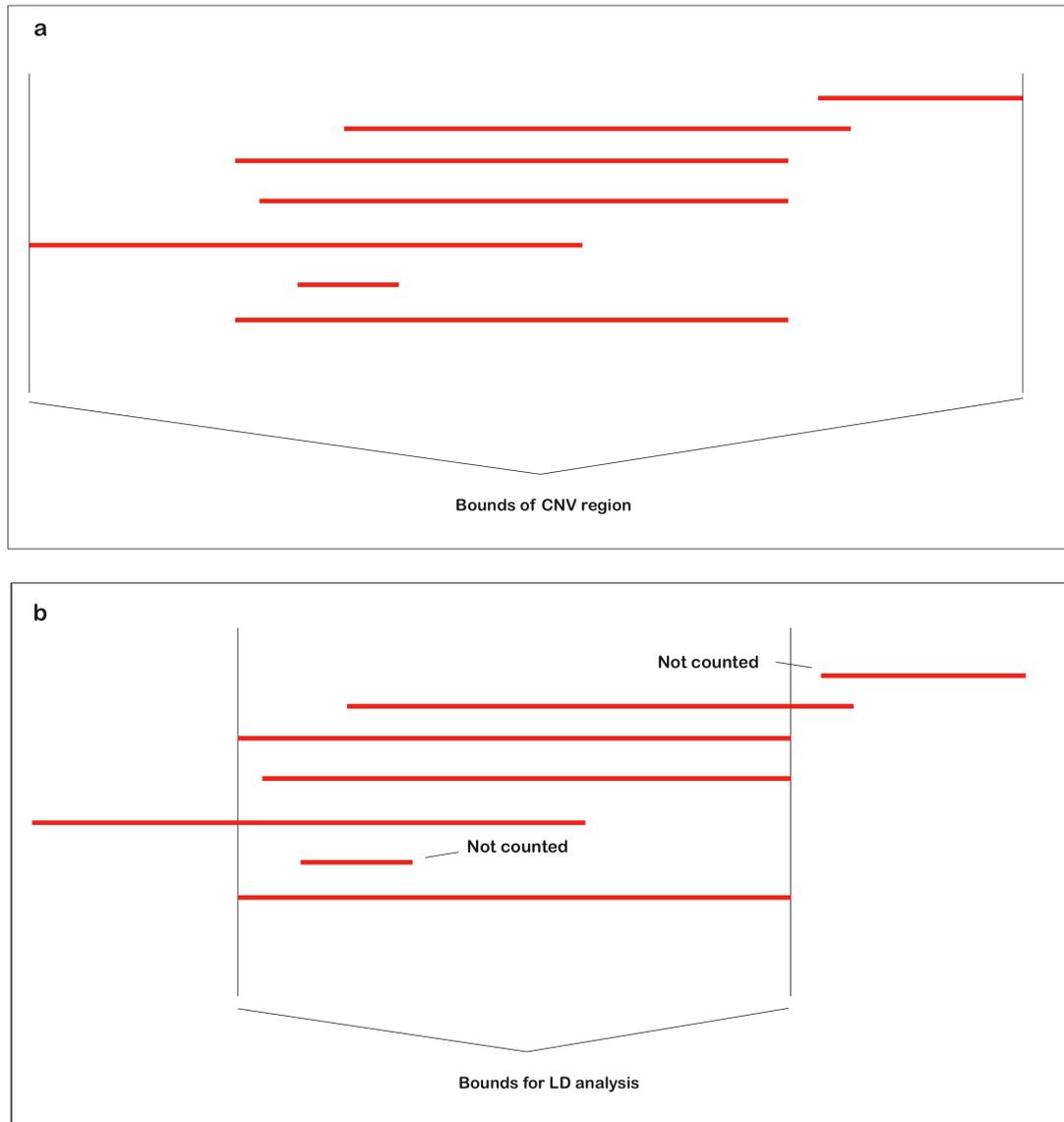


Figure 1.8: Definition of CNVRs. a) Schematic drawing showing how CNVRs were classified for genomic plotting and frequency analysis. b) Schematic drawing showing how CNVRs were classified for LD analysis.

For example, a region found in five dogs with three showing 100 kb deletions and two showing 50 kb overlapping deletions will be counted as being found in five dogs. For the analyses of LD of CNVs we took the regions where the majority of dogs that were segregating for that event were concordant as the breakpoints across all dogs. That is, given a position with many CNVs across

many dogs, we defined the breakpoints to be the maximal region where >50% of dogs that were segregating for that event continued to show a non-copy-neutral state as the highest posterior probability state.

Heat-map representation

To generate the heat-map representation of the genomic CNVs, the posterior probability of duplication and deletion states were transformed to give duplications a positive value proportional to the posterior probability and deletions a negative value. Regions of no CNV were attributed a value near 0. All dogs analyzed were then placed in a single matrix per chromosome and the function **image.plot** in **R** was used to obtain the heat-map.

Linkage Disequilibrium Comparison

To calculate r^2 between SNPs and CNVs we used the SNPs called as in Boyko et al. (in prep) and encoded CNVs as a homozygous genotype located at the center of the most common breakpoint of the event when assessed across all dogs. We then calculated the genotype r^2 in Plink using the flags

```
--allow-no-sex  
--dog  
--r2  
--ld-window-r2 0  
--ld-window-kb 1000  
--ld-window 99999
```

Bioinformatics and Data Base versions

The Ensembl release 56 gene set was obtained and the overlap assessed using the UCSC genome browser CanFam2 (<http://genome.ucsc.edu>)

CanFam2 segmental duplication coordinates were obtained from: (<http://humanparalogy.gs.washington.edu/canFam2seqdup/canFam2.html>).

REFERENCES

1. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. (2006) Global variation in copy number in the human genome. *Nature* 444(7118): 444-454.
2. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, et al. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* 40(10): 1166-1174.
3. Lee AS, Gutierrez-Arcelus M, Perry GH, Vallender EJ, Johnson WE, et al. (2008) Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. *Hum Mol Genet* 17(8): 1127-1136.
4. Graubert TA, Cahan P, Edwin D, Selzer RR, Richmond TA, et al. (2007) A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genet* 3(1): e3.
5. Liu W, Sun J, Li G, Zhu Y, Zhang S, et al. (2009) Association of a germline copy number variation at 2p24.3 and risk for aggressive prostate cancer. *Cancer Res* 69(6): 2176-2179.
6. Szigeti K, Lupski JR. (2009) Charcot-marie-tooth disease. *Eur J Hum Genet* 17(6): 703-710.
7. Itsara A, Cooper GM, Baker C, Girirajan S, Li J, et al. (2009) Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet* 84(2): 148-161.
8. Germonpréa M, Sablinb MV, Stevensc RE, Hedgesd REM, Hofreitere M, et al. (2009) Fossil dogs and wolves from palaeolithic sites in belgium, the ukraine and russia: Osteometry, ancient DNA and stable isotopes *Journal of Archaeological Science* 36(2): 473-490.

9. Ostrander EA, Kruglyak L. (2000) Unleashing the canine genome. *Genome Res* 10(9): 1271-1274.
10. Sutter NB, Eberle MA, Parker HG, Pullar BJ, Kirkness EF, et al. (2004) Extensive and breed-specific linkage disequilibrium in *canis familiaris*. *Genome Res* 14(12): 2388-2396.
11. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438(7069): 803-819.
12. Parker HG, Kim LV, Sutter NB, Carlson S, Lorentzen TD, et al. (2004) Genetic structure of the purebred domestic dog. *Science* 304(5674): 1160-1164.
13. Wang K, Li M, Hadley D, Liu R, Glessner J, et al. (2007) PennCNV: An integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 17(11): 1665-1674.
14. Li C, Beroukhi R, Weir BA, Winckler W, Garraway LA, et al. (2008) Major copy proportion analysis of tumor samples using SNP arrays. *BMC Bioinformatics* 9: 204.
15. Chen WK, Swartz JD, Rush LJ, Alvarez CE. (2009) Mapping DNA structural variation in dogs. *Genome Res* 19(3): 500-509.
16. Nicholas TJ, Cheng Z, Ventura M, Mealey K, Eichler EE, et al. (2009) The genomic architecture of segmental duplications and associated copy number variants in dogs. *Genome Res* 19(3): 491-499.
17. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, et al. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451(7181): 998-1003.

18. vonHoldt BM, Han E, Pollinger JP, Lohmueller KE, Parker HG, et al. (2010) Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* .
19. Bengtsson H, Irizarry R, Carvalho B, Speed TP. (2008) Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* 24(6): 759-767.
20. Wang K, Li M, Hadley D, Liu R, Glessner J, et al. (2007) PennCNV: An integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 17(11): 1665-1674.
21. Li N, Stephens M. (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165(4): 2213-2233.

CHAPTER 2

Copy number variation of *CCL3*-like genes affects rate of progression to simian-AIDS in rhesus macaques (*Macaca mulatta*)²

2.1 Abstract

Variation in genes underlying host immunity can lead to marked differences in susceptibility to HIV infection among humans. Despite heavy reliance on non-human primates as models for HIV/AIDS, little is known about which host factors are shared and which are unique to a given primate lineage. Here, we investigate whether copy number variation (CNV) at *CCL3-like* genes (*CCL3L*), a key genetic host factor for HIV/AIDS susceptibility and cell-mediated immune response in humans, is also a determinant of time until onset of simian-AIDS in rhesus macaques. Using a retrospective study of 57 rhesus macaques experimentally infected with SIVmac, we find that *CCL3L* CNV explains approximately 18% of the variance in time to simian-AIDS ($p < 0.001$) with lower *CCL3L* copy number associating with more rapid disease course. We also find that *CCL3L* copy number varies significantly ($p < 10^{-6}$) among rhesus subpopulations, with Indian-origin macaques having, on average, half as many *CCL3L* gene copies as Chinese-origin macaques. Lastly, we confirm *CCL3L* shows variable copy number in humans and chimpanzees and report on *CCL3L* CNV within and among three additional primate species. Based on our findings we suggest that 1) the difference in population level copy number may explain previously reported observations of longer post-infection survivorship of Chinese-origin rhesus macaques, 2) stratification by *CCL3L* copy number in rhesus SIV vaccine trials will increase

² Previously published in Degenhardt et. al. (2009) under CCAL.

power and reduce noise due to non-vaccine related differences in survival, and 3) *CCL3L* CNV is an ancestral component of the primate immune response and therefore, copy number variation has not been driven by HIV or SIV *per se*.

2.2 Introduction

Rhesus macaques are the most widely used non-human-primate model of HIV/AIDS^[1]. We and several other research groups have reported substantial inter-individual variation in progression rates to simian-AIDS as well as population level differences between Chinese and Indian origin macaques^[2-5]. Understanding the genetic basis of these individual and population differences is critical to building reliable animal models of human HIV infection and AIDS progression.

In humans, an important host factor for HIV susceptibility is copy number variation at *CCL3L1*, a paralog of the *CCL3* gene^[6-12]. *CCL3* and *CCL3L1* encode chemokine ligands of *CCR5*, the main co-receptor used by HIV-1 for entry into host cells^[10-11]. Reduced *CCL3L1* copy number relative to the population median correlates with increased risk of acquiring HIV^[13], increased progression rate to AIDS^[13], and increased risk of maternal-fetal HIV transmission^[13-16]. After the discovery of copy number variation of *CCL3-like* genes, there have been a large number of studies in humans, expanding our understanding of the role of this variation in differential HIV susceptibility and progression. It has been shown that *CCL3-like* gene CNV plays a role in the level of chemokine production and chemotaxis^[9,13], controlling viral load^[13,15], cell-mediated immune response^[17], and most recently, HIV-specific gag response^[18]. However, currently it is not known whether copy number variation of *CCL3-like* genes plays a role in S/HIV immunity in other primates, although

it has been shown that copy number variation exists at these loci in chimpanzees^[13] and that this locus is duplicated in a rhesus macaque^[19].

2.3 Results

To investigate whether *CCL3-like* genes show variable copy number in rhesus macaques and more specifically, to study the role of the *CCL3-like* genes in SIV survivorship among rhesus macaques, we assayed copy number

Table 2.1: Results of necropsy results for 57 animals used in the retrospective study.

Animal	Survival time (months)	Clinical findings of necropsy (if not alive)	Origin
V272	.27	Severe lymphoid hyperplasia of small intestine lymphoid follicle	Indian
BA20	1.5	Hepatic steatosis, gastroenterocolitis	Indian
AT56	3	Amyloidosis, enterocolitis	Indian
BA19	3	Colitis (SIV)	Indian
T600	3.7	Cytomegalovirus infection	Indian
DE99	4	SIV infection	Indian
T590	4	Opportunistic infection	Indian
H405	4.5	Colitis, pancreatic amyloidosis	Indian
DG96	5	Pneumonia, CMV, cryptosporidium	Indian
R432	5.5	Giant Cell Disease (SIV)	Indian
AV90	6.5	Colitis, pneumonia, encephalitis, giant cell	Indian
AJ82	7	Enterocolitis	Indian
BE86	7	SIV infection	Indian
DE70	7	Pneumocystis, colitis, giant cell	Indian
DR43	7	SIV infection	Indian
DT52	7.2	Colitis (SIV)	Chinese
P205	7.6	Amyloidosis/intestinal	Indian
AE14	8	Enteritis (SIV)	Indian
DT69	8.3	Pneumonia	Chinese
N107	9	Cryptosporidium infection	Indian
DT67	9.7	Pneumonia	Chinese
FA97	10.3	Cytomegalovirus infection	Chinese
CK76	12	Lymphoid hyperplasia, Giant cell	Indian
CV94	12	Pneumocystis carinii	Chinese
DD88	12	Amyloidosis, colitis	Indian
EE54	12	SIV infection, pneumocystis	Indian
CN851	13	Pulmonary infarct, pneumonia, colitis	Indian
DD95	15	SIV infection	Indian
CF52	16	Lymphoid hyperplasia	Indian
CE45	17	Pneumonia	Chinese
FB04	18	Alive at time of Sampling	Chinese
V248	18	Colitis, amyloidosis (SIV)	Indian
V754	18	Alive at time of Sampling	Chinese
J304	19.2	Neoplasm/Lymphoma	Chinese
I553	21.5	Pneumonia	Indian
V515	21.6	Mycobacterium avium/intracellular	Chinese
AL26	22.8	Pneumonia/interstitial	Chinese

Table 2.1: (Continued)

P503	24	Lymphoma, hepatic amyloidosis, opportunistic infection	Indian
T687	24	Undetermined but SIV related	Indian
V190	24	Mycobacterium avium/intracellular	Chinese
V205	24	Intestinal amyloidosis	Indian
BE64	25	Pericarditis, vasculitis, cardiac and pulmonary thrombi	Indian
P045	26	Pneumocystis pneumonia, giant cell	Indian
P772	27.6	Pneumocystis carni infection	Chinese
T078	27.6	Mycobacterium avium/intracellular	Chinese
AV89	29	Lymphoma	Indian
BI33	30	Pneumocystis pneumonia	Indian
M008	32	Alive at time of Sampling	Chinese
AP09	33	Alive at time of Sampling	Indian
BG21	33	Alive at time of Sampling	Indian
BM47	33	Alive at time of Sampling	Chinese
DT46	33	Alive at time of Sampling	Chinese
L618	33	Alive at time of Sampling	Chinese
T153	33	Alive at time of Sampling	Indian
BE65	36	Mycobacterium avium infection	Indian
DT92	43	Alive at time of Sampling	Chinese
AJ07	84	Alive at time of Sampling	Chinese

variation at these genes in a cohort of 37 Indian origin and 20 Chinese origin animals previously infected with SIVmac at the Tulane National Primate Research Center. Individual animals were included in our retrospective study only if the clinical results of a necropsy confirmed health complications due to simian-AIDS at the time of euthanasia or if the animal remained AIDS free for at least 18 months post-infection (see Table 2.1).

An analysis of the shotgun and BAC reads of the *CCL3* and *CCL3-like* gene regions of the macaque genome revealed no fixed differences that would enable us to design a *CCL3-like* gene-specific primer or probe in this species (results not shown). Therefore, our assay, as designed, will detect both *CCL3* and all *CCL3-like* gene paralogs in rhesus as well as in chimpanzee and human cells, and we refer to the combined loci detected as *CCL3L*.

In order to estimate *CCL3L* copy numbers we used real-time PCR (rtPCR), and determined absolute copy numbers using two reference samples

(see Additional Methods and Figure 2.1 for calibration curve and methods for more details). The first is the human cell line A431, which has two copies per diploid genome of *CCL3* and two copies of *CCL3L1* (by fluorescent in-situ hybridization (FISH); Figure 2.1; Figure 2.2A; see also reference^[9]).

Figure 2.1

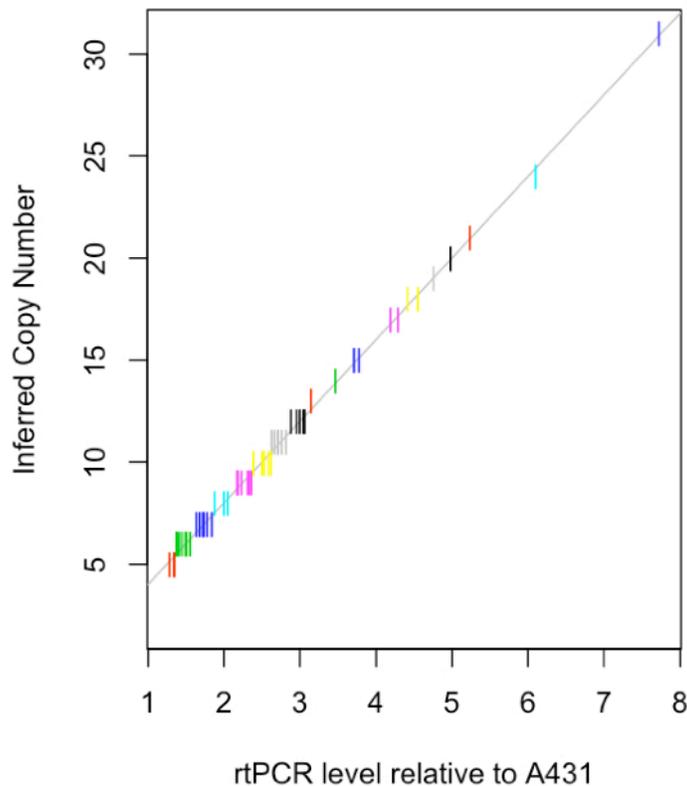


Figure 2.1. Calibration curve for rtPCR assay using A431 cell line as a standard. Since the A431 cell line has four copies of *CCL3L* (see Figure1A), *CCL3L* copy number is inferred as the relative rtPCR level for a sample, multiplied by 4 and rounded to the nearest integer. Each color represents a transition in copy number variation call (i.e., the break between 5 copies and 6 copies is denoted by a transition of red to green, and the break between 6 and 7 copies by a transition from green to dark blue).

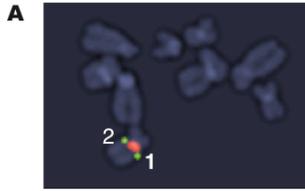
The second reference sample is the rhesus macaque genome donor, which

whole-genome shotgun sequencing analysis (WGSA) found to have between six and eight copies per diploid genome of *CCL3L* (Figure 2.2B). The use of two independent references allowed us to cross-validate our copy number estimates. Support for our CNV estimates also comes from a comparison of the rtPCR result to interphase FISH of a macaque cell line (MMU2 9133) (see Figure 2.2C&D).

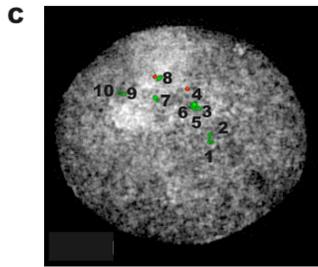
Using the rtPCR assay, we observed extensive variation in copy number of the *CCL3L* region among animals in our study, with a range of 5 to 31 copies per diploid genome (mean 11.05 ± 5.16 [sd]; Figure 2.3). Tables 1 and 2 summarize the results of Cox proportional hazard models^[20] for the survivorship data using *CCL3L* copy number and population-of-origin as potential covariates (see methods). Overall, we found strong evidence that reduced *CCL3L* copy number correlates with increased rate of progression to simian AIDS. Specifically, a model that includes *CCL3L* as a covariate (m_1) provides a significantly better fit to the data than the model (m_0) without *CCL3L* (LRT m_0 v. m_1 = 11.6; $p < 0.001$; Table 1; Figure 2.4A).

Population substructure is a potential confounding variable for our analysis, since it has previously been shown that Chinese origin animals tend to exhibit slower progression rates post-infection than Indian-origin animals^[2-5]. In order to address this issue, we first validated population assignments of all individuals in our sample by genotyping 53 unlinked microsatellites and analyzing the data using the Bayesian clustering algorithm STRUCTURE^[21] and Principle Component Analysis (see Additional Methods; Figures 2.5 & 2.6).

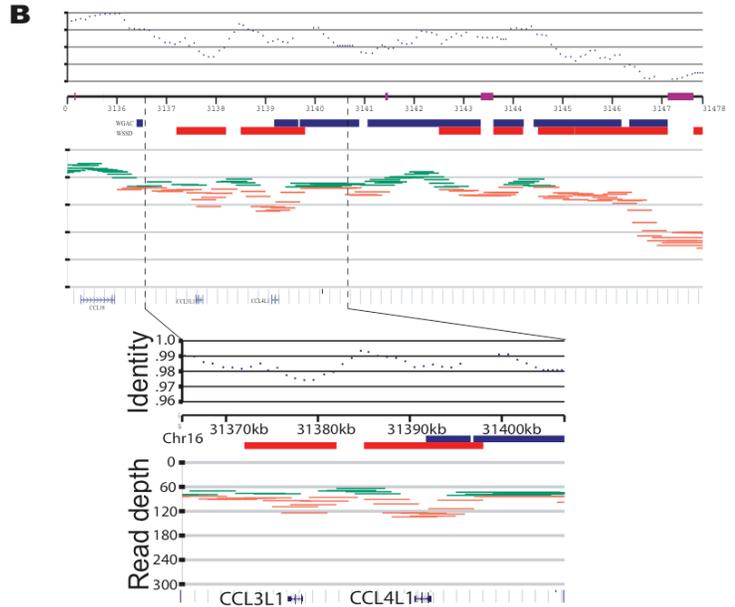
Figure 2.2: Calibration and Verification of rtPCR copy number. A) Metaphase FISH image of A431 cell line confirming diploid copy number of two *CCL3L1* genes (Note: Therefore using our assay we consider the A431 cell line to have a diploid copy number of four genes since it also contains two copies of *CCL3*). B) Whole-genome shotgun read depth analysis showing estimation of *CCL3L* copy number in the rhesus macaque genome donor as 6 copies per diploid genome based on rheMac2 assembly. Green and orange lines denote shotgun reads aligned to *CCL3L* region of the January 2006 assembly of the rhesus macaque reference genome with orange lines showing those that likely represent regions of duplications based on the read-depth analysis (See Methods). C) Interphase FISH image of the MMU2 9133 rhesus macaque cell line, which has an estimated diploid copy number of 10 copies of *CCL3L*. D) Validation of rtPCR estimates of *CCL3L* copy number. Black dots represent rtPCR copy number estimates for the A431 human cell line, the rhesus genome donor, and MMU2 9133 rhesus cell line. Red dots represent an independent estimate of copy number for all three samples based on either FISH or WGS analysis. See supplemental material for additional information regarding the copy number estimation methods.



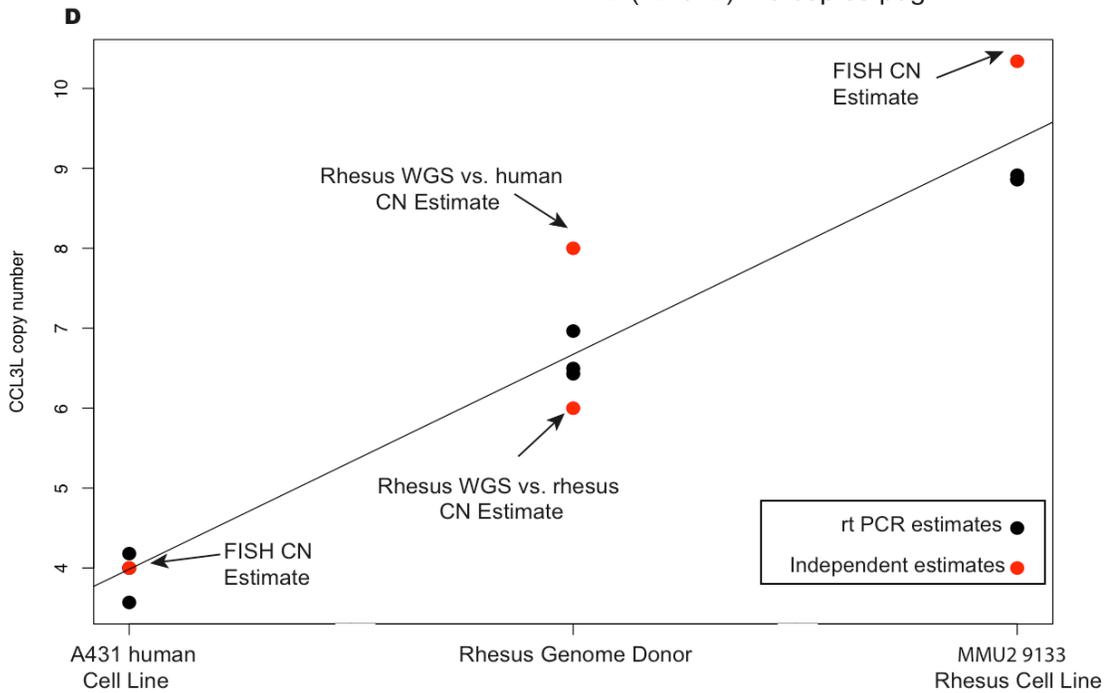
Metaphase FISH for CCL3L1 on human cell line A431



Interphase FISH for CCL3L on rhesus cell line MMU2 9133



$$2 * (124 / 42) = 6 \text{ copies pdg}$$



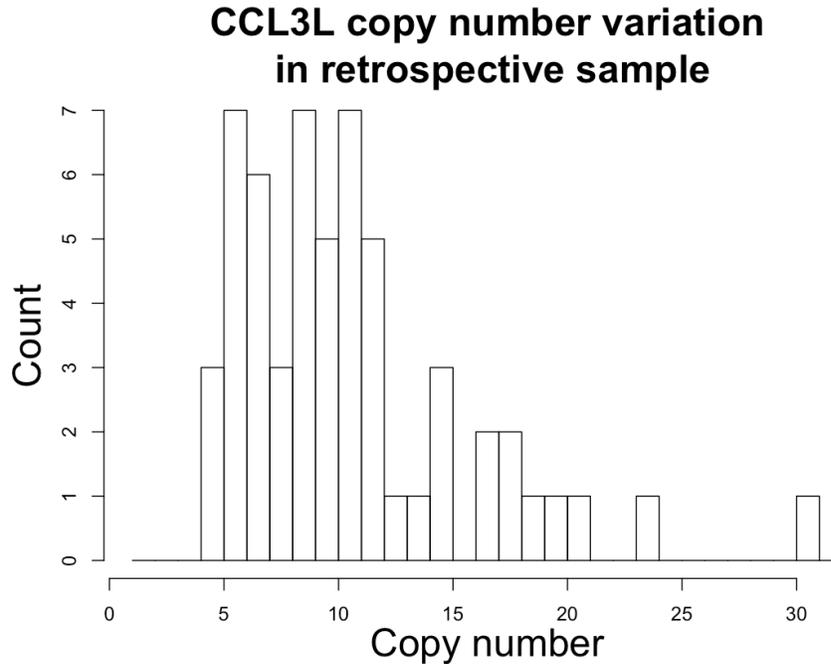


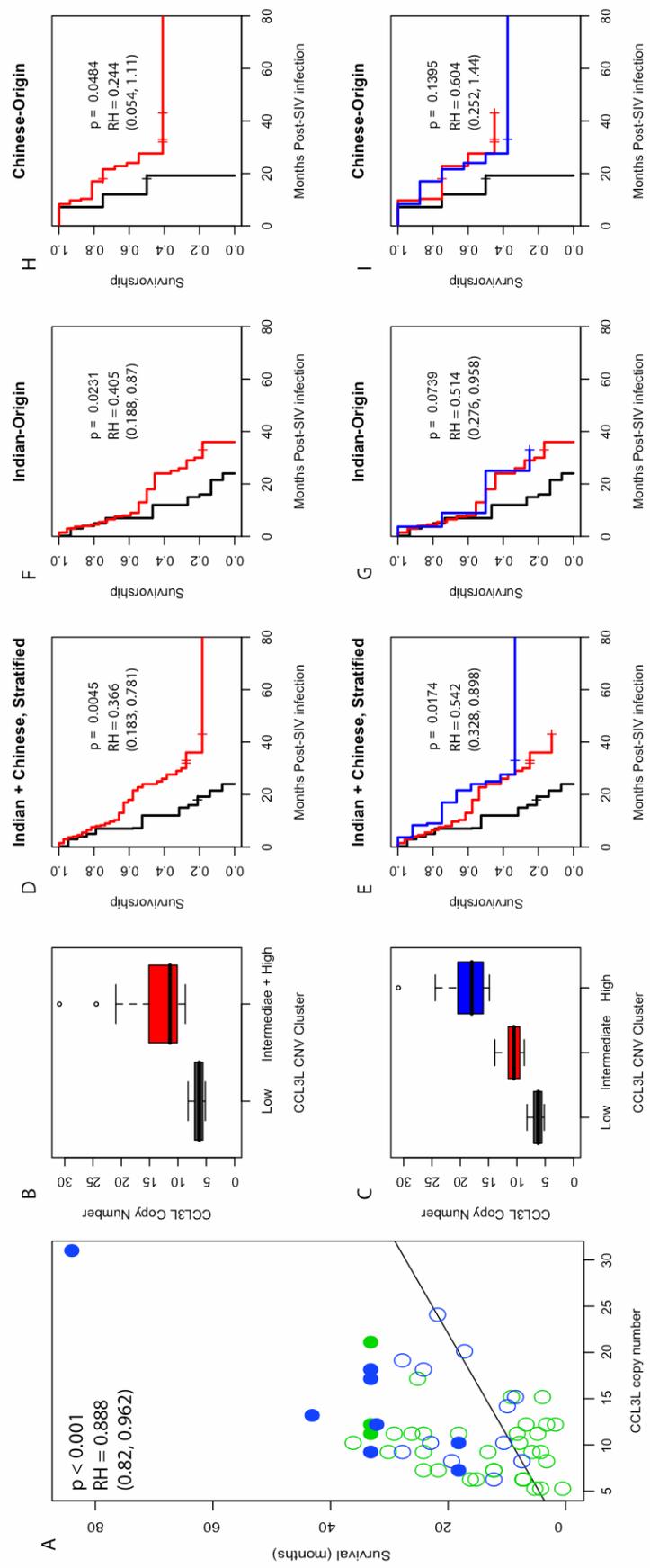
Figure 2.3 Histogram of copy number estimates. Histogram of the rtPCR estimated copy number of *CCL3-like* genes for the 57 retrospective samples of rhesus macaque. The histogram shows a large range of copy numbers found in this sample with copy number estimates from 5 to 31 copies per diploid genome.

Both analyses clearly suggest two (and only two) sub-populations in our data with no evidence of admixture. We also calculated Queller-Goodnight²²⁻²³ estimates of genetic relatedness from the microsatellite data and found only low levels of cryptic relatedness within both populations (see Additional Methods; Figure 2.7). The finding that there is some level of relatedness is expected given that the animals used in our study were sampled from US colonies, however, genomic control analysis of the microsatellite data suggests that these low levels of cryptic relatedness do not markedly affect our *p*-value estimates (see Additional Methods; Figure 2.8).

Table 2.2. Likelihood ratio test statistics for analysis of multiple variables contributing to survivorship based on Cox proportional hazard model. The test statistics are asymptotically χ^2 distributed.

Population	Model	Log-likelihood	R ²	Model Comparison	LRT statistic	p-value
Combined (n = 57)	m_0 : No covariates	-155.9131	--	--	--	--
	m_1 : <i>CCL3L</i> copy number	-150.1400	18.3%	M1 vs. M0 (df = 1)	11.6	<0.0007
	m_2 : Population of origin	-151.7286	13.7%	M2 vs. M0 (df = 1)	8.37	0.0038
Indian-only (n = 37)	m_3 : <i>CCL3L</i> copy number + Population of origin	-148.5120	22.9%	M3 vs. M0 (df = 2)	14.8	0.0006
				M3 vs. M1 (df = 1)	3.25	0.0710
				M3 vs. M2 (df = 1)	6.43	0.0110
Chinese-only (n = 20)	m_0 : No covariates	-96.15256	--	--	--	--
	m_1 : <i>CCL3L</i> copy number	-93.01357	15.6%	M1 vs. M0 (df = 1)	6.28	0.0122
	m_0 : No covariates	-30.55236	--	--	--	--
	m_1 : <i>CCL3L</i> copy number	-30.07515	4.7%	M1 vs. M0 (df = 1)	0.95	0.3290

Figure 2.4: Rhesus macaque survival analysis. A) Scatter plot of post SIV infection survival time (or censor time if animal is alive) by *CCL3L* copy number. Blue dots represent Chinese origin rhesus macaques while green dots represent Indian origin. Filled in dots represent animals still alive at time of sampling. Fitted regression curve, p -value and relative-hazard (RH) from Cox proportional hazard model (model 1 in text). B-C) Boxplots of *CCL3L* copy-number defining “low” copy number to be fewer than or equal to 8 copies per diploid genome, “intermediate” to be 9 and 14, and “high” to be more than 14 copies or low vs. intermediate + high. D-I) Estimated Kaplan-Meier survival curve for SIV-infected macaques with time measured from date of infection. The black curve represents “low”, the red curve “intermediate” or “intermediate + high”, and the blue curve “high” copy number for KM curves based on all animals (D,E), Indian-origin only (F,G), and Chinese-origin only (H,I). The p -values correspond to Harrington-Fleming tests of equality for survivorship curve using $r = 0$ which is equivalent to a log-rank or Mantel-Haenszel test. Relative-hazard (RH) for equivalent Cox proportional hazard model are also presented.



Once population assignments for all individuals had been confirmed, we considered several statistical models for the progression data that included population-of-origin as a potential covariate. When considered alone, we found that population-of-origin impacts survivorship with Indian-origin, correlating with increased rate of progression to simian-AIDS as previously reported (LRT m_0 v. $m_2 = 8.37$; $p < 0.01$; Table 1). However, once *CCL3L* is included in the model, population-of-origin makes only a marginally significant improvement (LRT m_1 v. $m_3 = 3.25$; $p = 0.071$; Table 2.2). This analysis suggests that *CCL3L* is the predominant factor impacting survivorship differences among individuals, and predicts that differences in the distribution of *CCL3L* copy number among Indian and Chinese populations may explain the population-level differences in survivorship.

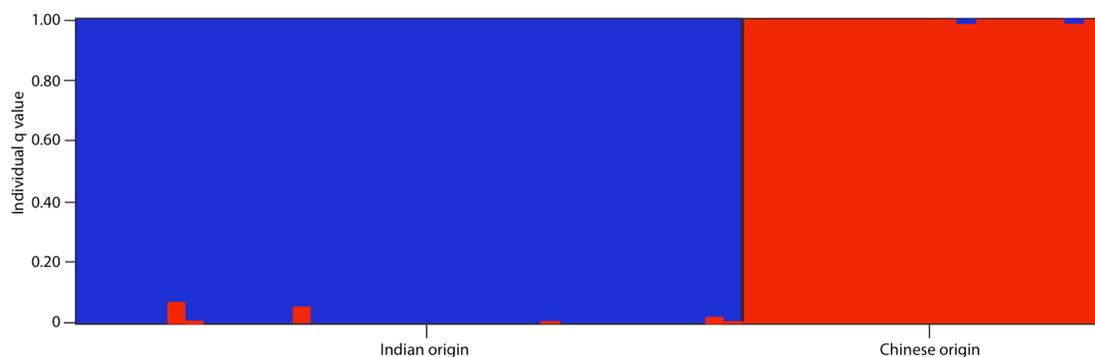


Figure 2.5: Structure results of the retrospective individuals from the 53 microsatellite loci sorted by assumed population. Blue are Indian origin animals and red are Chinese origin animals.

We further tested the impact of population substructure by repeating our analysis using only Indian origin rhesus macaques. (The sample size and proportion of censored data in the Chinese origin sample rendered the power of the test too low to detect a significant result; see Additional Methods, Figure 2.9). We found that including *CCL3L* CNV in the model explains a significant proportion of the survival time variation among Indian origin macaques alone

($R^2 = 15.6\%$; $p = 0.0122$), and that the estimated effect size of *CCL3L* copy-number variation (b) on survivorship is highly comparable across subsets of the data (see Table 2.3 and 95% confidence intervals for $\exp(b)$). This observation suggests that *CCL3L* CNV has a similar effect across both populations, whereby each copy of *CCL3L* decreases the baseline risk by a constant factor of approximately $\exp(b) = 0.907$ relative to the mean copy number (e.g., having 16 copies decreases the hazard by a factor of $0.907^5 = 0.61$, and having 8 copies increases the hazard by a factor $0.907^{-3} = 1.34$).

Further support for the protective effects of increased *CCL3-like* gene copy number is provided by Harrington-Fleming tests of equality for Kaplan-Meier survival curves²⁴. Comparisons of the survival curves across all observed *CCL3L* copy number levels clearly reject equality, whether analyzing all individuals together ($\chi^2 = 51.3$; $p < 0.001$, $df = 17$) or stratifying by population of origin ($\chi^2 = 48.1$; $p < 0.001$, $df = 17$). Additionally, we considered dividing the data into qualitative copy-number categories: “low” having less than 9 *CCL3L* copies pdg, “intermediate” having 9-14 *CCL3L* copies pdg, and “high” having greater than 14 *CCL3L* copies pdg. We also considered a two-class classification that combined the “intermediate” and “high” copy number classes into a single class. Overall, we observe a highly significant difference in survivorship between *CCL3L* copy classes in the combined data stratified by origin ($p = 0.0045$ for two categories and $p = 0.0174$ for three categories; see Figure 2.4B-D).

Figure 2.6: PCA results for the retrospective sample. Blue are Indian origin and red are Chinese origin. A) Box-plot of PC1 values. B) Bi-plot of PC1 vs. PC2 showing distinct clustering of animals into proper sub-populations.

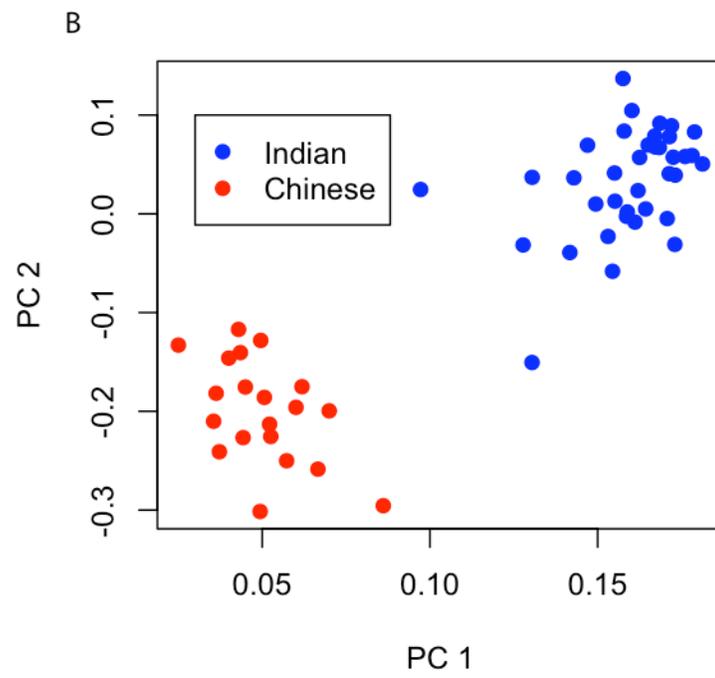
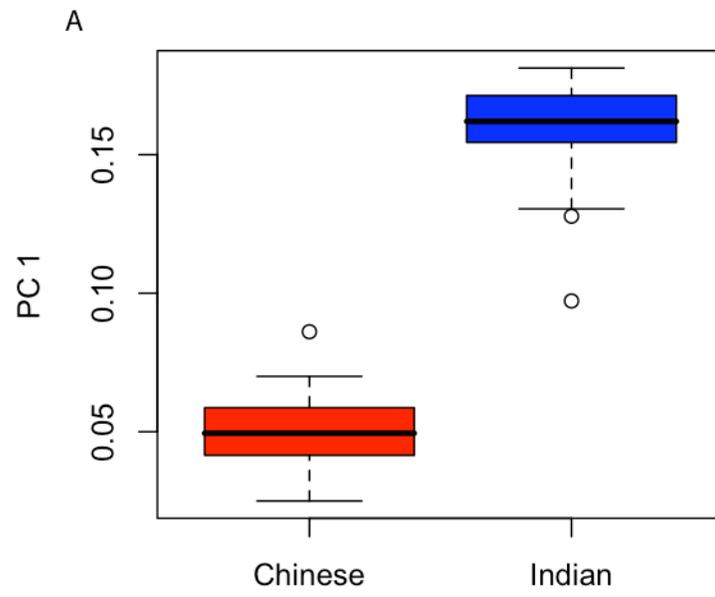
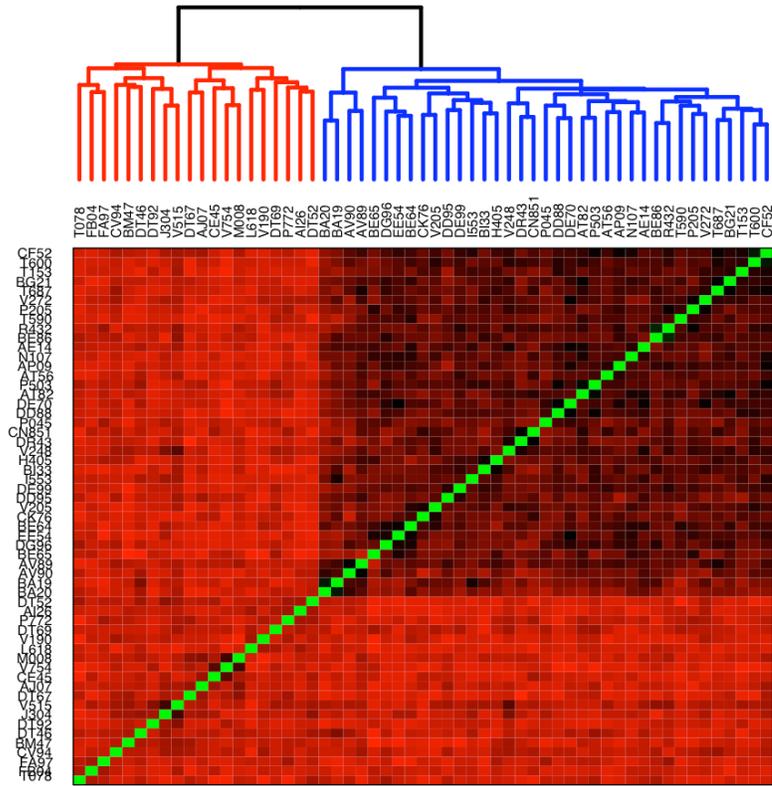
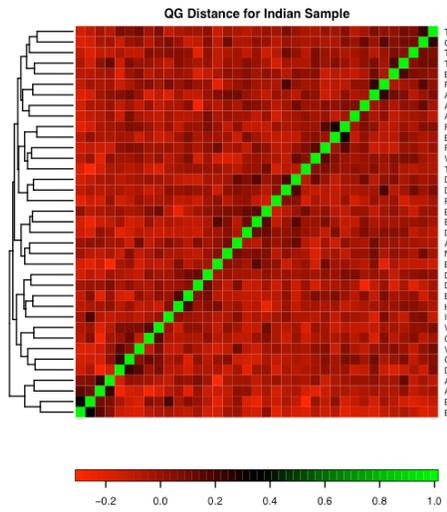


Figure 2.7: Heat plots summarizing genetic relatedness in the sample based on 53 unlinked microsatellite loci. A) Pearson product-moment correlation of genotypic state for all individuals in the sample; B) Queller-Goodnight r distance between pairs of individuals in the Indian-origin sample; C) QG distances for individuals in the Chinese-origin sample.

A



B



C

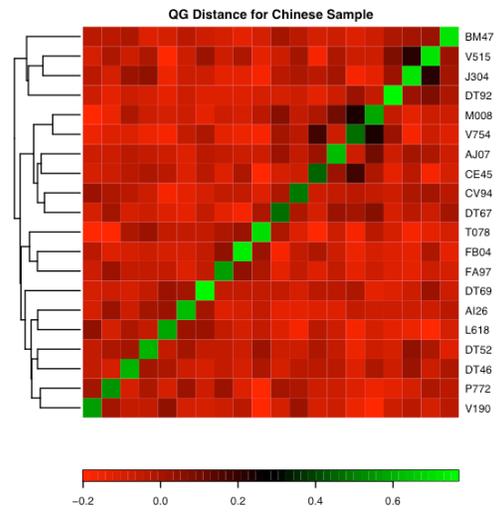
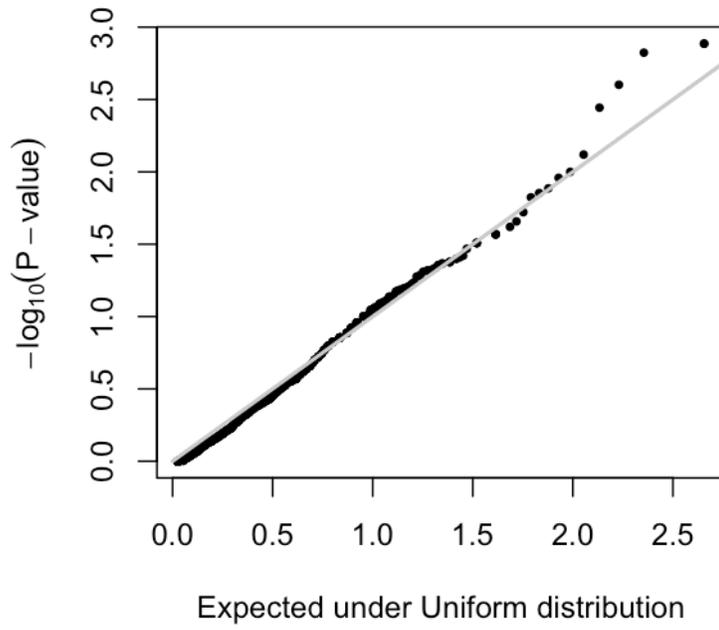


Figure 2.8: Test of genome-wide significance of *CCL3L* CNV A) Quantile-Quantile plot of the empirical p -value distribution from the 53 unlinked microsatellites versus that expected under a uniform distribution. B) Histogram of the $-\log_{10}$ p -values from the microsatellite data with arrow showing the position of the p -value for the association with \log_2 *CCL3L* copy number and survival.

A Empirical p-value distribution



B

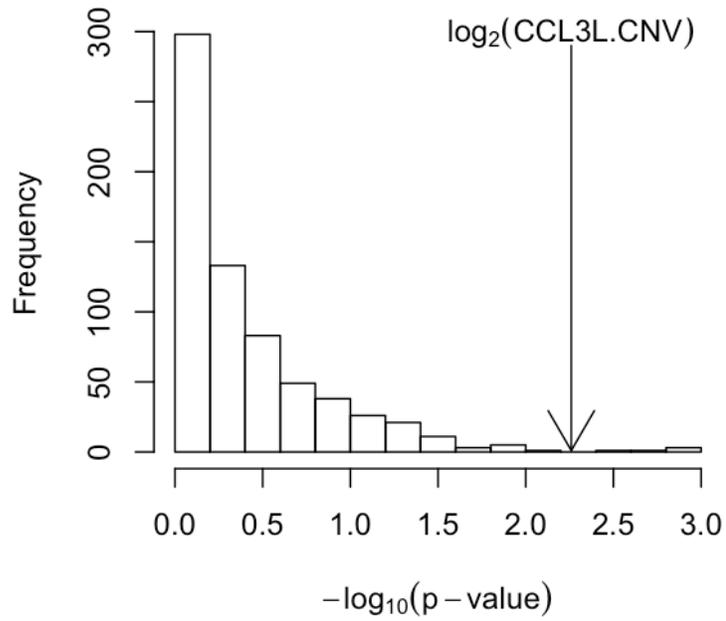


Table 2.3. Regression coefficient estimates (b), standard errors on the regression coefficient estimates, confidence intervals, and significance for terms in the Cox proportional hazard models summarized in Table 1. “Variable” refers to a particular term in the regression model (i.e., CCL3L copy number, log of CCL3L copy number, or population of origin), “Data” refers to which subset of the data is considered (i.e., Combined = Indian + Chinese origin animals or Indian animals alone), and “Other factors in the model” refer to whether the regression coefficient is estimated alone or in the presence of other terms.

Variable	Data	Other factors in model	β	RH	se (β)	95% CI on Exp(β)	p-value
<i>CCL3L</i>	Combined		-0.119	0.888	0.04	(0.82, 0.96)	0.0038
	Indian-only		-0.149	0.861	0.07	(0.76, 0.98)	0.0260
	Combined	Origin	-0.097	0.907	0.04	(0.84, 0.99)	0.0220
$\log_2(\text{CCL3L})$	Combined		-1.10	0.333	0.32	(0.18, 0.62)	0.0006
	Indian-only		-1.12	0.327	0.44	(0.14, 0.77)	0.0110
	Combined	Origin	-0.93	0.393	0.34	(0.20, 0.76)	0.0055
Origin	Combined		-0.92	0.398	0.34	(0.21, 0.77)	0.0064
	Indian-only	<i>CCL3L</i>	-0.61	0.54	0.35	(0.27, 1.08)	0.0830
	Combined	$\log_2(\text{CCL3L})$	-0.58	0.56	0.35	(0.20, 0.76)	0.1000

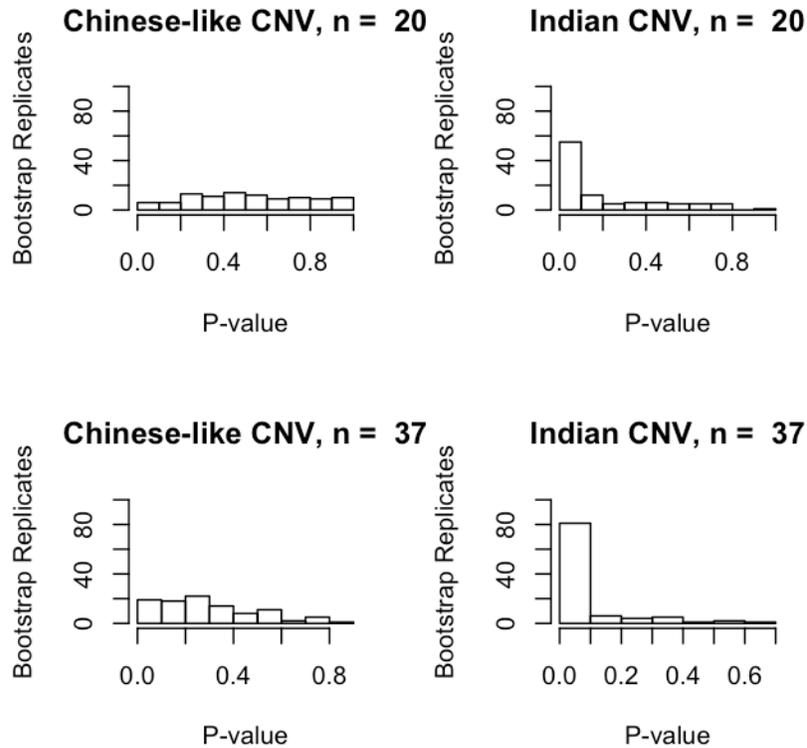


Figure 2.9. Bootstrap simulations to assess power of Cox proportional hazard regression of survivorship on *CCL3L* copy number applied to each population separately.

Likewise, if we consider survivorship curves within each population separately, a significant difference is observed between animals with low copy number relative to those with intermediate or high copy number ($p = 0.0231$ for Indian; $p = 0.0484$ for Chinese). These results taken together suggest that it is low *CCL3L* copy number, in particular, that is correlated with increased rate of progression.

Next we investigated whether differences in the distribution of *CCL3L* copy number alleles between populations could explain the previously

reported slower simian-AIDS progression rates of Chinese origin animals^[2-5]. That is, given the association between higher *CCL3L* copy number and slower progression, we would expect Indian origin macaques to have, on average, lower *CCL3L* copy numbers as compared with Chinese origin macaques. Within the samples used for the retrospective study, animals designated as Indian-origin did, in fact, have a significantly lower mean copy number (mean = 9.51, sd = 3.57, s.e.m = 0.587), than those designated as Chinese-origin (mean 13.90, sd = 6.41, s.e.m = 1.43) as measured by a Mann-Whitney *U* test using either relative copy number estimates from rtPCR ($p = 0.0088$) or binned and rounded CNV calls ($p = 0.0077$; see also Figure 2.10A). We also assayed *CCL3L* CNV in an independent panel of SIV-free Indian origin and Chinese origin rhesus macaques to ensure that the relationship between origin and *CCL3L* was not a peculiar artifact of the animals we utilized from the SIV vaccine trials. This independent panel included 15 wild-caught Chinese-origin macaque samples collected as part of the Rhesus Macaque Genome project^[19] and 16 colony-born Indian origin^[19] macaques provided by Yerkes National Primate Center. In this second panel, we found an even higher difference in *CCL3L* CNV between the two populations ($p < 9 \times 10^{-7}$ Mann-Whitney *U* test; also see Figure 2.10A). Chinese origin animals had, on average, twice as many copies of *CCL3L* as Indian origin animals (Chinese origin mean = 17.6, s.d. = 3.56, s.e.m = 0.91; Indian origin mean = 9.41, s.d = 3.4, s.e.m = 0.91), consistent with the average slower progression rates of Chinese vs. Indian-origin animals.

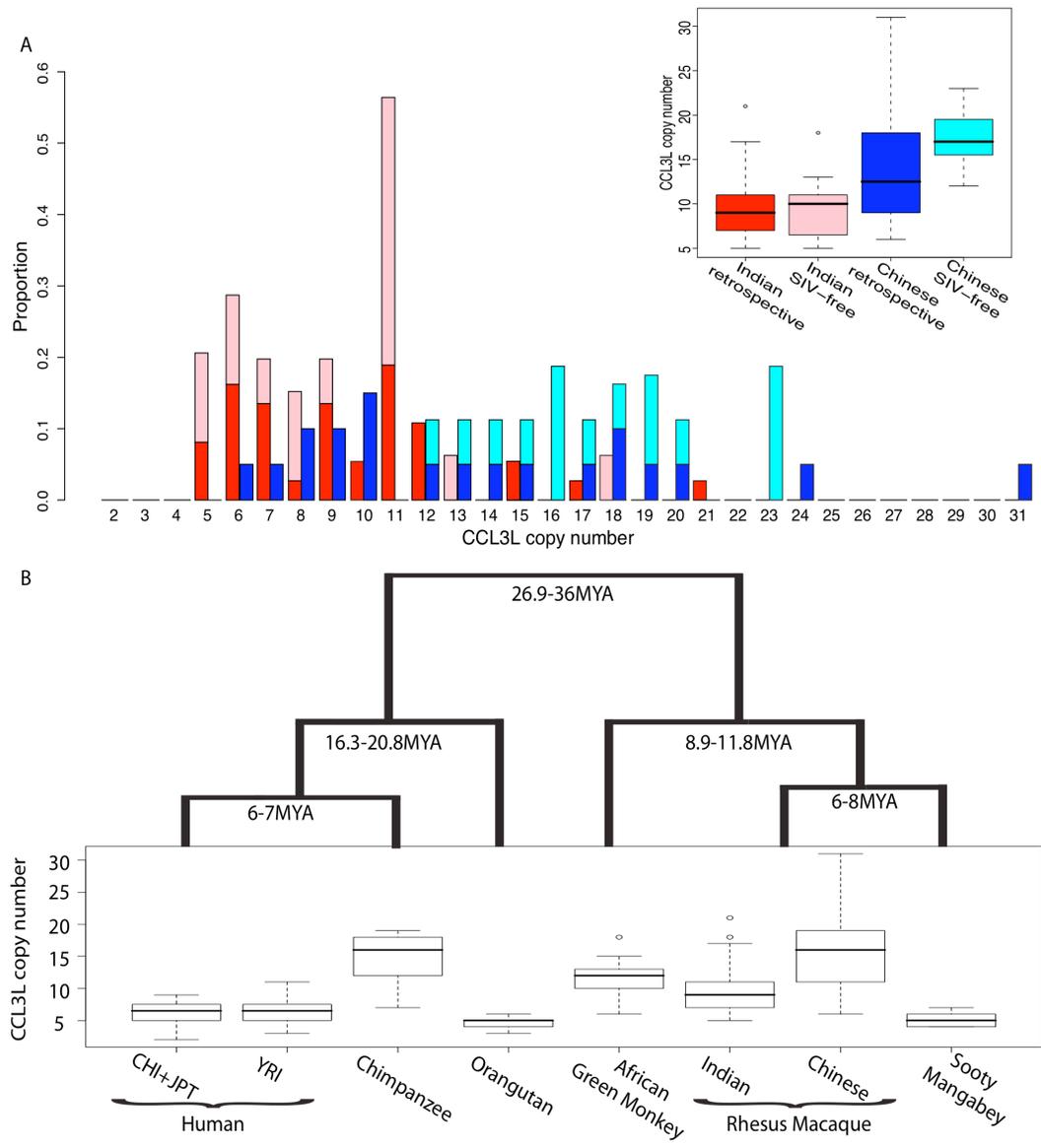
2.4 Discussion

Analysis of the retrospective data provides strong support for the hypothesis that *CCL3L* CNV affects individual level SIV progression rates in

rhesus macaques. This is particularly evident in the Indian origin rhesus macaque where lower copy numbers of *CCL3L* are more common, putatively leading to an overall increase in progression rates in this population. Due to the limited power in our analysis of the Chinese-only sample we recommend further studies to confirm the role of *CCL3L* in this population. To our knowledge, the current study provides the first example of an association between copy number variation and disease in a non-human primate. These results broaden our understanding of the role copy number variation in disease susceptibility and point to the importance of utilizing methods, which allow for detecting this type of variation in genome-wide scans of disease association.

When taken together with the results of the retrospective progression study, the population level analysis suggests that differences in the distribution of *CCL3L* copy number may explain a large portion of the differences in progression rates between Indian and Chinese origin macaques. Using the results of the Cox proportional hazard model, and the observed *CCL3L* distribution between subpopulations, we have generated predictions for expected survivorship at different levels of *CCL3L* copy-number variation and population-of-origin designation (provided in Addition Methods Figure 2.11). These calculations may prove useful in the efficient design of vaccine trials. For example, we predict less than 15-20% of Indian or Chinese-origin animals with six or fewer copies of *CCL3L* will survive past 24 months post-SIV infection. In contrast, the vast majority of animals with 25 or more copies are expected to survive well past 36 months, regardless of whether they are of Indian or Chinese origin.

Figure 2.10: Population and species level copy number variation. A) Histograms and boxplots of *CCL3L* copy number distribution among the $n = 57$ animals used in the retrospective study as well as for a sample SIV-free Indian origin ($n = 16$) and Chinese origin rhesus macaques ($n = 15$). Red and light-red bars indicate Indian origin for the SIV and SIV-free populations, and blue and light-blue bars indicate the analogous for Chinese origin animals. B) Box plot of copy number variation for 6 primate species: Human, Chimpanzee (*Pan troglodytes*), Orangutan (*Pongo pygmaeus*), Rhesus macaque (*Macaca mulatta*), African green monkey (*Cercocebus aethiops*), and Sooty mangabey (*Chlorocebus atys*). Whiskers indicate the upper and lower quartile with dots showing outliers. Estimates of species divergence times are from reference [30].



In this context, it is important to note that the determination of absolute copy numbers using rtPCR completely depends on the quality of the reference. Moreover, determination of absolute high copy numbers is less accurate than low copy number, because noise accumulates during the progression of the amplification reaction. That said, since our absolute copy number results are based on two validated references, it is likely that they are accurate. In addition, importantly, we note that the conclusions of this study are not contingent on obtaining accurate *absolute* copy numbers for each sample. Rather, our conclusions are based on the *relative* copy number of *CCL3L* between samples, a measure that qualitatively is not sensitive to the specific reference used. Specifically, our results are robust with respect to how *CCL3L* copy number is defined. In other words, if we consider \log_2 of *CCL3L* copy number, or relative estimates of *CCL3L* copy numbers instead of absolute copy numbers, our conclusions are unchanged (Table 2.4).

Our findings, together with previous findings^[13-17] suggest that *CCL3L* copy number variation is a shared genetic mechanism of slower disease progression between humans and macaques. This result is surprising given the long evolutionary time separating the two species. Population genetic theory suggests that little genetic variation currently in the human population should be shared ancestrally with rhesus macaques, so there is no *a priori* reason to suspect a shared mechanism due to a common polymorphism.

Figure 2.11: Predicted Kaplan-Meier survival curves based on Cox-Proportional hazard model of post-SIV survivorship including *CCL3L* copy number and population-of-origin as covariates. Dashed lines indicate 95% prediction intervals based on application of the function `survfit` in the `survival` R package.

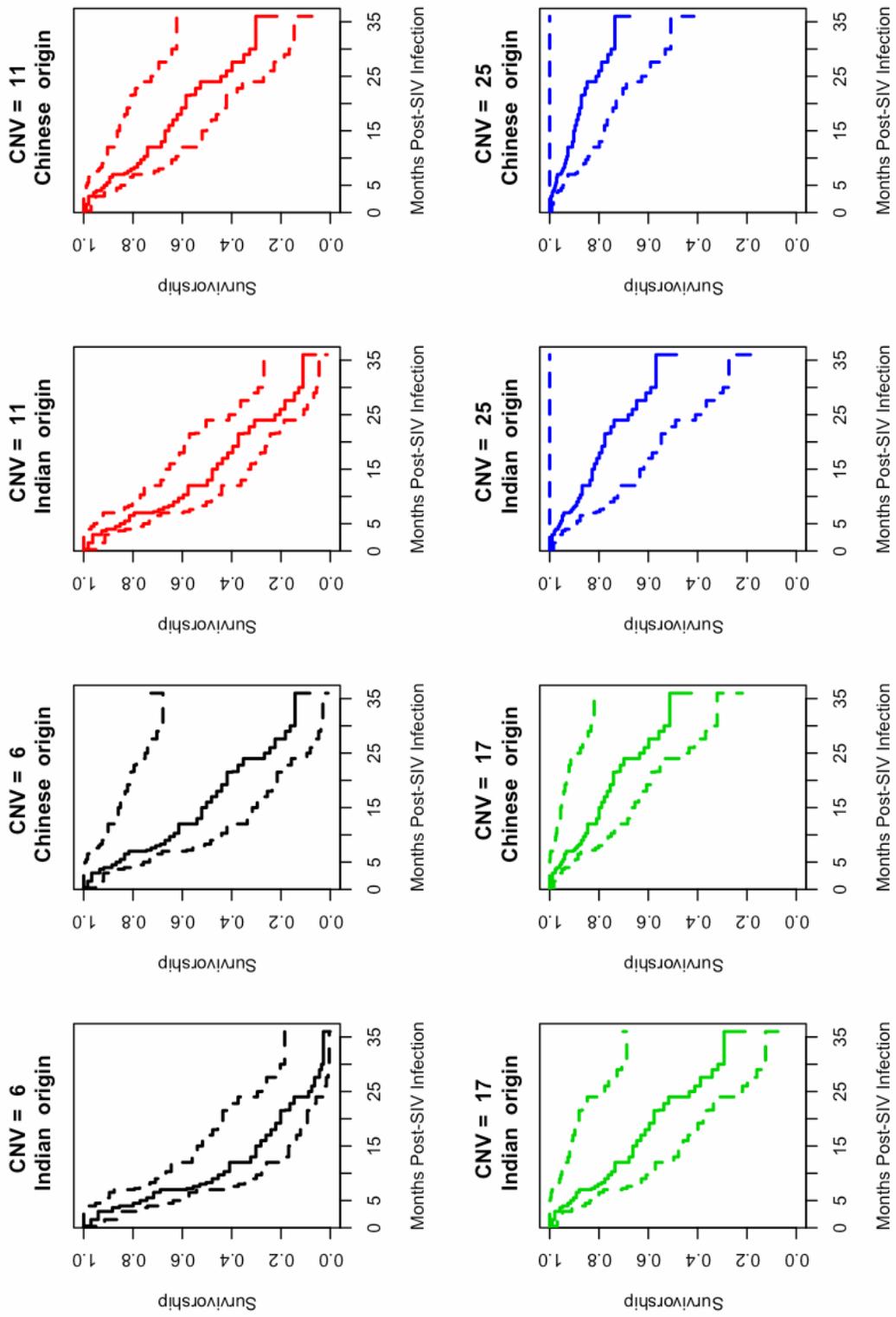


Table 2.4: Summary statistics for CCL3L copy number distribution among primate species and populations

Population	<i>n</i>	Mean	SD	Median
Human (Yoruban [YRI])	8	6.125	2.17	6.5
Human (Chinese [CHI]+ Japanese [JPT])	8	6.5	2.45	6.5
Chimpanzee	12	14.58	4.27	16
Orangutan	7	4.57	0.976	5
Rhesus macaque (Chinese – SIV free)	15	17.6	3.56	17
Rhesus macaque (Chinese – SIV retrospective)	20	13.9	6.40	12.5
Rhesus macaque (Indian – SIV free)	16	9.44	3.4	10
Rhesus macaque (Indian – SIV retrospective)	37	9.41	3.56	9.0
African green monkey	12	11.58	3.15	12
Sooty Mangabey	10	5.2	1.32	5

To determine if *CCL3L* CNV is indeed shared ancestrally, we examined *CCL3L* copy number in five other species: African green monkey [AGM] (*Chlorocebus atheops*, $n = 12$), sooty mangabey [SM] (*Cercocebus atys*, $n = 10$), orangutan [PP] (*Pongo pygmaeus*, $n = 7$), chimpanzee [PT] (*Pan troglodytes*, $n = 12$), and humans [HS] (8 Yoruban, 4 Chinese and 4 Japanese from the phase I HapMap set). Our results confirm a previous observation^[13] and reveal the presence of extensive variability in *CCL3L* copy number in all primate species examined (Table 2.2; Figure 2.10B), suggesting that *CCL3L* CNV has likely been segregating in Old World monkeys and apes for at least 25 million years through recurrent duplication and deletion of the locally unstable genomic segments containing the *CCL3L* genes.

In summary, our findings further support the hypothesis that *CCL3L* gene copy number variation is an important host factor for explaining variation in HIV/SIV progression rates^[13-17]. Our results also provide an example of a common mechanism of increased survival time after infection with HIV or SIV in humans and another primate species respectively. There are two immediate predictions from our observations. First, stratifying by *CCL3-like*

gene copy number in macaque vaccine studies will allow researchers to remove *CCL3L* as a confounding effect, thereby increasing the power of vaccine trials. Second, based on our observations, we suggest that rhesus macaque is a valuable model organism for further studies of the specific mechanism by which *CCL3-like* gene copy number affects rates of HIV progression in humans.

Finally, we acknowledge that an important caveat of our work is that we have used a candidate gene approach in our analysis of the association of genetic variation with simian-AIDS progression. We have addressed this to some extent by conducting replicate association analyses with the 53 genome-wide, unlinked microsatellite loci (see Additional Methods; Figure 2.8). We find that copy number variation at the *CCL3L* locus falls in the 1% tail of this distribution (after accounting for population substructure) and is therefore, likely a true positive. It is important to remember that there are many other host factors aside from chemokines and their receptors known to influence HIV susceptibility and pathogenesis in humans^[24,25]. We believe these other factors should be characterized in rhesus along with discovery of rhesus-specific genetic variation before conclusions can be drawn on the relative importance of shared versus species-specific factors influencing retroviral susceptibility and disease progression.

2.5 Material and Methods

Retrospective progression samples

The rhesus macaques used in the retrospective analysis were all inoculated with SIVmac as part of previous SIV research programs at the Tulane National Primate Research Center. All macaques were infected with

SIVmac239 or SIVmac251 with SIV inoculum given under a standard protocol and at similar mid-level dose. Doses in this range and the strain used have been previously shown not to affect the outcome of disease course^[26].

All animals used in our study were euthanized under the same set of guidelines if they did not remain healthy after infection. Specifically, euthanasia was carried out if life threatening clinical conditions indicated that the life expectancy of the animal was less than 7 days. Following euthanasia, a necropsy was performed, and animals were only included in the current study if the necropsy confirmed SIV as the underlying cause of the clinical state. Under these protocols, the time of euthanasia will give a reasonable approximation to both time to progression to simian-AIDS and survival time, as the presence of AIDS defining illnesses met the criterion for euthanasia. (See Table 2.1 for clinical findings of necropsy and Additional Methods).

Additional Primate Samples

DNA extractions from the uninfected Chinese origin rhesus samples were obtained from the Rhesus macaque genome consortium. The chimpanzee, orangutan, sooty mangabey, and uninfected Indian origin rhesus macaque DNA samples were obtained from the Yerkes National Primate Research Center and the African green monkey DNA samples were obtained from the University of California Los Angeles.

Real-time PCR *CCL3L* copy number estimation

CCL3L gene copy number was determined using real-time Quantitative PCR (rtPCR) on a 7900HT Fast Real-Time PCR System (Applied Biosystems Inc.) with the JumpStart *Taq* ReadyMix (SIGMA) and *TaqMan* probes. The PCR included 18 ng total genomic DNA. Cycling conditions were: initial denaturation at 94°C for 2 min; followed by 40 cycles of 15 sec denaturation at

94°C and 1 minute annealing/extension at 60°C. The *Stat6* gene, found to be present in a single copy in rhesus macaque, chimpanzee and human reference genomes, was used as the internal control. Oligonucleotide sequences used for *CCL3L* were: Forward: 5' CCAGTGCTTAACCTTCCTCC 3', Reverse: 5' TCAGGCACTCAGCTCCAGGT 3', Probe: 5' AGGCCGGCAGGTCTGTGCTGACC 3'. For *Stat6*, sequences were: Forward: 5' CCAGATGCCTACCATGGTGC 3', Reverse: 5' CCATCTGCACAGACCACTCC 3', Probe: 5' CTGATTCCCTCCATGAGCATGCAGCTT 3'. This primer set does not distinguish between *CCL3* and the *CCL3-like* gene paralogs, as there are not sufficient fixed differences between these paralogs in rhesus macaque to design a specific assay. It is also unknown whether any pseudogenized copies of *CCL3L* genes exist in the rhesus macaque populations. As such, we here refer to *CCL3* and its paralogs as *CCL3L*. PCR results were analyzed using SDS v2.2.1 software package (Applied Biosystems Inc.). We performed rtPCR for each individual in triplicate and determined the normalized *relative* copy number by generating a standard curve and then normalizing across samples by the results of the *Stat6* control gene and dividing the value obtained by one of the reference individuals.

Analysis of *CCL3L* copy number based on reference samples

To estimate the absolute *CCL3L* copy number for each sample based on the rtPCR results described above, we used two reference samples: the A431 human cell line and the rhesus genome donor individual. The A431 cell line was chosen as it has previously been shown to have two copies of *CCL3L1* and two copies of *CCL3* per diploid genome (pdg)^[9], for a total copy number of four *CCL3L* using the rtPCR assay described above. To confirm the

CCL3L1 copy number of the particular A431 cell line culture used here, we performed florescent in situ hybridization (FISH) of metaphase chromosomes using the human fosmid probes WIBR2-3688L07 (*CCL3L1* specific; green spots on Figure 2.2A) and WI2-653M1 (chr. 17 single copy control; red spots on Figure 2.2A). Visualization of the FISH assay clearly shows that this cell line extract had 2 copies of *CCL3L1* pdg.

The second reference sample was the rhesus macaque genome donor sample. Copy number of the *CCL3L* locus for this sample was determined using whole genome shotgun (WGS) read depth analysis^[19,28]. All fragments of minimum 150 bp of non-repeat masked sequence were aligned to the to the macaque *CCL3L1* locus with a 95% identity threshold. We compared the average depth of WGS sequence coverage for unique (not-duplicated) sequence in 5 kb windows with the depth of coverage to the *CCL3L1* locus to estimate copy-number of the locus (Fig 2.2B). The experiment was repeated, using the human *CCL3L* locus as a reference with an 88% identity threshold (results not shown). From these analyses, we predicted the *CCL3L* copy number for the genome donor macaque to be 6-8 copies of *CCL3L* pdg depending on whether the rhesus or human genome is used for alignment. The difference in estimated copy number between the alignment to the rhesus genome and that of the human genome is likely due to alignment of non-*CCL3L* genes. Due to this, alignment to the rhesus genome is likely a better predictor of *CCL3L* copy number for this individual because it is less likely to include non-*CCL3* gene paralogs.

We determined the absolute *CCL3L* copy number in each sample by comparing rtPCR results between samples and the references. Specifically, the normalized rtPCR values were averaged across the three replicates

divided by the rtPCR averaged and multiplied by 4 (the diploid copy number of the A431 cell line including *CCL3L1* and *CCL3*) or 7 (the average diploid copy number of the rhesus macaque donor individual). The resulting number was then rounded to the nearest integer value to estimate absolute copy number. In Figure 2.1, we report the calibration curves for the A431 reference samples and demarcation of inferred copy number pdg for each sample. All statistical analyses were conducted using the rounded as well as the raw values.

Confirmation of rtPCR *CCL3L* copy number estimate

To confirm that the rtPCR absolute copy number estimates were accurate we estimated *CCL3L* copy number for an additional rhesus macaque cell line using both rtPCR and interphase FISH (Figure 2.2C). The rtPCR estimated diploid copy number for this macaque cell line is 9 using either reference sample. The estimated *CCL3L* copy number from the FISH experiment is 10.34 ± 3.00 (mean \pm standard error based on 54 replicate FISH experiments). The slight discrepancy between the rtPCR and FISH is likely due to the fact that the FISH probe used contains other, known structural variants which show higher copy number in the macaque reference genome (visible in WGS read depth analysis see Figure 2.2C). As well, the proximity of the *CCL3L* gene copies renders it difficult to distinguish distinct copies in some of the FISH images.

Primers in additional species

The same rtPCR primers and probe were used in all primate species. These primers are not specific to the other species and differences in both the chimpanzee and human reference priming sequences were observed. (No reference sequences are available for the orangutan, sooty mangabey or the African green monkey on which to design species-specific probes). While this

may lead to slight biases in the determination of the absolute copy number for any particular individual or species, it does not effect the overall conclusions of the study that all species surveyed show population variation in copy numbers.

Statistical analysis

All statistical analysis was conducted using the R statistics package.

Significance of copy number differences between Indian-origin and Chinese-origin populations of SIV and non-SIV infected rhesus macaque was evaluated using a Mann-Whitney U test. Survival analyses of the SIV infected macaque data were conducted using the survival package in R.

The Cox proportional hazard model was chosen, as it is a flexible semi-parametric regression model that accounts censored data. Let $i = 1 \dots n$ index individuals and $j = 1 \dots p$ index variables of the regression model. The Cox proportional hazard rate of individual i at time t has the form:

$$h_i(t) = h_0(t) \exp\left(\sum_{j=1}^p \beta_j x_{ij}\right)$$

where $h_0(t)$ is the base line hazard function, the x_{ij} 's for $j = 1 \dots p$ are the covariates for individual i , and the β_j 's are regression coefficients. An underlying assumption of this model is that the covariates act additively on the log of the hazard function and that the log hazard function changes linearly with the β terms. These are referred to as the proportionality assumptions. We tested this assumption using the method proposed by Grambsch and Therneau^[29] as implemented in the survival package in R and found that the assumption holds for these data. It is important to note that there no assumption is made regarding the functional form of base line hazard function $h_0(t)$. The reason for this is that our object of analysis is the *proportional*

hazards among individuals that at time t are independent of h_0 . For example, considering a pair of individuals i and i' , the hazard ratios are:

$$\frac{h_i(t)}{h_{i'}(t)} = \frac{\exp\left(\sum_{j=1}^p \beta_j x_{ij}\right)}{\exp\left(\sum_{j=1}^p \beta_j x_{i'j}\right)}$$

The model parameters $b_1 \dots b_p$ are estimated given the *ranked* observed failure times $y_1 < y_2 < \dots < y_n$ using the partial likelihood method proposed by Cox^[20] as implemented in the `coxph` function in R. Since some data are censored, we introduce n' to denote the number of uncensored observations. The partial likelihood is given by:

$$L(\beta | y_1, \dots, y_n) = \prod_{i=1}^{n'} \frac{\exp\left(\sum_{j=1}^p x_{ij} \beta_j\right)}{\sum_{i'>i}^n \exp\left(\sum_{j=1}^p x_{i'j} \beta_j\right)}$$

Four models are considered; m_0 , which includes no covariates; m_1 , which includes only *CCL3L* copy number as a potential covariate; m_2 , which considers only population-of-origin as a factor, and m_3 , which considers both *CCL3L* copy number and population of origin. To choose among nested regression models for the SIV infected macaque survival data, we used twice the difference in log-likelihood and assessed significance using standard c^2 approximations.

The Harrington and Fleming procedure was used to assess differences among Kaplan-Meier survival curve. This method was also implemented in the

survdiff function of the R survival package. All analysis labeled “stratified” were conducted by including the term strata(origin) in the right hand side of the regression equation where origin is an indicator variable of Chinese-origin (i.e., 1 if Chinese, 0 if Indian). The survfit routine to generate predicted Kaplan-Meier survival curves as a function of *CCL3L* copy number and population-of-origin. All R scripts used for analysis and production of Figures are available from the investigators upon request.

Confirming Primer sequences

Real-time PCR primer and probe sequences were designed against the publicly available rhesus macaque genome sequence. As the sequence is based on the genome of one Indian origin individual we tested the specificity of the primers by sub-cloning and sequencing PCR products from two Chinese and two Indian individuals. To do so, we designed primers that flank the original rtPCR primers, thereby amplifying a product that includes the entire original rtPCR product. We used TA-cloning to clone individual PCR products, and performed touch-down PCR followed by direct sequencing of 60 clones from each individual. The sequences were aligned and sequence differences called, using the Sequencher software (Gene Codes Corp. Ann Arbor, MI). While some polymorphisms were observed between the clones, there were no fixed differences observed between the groups (Table 2.4). Since most of the differences were observed in the Chinese origin individuals, the possible bias of the rtPCR assay is in the direction of underestimating *CCL3L* copy number of Chinese origin individuals, a conservative bias with respect to our conclusions.

Table 2.5. Total number of polymorphic sites found per primer/probe/individual for *CCL3L* rtPCR assay. CH1 and CH2 are two macaque individuals of Chinese origin. IN1 and IN2 are Indian-origin macaques.

Individual	Forward primer	Reverse Primer	Probe
CH1	1	4	4
CH2	1	2	0
IN1	3	1	1
IN2	0	0	0

Analysis of microsatellite data

In order to test for population structure and relatedness between individuals we typed 53 microsatellite loci (developed as part of the rhesus genome map at the Southwest National Primate Research Center; Rogers et al. 2006 <http://www.snprc.org/linkage/index.html>) in each of the 57 previously infected rhesus monkeys (see Table 2.5).

PCR amplifications used 25 ng of genomic DNA as template, plus standard buffers, one unlabelled primer and one fluorescently labeled primer in reactions of 6 μ l total volume. Thermocycling parameters differ among loci, but are available at <http://www.snprc.org/linkage/index.html>. Seven to ten PCR products were combined into single pools, an aliquot of LIZ-600 size standard (ABI) plus formamide was added, and this mixture loaded into the ABI 3730 instrument for capillary electrophoresis. Standard methods were used to determine each genotype for each individual sample. An image file of the raw data for multiplexed genotypes is created by installed ABI collection software. This image file was then analyzed using ABI GeneMapper software

(Applied Biosystems Inc., Foster City, CA), employing the local Southern method and the internal ABI size standards to estimate fragment lengths.

These data were used for three analyses. First, we conducted STRUCTURE (Pritchard et al 2000) and Principle Component Analysis (PCA) to ensure population-of-origin had been correctly assigned for animals in the retrospective samples. Second, we estimated pair-wise relatedness among individuals to detect cryptic relatedness in the samples and lastly, we conducted replicate association analyses as a form of genomic controls to assess the significance of the *CCL3L* association.

2.6 Additional Methods

Assessing population of origin

To ensure the individuals in the retrospective sample were correctly attributed to their population of origin and to ensure none were of admixed heritage, we used STRUCTURE and PCA. STRUCTURE was run on the complete set of microsatellites for 500,000 iterations with a burn-in of 100,000 iterations. Three independent runs of $K=2$ were run using the admixture model with correlated allele frequencies and pop-flags off. Convergence of each run was evaluated by visual inspection of the Ln(PID) plots. The results of the STRUCTURE runs confirmed that all individuals in the retrospective sample had been correctly assigned to their population. With the exception of two individuals, all showed q-values greater than 98%. These two individuals showed the greatest extent of admixture with q-values of ~94% (Figure 2.5).

Table 2.6: Microsatellite ID, number of alleles found in the retrospective sample and heterozygosity for the 53 typed microsatellites.

MARKER	Number of alleles	Heterozygosity
D11S2002	7	0.76
D12S67	23	0.91
D13S280	14	0.87
D15S108	14	0.85
D17S1605	13	0.81
D18S1140	12	0.87
D18S1371	12	0.87
D1S231	12	0.85
D22S280	15	0.86
D2S296	22	0.91
D3S1768	13	0.88
D5S1989	12	0.86
D6S266	11	0.88
MML10S27	12	0.82
MML11S2	13	0.81
MML12S29	13	0.80
MML12S9	11	0.76
MML13S7	13	0.74
MML14S21	13	0.87
MML14S27	12	0.72
MML15S21	15	0.89
MML15S3	19	0.88
MML16S27	16	0.79
MML16S46	9	0.74
MML17S39	8	0.76
MML19S19	7	0.81
MML19S27	13	0.82
MML1S1	5	0.53
MML1S42	12	0.82
MML1S8	15	0.78
MML20S35	13	0.73
MML20S36	12	0.85
MML2S34	11	0.84
MML2S41	17	0.88
MML3S16	6	0.69
MML3S4	12	0.84
MML3S43	11	0.82
MML3S9	21	0.93
MML4S21	18	0.90
MML4S8	13	0.85
MML5S25	15	0.87
MML5S32	14	0.82
MML5S38	10	0.85
MML6S27	10	0.73
MML7S2	11	0.75
MML7S9	14	0.87
MML8S40	7	0.60
MML8S56	10	0.84
MML8S7	9	0.76
MML9S30	11	0.81
MML9S44	18	0.88
MML9S6	12	0.83
MMI1S6	12	0.71

Additional confirmation of population assignment is provided by a principle component analysis (Figure 2.6A & 2.6B) conducted using the `prcomp` function in R. This analysis confirms that all samples have been correctly attributed to their population of origin.

Cryptic relatedness

As the individuals used in the retrospective analysis were all obtained from colonies, we used the microsatellite data to determine if there are significant levels of relatedness in the sample. Briefly, we used the modified Queller and Goodnight (QG; Queller and Goodnight 1989, Lynch and Ritland 1999) estimator of r to evaluate pair-wise relatedness of all individuals in the sample. We find that the Indian origin samples show, on average, higher levels of relatedness than the Chinese origin animals (Figure 2.7A). This result is expected, as export of Indian-origin animals has been banned since 1978. Therefore, the Indian-origin animals have been isolated for a longer period time than the Chinese origin animals. Both populations show modest levels of cryptic relatedness (e.g., several pairs of individuals show QG distances between half sib relationships [0.25] and first cousins [0.0625] corresponding to dark black squares in Figure 2.7B & Figure 2.7C). This level of relatedness, however, appears to cause only a slight bias (if at all) in our estimates of the significance of the association; see below.

Genomic control

To further assess the significance of CNV in the *CCL3L* locus with time until onset, we conducted replicate association analysis with each microsatellite allele. For each allele at each locus we recoded the microsatellite as 2 (homozygous for the allele), 1 (heterozygous with the

particular allele to be analyzed and another allele), and 0 (two alleles which are not that allele being analyzed). For each recoded allele we then conducted a Cox proportional hazard analysis with population origin as a covariate and recorded the p -value. We found that the distribution of the p -values closely follows a uniform distribution by visualizing the Q-Q plot (Figure 2.8A). The low level of relatedness seen in the above analysis may be causing the slight uptick in the distribution of the most extreme p -values. However, we find that the p -value for the association of *CCL3L* copy number with time until onset still falls in the extreme right tail of the p -value distribution (Figure 2.8B). Therefore, the relatedness seen in the above analysis is likely not significantly affecting the estimation of the association.

Power assessment and simulations for Chinese-origin Cox proportional Hazard model

In order to test whether the lack of an observed significant association between survivorship and *CCL3L* copy number for the Chinese-origin sample was due to power, we used a variant of non-parametric bootstrap re-sampling. In particular, we were interested in understanding whether the smaller sample size of the Chinese-origin sub-sample ($n = 20$), coupled with higher overall population mean *CCL3L* copy number accounted for the lack of a significant regression coefficient. In order to test this hypothesis, we generated $B = 100$ bootstrap data sets of size $n = 20$ or size $n = 37$ using the *Indian-origin* individuals sampled in proportion to the observed *CCL3L* copy number distribution in either the Chinese-origin sample (designated as “Chinese-like”) or the Indian-origin sample. For each data set, we sample with replacement triplets of (*CCL3L* copy number, time since infection, survivorship status) until either $n = 20$ or $n = 37$ individuals had been sampled. Each data set was then

run through the same Cox proportional hazard regression analysis, and the p -value of the likelihood ratio test comparing models m_0 (no factors) vs. m_1 (*CCL3L* as a factor) were retained.

We observe that the distribution of p -values among replicates with “Chinese-like” *CCL3L* CNV distribution is markedly uniform in comparison to the p -value distribution for animals with “Indian-like” *CCL3L* CNV distribution. This indicates a much lower power in the former as compared to the latter. In other words, in the “Chinese-like” simulations we observe few data sets with significant p -values, where as in the “Indian-like” simulations upwards of 50% - 80% of simulations where in the lowest p -value bin. This result holds regardless of whether $n = 20$ or $n = 37$ individuals are sampled, although the power for $n = 20$ in the Indian bootstrap simulations is also reduced. Our interpretation of this finding is that the Chinese-only sample has little power to detect an effect of *CCL3L* copy number on survivorship due to a lack of animals with low number of copies of *CCL3L* and smaller overall sample size.

REFERENCES

1. Goldstein, S., Brown, C.R., Dehghani, H., Lifson J.D., Hirsch, V.M. (2000) Intrinsic Susceptibility of Rhesus Macaque Peripheral CD4+ T Cells to Simian Immunodeficiency Virus In Vitro Is Predictive of In Vivo Viral Replication. *J. Virol.* 74(20): 9388-9395.
2. Trichel A.M. Rajakumar P.A., Murphey-Corb M. (2002) Species-specific variation in SIV disease progression between Chinese and Indian subspecies of rhesus macaque. *J Med Primatol* 31(4-5): 171-178.
3. Joag, S.V. Stephens EB, Adams RJ, Foresman L, Narayan O. (1994) Pathogenesis of SIVmac infection in Chinese and Indian rhesus macaques: effects of splenectomy on virus burden. *J. Virol.* 200(2): 436–446.
4. Ling, B., Veazey R.S., Luckay A., Penedo C., Xu K., et al. (2002) SIV(mac) pathogenesis in rhesus macaques of Chinese and Indian origin compared with primary HIV infections in humans. *AIDS* 16(11): 1489–1496.
5. Ling, B., Veazey R.S., Penedo C., Xu K., Lifson J.D., et al. (2002) Longitudinal follow up of SIVmac pathogenesis in rhesus macaques of Chinese origin: emergence of B cell lymphoma. *J. Med. Primatol.* 31(4-5): 154-163.
6. Irving, S.G., Zipfel P.F., Balke J., McBride O.W., Morton C.C., et al. (1990) Two inflammatory mediator cytokine genes are closely linked and variably amplified on chromosome 17q. *Nucleic Acids Res.* 18(11): 3261-3270.
7. Nakao, M., Nomiya, H., Shimada, K. (1990) Structures of human genes coding for cytokine LD78 and their expression. *Mol. Cell. Biol.* 10(7): 3646-3658.
8. Hirashima, M., Ono T., Nakao M., Nishi H., Kimura A., et al. (1992)

- Nucleotide sequence of the third cytokine LD78 gene and mapping of all three LD78 gene loci to human chromosome 17. *DNA Seq.* 3(4): 203-212.
9. Townson, J.R. Barcellos, L.F. Nibbs, R.J. (2002) Gene copy number regulates the production of the human chemokine *CCL3-L1* *Eur J Immunol* 32(10): 3016-3026.
 10. Nibbs, J.B., Yang, J., Landau, N., Moa, J., Graham, G.J. (1999) LD78 β , A Non-allelic Variant of Human MIP-1 α (LD78 α), Has Enhanced Receptor Interactions and Potent HIV Suppressive Activity. *J Biol Chem* 274(25): 17478-17483.
 11. Menten, P. Wuyts, A. Van Damme, J. (2002) Macrophage inflammatory protein-1. *Cytokine Growth Factor Rev.* 13(6): 455–481.
 12. Proost, P., Menten P., Struyf S., Schutyser E., De Meester I., et al. (2000) Cleavage by CD26/dipeptidyl peptidase IV converts the chemokine LD78beta into a most efficient monocyte attractant and CCR1 agonist. *Blood* 96(5): 1674–1680.
 13. Gonzalez, E. Kulkarni H., Bolivar H., Mangano A., Sanchez R., et al. (2005) The Influence of *CCL3L1* Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility. *Science* 307(5714):1434-1440.
 14. Meddows-Taylor, S., Donninger S.L., Paximadis M., Schramm D.B., Anthony F.S., et al. (2006) Reduced ability of newborns to produce *CCL3* is associated with increased susceptibility to perinatal human immunodeficiency virus 1 transmission. *J. Gen. Virol.* 87: 2055-2065.
 15. Kuhn, L. Schramm D.B., Donninger S., Meddows-Taylor S., Coovadia A.H., et al. (2007) African infants' *CCL3* gene copies influence perinatal HIV transmission in the absence of maternal nevirapine. *AIDS* 21(13): 1753-1761.

16. Tiemessen, C.T. & Kuhn, L. (2007) CC chemokines and protective immunity: insights gained from mother-to-child transmission of HIV. *Nat. Immunol.* 8(3): 219-222.
17. Dolan, M.J., Kulkarni H., Camargo J.F., He W., Smith A., et al. (2007) *CCL3L1* and CCR5 influence cell-mediated immunity and affect HIV-AIDS pathogenesis via viral entry-independent mechanisms. *Nat. Immunol.* 8(12): 1324-1336.
18. Shalekoff, S., Meddows-Taylor S., Schramm D.B., Donninger S.L., Gray G.E., et al. (2008) Host *CCL3L1* gene copy number in relation to HIV-1-specific CD4+ and CD8+ T-cell responses and viral load in South African women. *JAIDS* 48(3): 245-254.
19. Gibbs R., Rogers J., Katze M.G., Bumgarner R., Weinstock G.M., Mardis E.R., et al. (2007) Evolutionary and Biomedical Insights from the Rhesus Macaque Genome. *Science* 316(5822): 222-234.
20. Cox, D. R. (1972) Regression Models and Life Tables (with Discussion). *J. R. Stat. Soc. Series B* 34(2): 187-220.
21. Pritchard, J. K., Stephens, M., and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics* 155: 945-959.
22. Queller D.C., Goodnight, K.F. (1989) Estimating relatedness using genetic markers. *Evolution* 43: 258-275.
23. Lynch M, Ritland K. (1999) Estimation of pairwise relatedness with molecular markers. *Genetics* 152: 1753-1766.
24. Kaplan, E.L. & Meier, P. (1958) Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* 53: 457-481.
25. Geczy A.F., Kuipers H., Coolen M., Ashton L.J., Kennedy C., et al. (2000)

- HLA and other host factors in transfusion-acquired HIV-1 infection. *Human Immo.* 61: 172-176.
26. Lama, J. and Planellas, V. (2007) Host Factors influencing susceptibility to HIV and AIDS progression. *Retrovirology* 4(52).
27. Smith, S.M., Holland, B., Russo, C., Dailey, P.J., Marx, P., Connor, R.I. (1999) Retrospective analysis of viral load and SIV antibody response in rhesus macaques infected with pathogenic SIV: Predictive value for disease progression. *AIDS Res Hum Retroviruses* 15(8): 1691-1701.
28. Bailey, J.A., Gu Z., Clark R.A., Reinert K., Samonte R.V., et al. (2002) Recent Segmental Duplications in the Human Genome. *Science* 297(5583): 1003-1007.
29. Grambsch, P. and Therneau, T. (1994) Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 81(3): 515-526.
30. Steiper, M.E., Young, N.M. (2006) Primate molecular Divergence dates. *Mol Phylogenet Evol.* 41(2): 384-394.

CHAPTER 3

A novel method for mapping the pseudoautosomal boundary and application to 5 canid species and Rhesus Macaque³

3.1 Abstract

To date, the pseudoautosomal region of the sex chromosomes has been mapped in only a handful of eutherian mammals including humans, laboratory mouse, horse, cow and a single domestic dog. However, no previous study has examined a large number of individuals from a single species or several closely related species within a genus to determine the stability of the PAR boundary. Here we have developed a novel method for the precise mapping and identification of the pseudoautosomal region and its boundary using data from oligonucleotide arrays. Compared to traditional methods, which require development of cell lines, BAC libraries and FISH probes, or alternatively the complete sequencing of the Y-chromosome, our method is simple, efficient and can be applied to a large number of individuals. We verify that the method is able to detect the PAR boundary accurately by applying this novel method to human HapMap data. We use the method to map the boundary in 75 breeds of dog as well as grey and red wolves, coyotes, and black-backed jackals and find that the PAR boundary has been remarkably stable over the evolution of carnivores. We also apply our method to mapping the PAR boundary in Indian rhesus macaques, completing the mapping in all mammalian species for which adequate genomic information exists. Combining the results of our current work with previous work on mapping the PAR boundary in other species

³ Degenhardt et al. in preparation.

suggests that while the PAR boundary has changed over long evolutionary time periods, it has been remarkably stable in the short term.

3.2 Introduction

The pseudoautosomal regions (PAR) of the sex chromosomes of eutherian mammals is a small region of homology between the X and the Y-chromosomes. This region, a remnant of the anciently homologous chromosomes from which the sex chromosomes were born^[1], is the sole region of recombination between the X and the Y-chromosomes^[2, 3]. Genes in this region therefore show patterns of segregation and expression similar to autosomal genes and also avoid X-inactivation in females. The evolution of this region and specifically of the genic content and the boundary of this region is still poorly understood^[4].

While this region has been mapped in only a few species^[5-10], it is apparent that the position of the boundary has moved several times during the evolution of mammals. Additionally, previous work looking at the divergence between the X and Y-chromosomes has shown the existence of 5 strata with varying levels of divergence in humans suggesting that the PAR has evolved through a sequence of size reduction events^[7, 11-13]. The divergence within these regions has been used to approximately date the events, however, it is unclear if these dates are accurate due to potential rearrangements within these regions^[7]. A better understanding the specifics of how and when the PAR boundary has changed will require mapping the boundary in multiple additional species. Part of the lack of progress of mapping the PAR boundary in these additional species is that mapping the PAR, to this point, has required a multi-step process involving the development of cell lines from a minimum of a single male and female from each species, mapped BAC libraries and FISH

and PCR probes. These steps are labor-intensive and restrictive in the number of individuals that can be tested. With the advent of next-generation sequencing techniques, whole genome sequencing is becoming cheaper and faster and additional species will be sequenced at an ever-increasing rate. However, to the best of our knowledge, no one has used this wealth of genomic data to map PAR boundaries.

Here, we introduce a novel and simple method for mapping the position of PAR boundary using genotyping or aCGH arrays. These arrays both work by measuring the amount of DNA bound to the chips from a particular genomic location. In the case of the genotyping chips this intensity is measured for a single individual at a time, whereas for the aCGH chips, the intensity is a competitive measure from two individuals. Our method works by the same principles as CNV mapping, using a test and a reference set to define the region of intensity change on an aCGH or genotyping chip. In essence, within the PAR, males and females will show similar intensity while in the non-PAR, the signal in males will be lower than in the females. In mapping the PAR, we can know which samples should be used for the test (male) and reference (female) population *a priori*. We can then find the region on the haploid X-chromosome in males, which show elevated intensities. The elevated intensity regions correspond to regions, which are also found on the Y chromosome, and define the PAR. In addition, because our method is relatively inexpensive and simple, we are able to estimate of the PAR boundary for multiple individuals. In leveraging multiple individuals, we can attain higher confidence in the precise location boundary estimation than with single-individual FISH and BAC-based methods. Furthermore, by independently estimating the PAR across multiple individuals, we can determine the stability of the PAR

boundary within a species. Our method will continue to be applicable with new technology as it can also easily be modified to use information from next generation sequencing information to estimate the PAR, in a similar fashion to methods used for mapping CNVs using sequence data, or as for a recent method for mapping W-linked contigs in the chicken genome (Chen and Clark in prep).

We validate our method by application to human 500k SNP chip data run on CEU, YRI, CHB and JPN HapMap samples. We apply our method to data from Affymetrix Canine V2 SNP chips to estimate the PAR in domestic dogs, Grey Wolves, Red Wolves, Coyotes, and black-backed Jackals, as well as to NimbleGen 385k aCGH data to provide the first estimates of the PAR boundary in rhesus macaques. Finally, we place our results of the PAR mapping for individuals from multiple canid species and rhesus macaques into the broader context of mapped PAR boundaries to further the understanding of the evolution of this region.

3.3 Results

Here we have developed a novel method for precisely mapping the position of the pseudoautosomal boundary. The non-PAR regions of the sex chromosomes can be thought of as a massive copy number variable region with known phenotypic effect. That is, males have two copies of the PAR, but only a single copy of both the non-PAR X and Y-chromosomes whereas females have two copies of both the PAR and the non-PAR X and no copies of the non-PAR Y. Therefore, we can leverage known sex information to map the PAR boundary with intensity or read-depth data obtained from the X-chromosomes of a population of males and females. Our method works analogously to methods for mapping CNVs using array data ^[14]. In this case

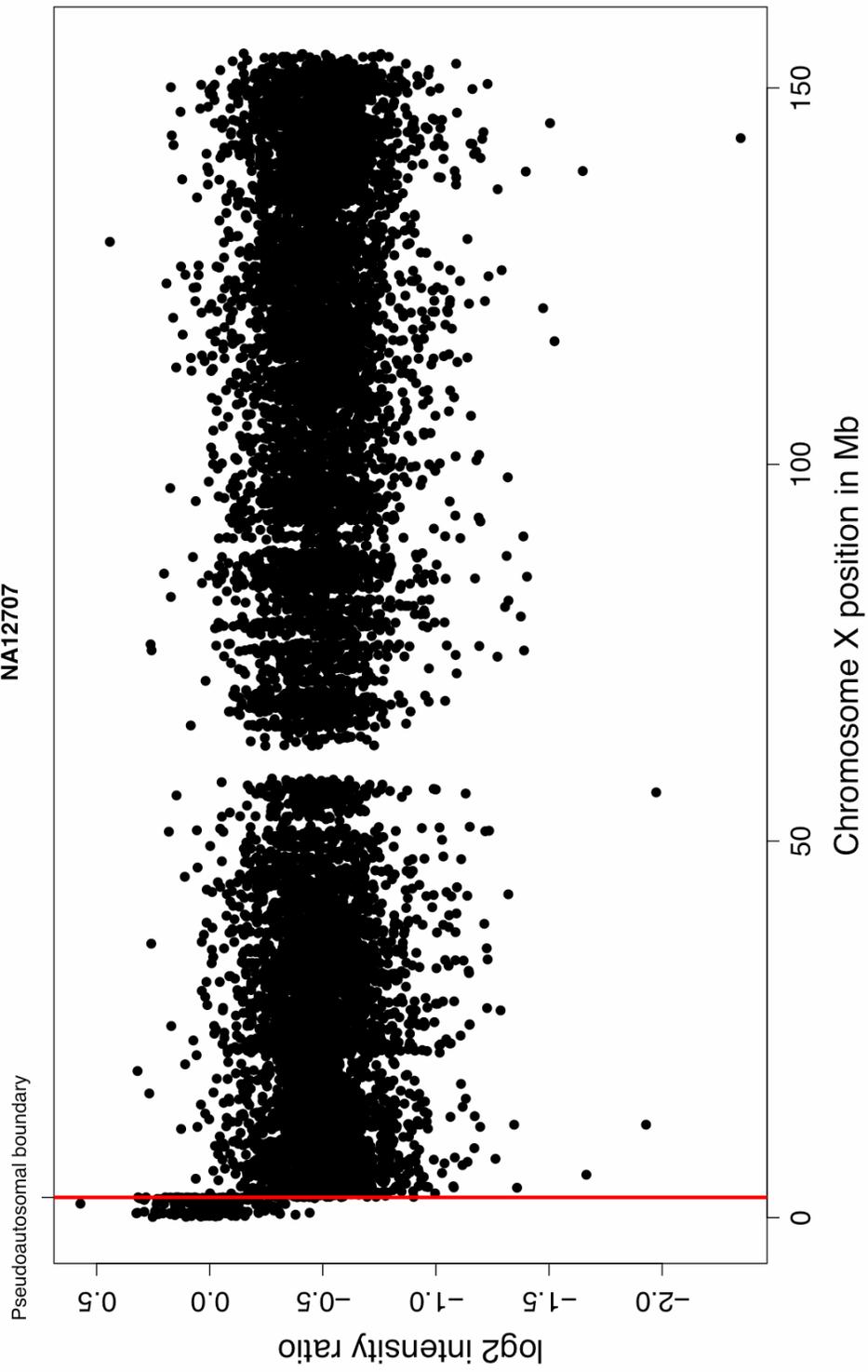
the test and reference populations can be determined *a priori*, using sex as the phenotype of interest. We have modified a Hidden Markov Model (HMM) CNV calling algorithm, developed previously (Degenhardt et al in submission), to identify the most likely change-point from PAR to non-PAR regions of the X-chromosome in our test samples. We call our new method spot_PAR, as it is a modification of our previously developed CNV detection method, spot_CNV. For more explicit details on the algorithm, see Methods.

To validate our new method, we applied spot_PAR to human Affymetrix 500k SNP-chip data and are able to precisely map the position of the PAR boundary in our HapMap test samples to a 9.1 kb (9,100 bp) window, which is centered on the known position of the boundary in humans, in the XG gene (Figure 3.1). The precision of the method is only limited by the density of the probes in the region of the PAR. Therefore, if for example an array has a probe density of one probe per 1 kb in a tiling-path array, we can place the position to within a 1 kb interval. This precise estimate as well as the intuitive ability of CNV-like detection to identify the PAR boundary supports the use of our method to estimate the PAR in species where the boundary is not known. Estimation of the Canine PAR:

Using the Affymetrix Canine V2 SNP chip data collected previously^[15] (Boyko et al accepted, Degenhardt et al. in submission) we applied our new method for identifying the PAR region (Figure 3.2). We used as the reference the set all of the female dogs from these studies to analyze the male individuals from 75 dog breeds. Our expectation, therefore, is that within the non-PAR X-chromosome the intensity for the reference individuals (female) will be twice that of the test individuals (male) and that the intensity within the PAR will be the same for males and females.

Figure 3.1: Validation of novel method with human HapMap data. Log_2 intensity ratio from the Affymetrix 500k SNP chip for the X chromosome of NA12707 using the female HapMap samples as the reference population. The x-axis shows the position in Mb along the X-chromosome. The red lines shows the known position of the PAR1 boundary. Note that the Affymetrix 500k chip contains no probes in the PAR2 region

HapMap male individual
NA12707



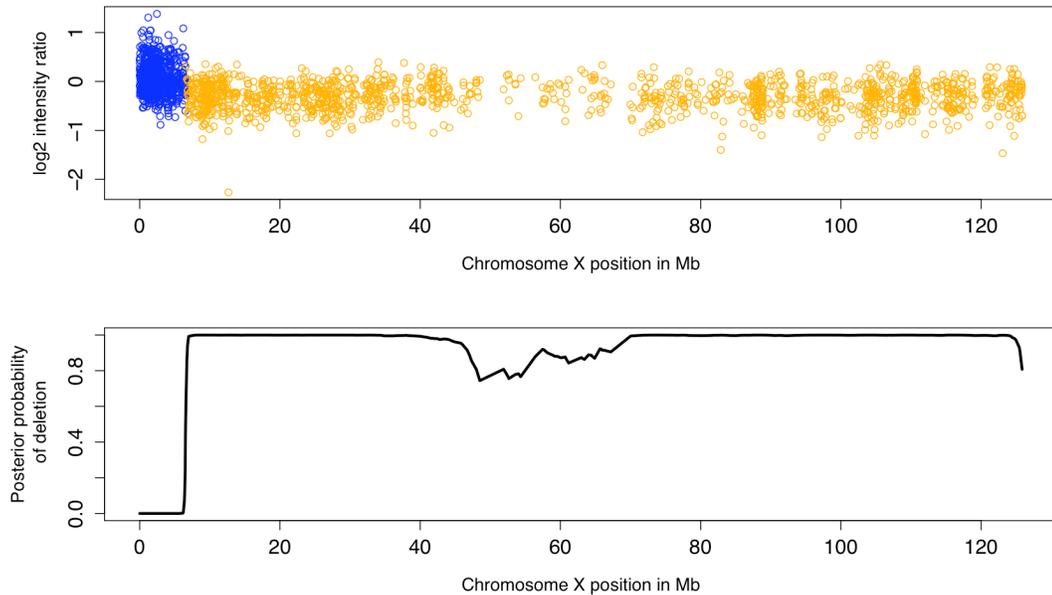


Figure 3.2: Example of novel method run on a single male dog chip. A) An example of the chromosome X \log_2 intensity ratio from a single male dog with the median of all female dogs used for the reference. The intensities are colored by the corresponding call (blue = diploid; gold = haploid) from spot_PAR. B) The corresponding posterior probability of haploid state from the spot_PAR.

This corresponds to a signal of a deletion across the non-PAR region of the X in males. The region of abrupt change in intensity levels is thereby the PAR boundary. We find that the intensity change occurs most frequently (425 out of 558 dogs) in domestic dogs in a ~ 53 kb window between probes at position 6,541,232 and 6,593,955 on the X chromosome. In the remaining 133 dogs the intensity change occurs between probes at positions 6,593,955 and 6,733,178 suggesting that the PAR boundary is contained within the 192 kb region between probes 6,541,232 and 6,733,178 (Figure 3.3). This agrees with the previous result ^[9], which found that the PAR boundary mapped to a 2

kb window within the gene *Shroom2* in a single male dog. Figure 2 shows that the position of the PAR boundary is likely located within this window; however, it may show some level of variability in the exact location within this window. Further, we find two lines of evidence to suggest that while the region of homology is fixed across all breeds, recombination may occur less frequently near the PAR boundary. The first observation is that the binding affinity of the probes in males is reduced on the Affymetrix SNP chip in the proximity of the PAR boundary. This leads to a weak signal of a deletion in this region (indicating it is non-PAR) and even a slight reduction of the estimated size of the PAR in some individuals. The binding affinity loss could be explained by greater divergence caused by lower recombination of the two chromosomes near the boundary. Concordant with this decrease in affinity, we also see an increase in heterozygosity within the high quality probes specifically in male dogs and not the female dogs (Figure 3.4; data from Boyko et al 2010). We suggest that the reason for this increase in male heterozygosity and decrease in binding affinity on the chips can be explained by a decrease in recombination between the X and Y-chromosomes in this region. As recombination is reduced, its homogenizing effect is also reduced and drift can lead to increasingly divergent haplotypes on the Y-chromosome in this region. This phenomenon, termed “PAR attrition” has been seen previously in the mapping of the PAR boundary in cattle ^[7].

We next examined four canid species for which the PAR had not been previously mapped. We analyzed intensity data from n=95 grey wolves, n=8 red wolves, n=26 coyotes and n=2 black-backed jackals. We find that the PAR boundary consistently maps to within the gene *Shroom2* in all of these species (Figure 3.3), showing that the PAR boundary is likely consistent within the

entire genus *Canis*. This position is also consistent with the genic position of the PAR boundary identified for domestic cats ^[16]. Therefore, this position is likely the ancestral position of the PAR boundary for the entire order of Carnivora. Examination of the genic content of the human PAR shows that it is largely consistent with the first ~1.4 Mb of the canine PAR ^[9]. However, based on the reference genome build, we find a single rearrangement in the three most distal genes on the canine X-chromosome with respect to the human X (Figure 3.5). Further analysis is necessary to confirm that this is not an artifact of the reference build.

We next compared the GC content of the PAR in the dog and human reference genomes. Previous studies of human, mouse, cattle and the horse have suggested that the GC content in the PAR is driven by an increase in recombination in this region. Therefore, we would expect to see the GC content of the PAR in dogs to gradually increase moving distally within the PAR. Similar to these previous studies, we do observe a slight increase in GC content within the PAR (Figure 3.6).

However, based on this hypothesis, we would also expect that the region of the canine PAR which has become X-specific in humans, should show a lower GC content in humans than that observed in dogs. As well, we would expect the patterns of GC content to differ between these two organisms as recombination in this region in humans likely ended more than 35 million years ago before the divergence of old-world monkeys and great apes with the relocation of the of the PAR boundary to the XG gene.

Figure 3.3: Heat-map representation of the transformed posterior probabilities from spot_PAR for all dogs and wild canids. Each row represents a single male dog and each column is a SNP probe position along the X. Only the first 10 Mb of the X is shown. This region contains all genes, which are thought to have occurred in the ancestral eutherian PAR.

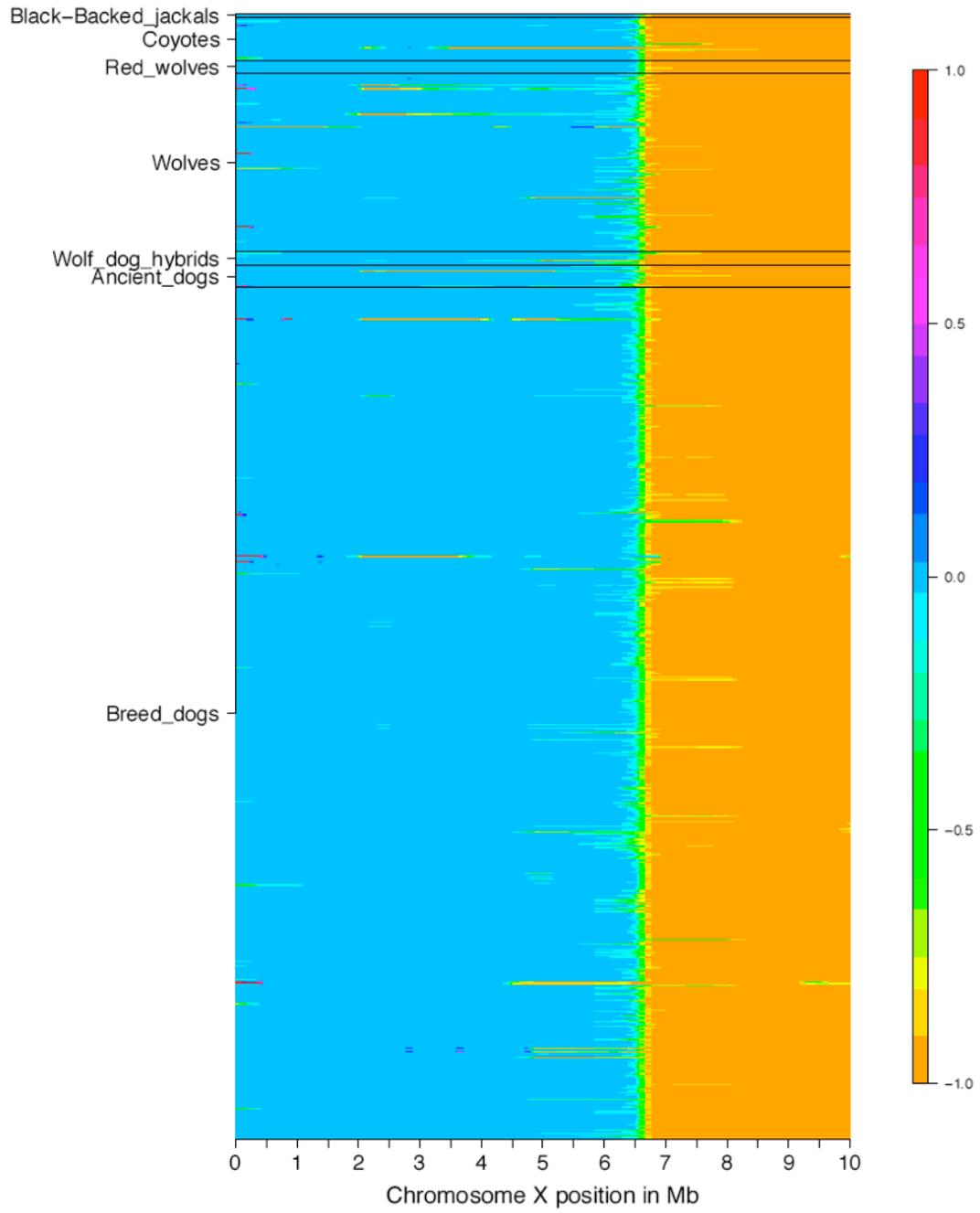


Figure 3.4: Extended PAR attrition in domestic dogs. Average male to female heterozygosity in breed dogs from the CanFam V2 chip in black dots (corresponding to the right y-axis) and the median (and 2 standard deviations around) of the \log_2 intensity ratio in the blue line (corresponding to the left y-axis). These values are shown for the first 8 Mb of the X chromosome. The figure shows a strong increase in male-specific heterozygosity (corresponding to differences between the two gonotypes) between ~ 5 Mb and ~ 6.6 Mb.

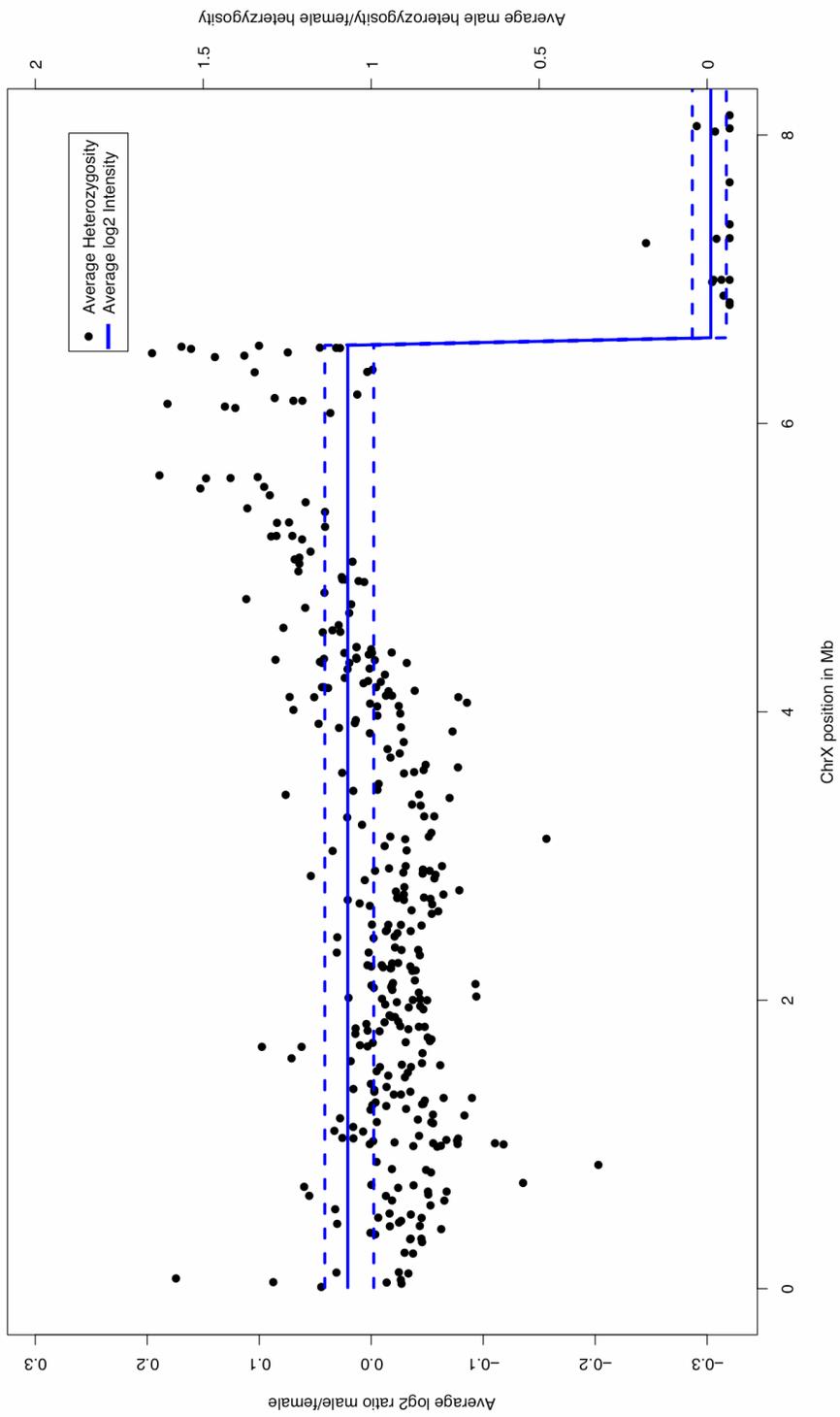
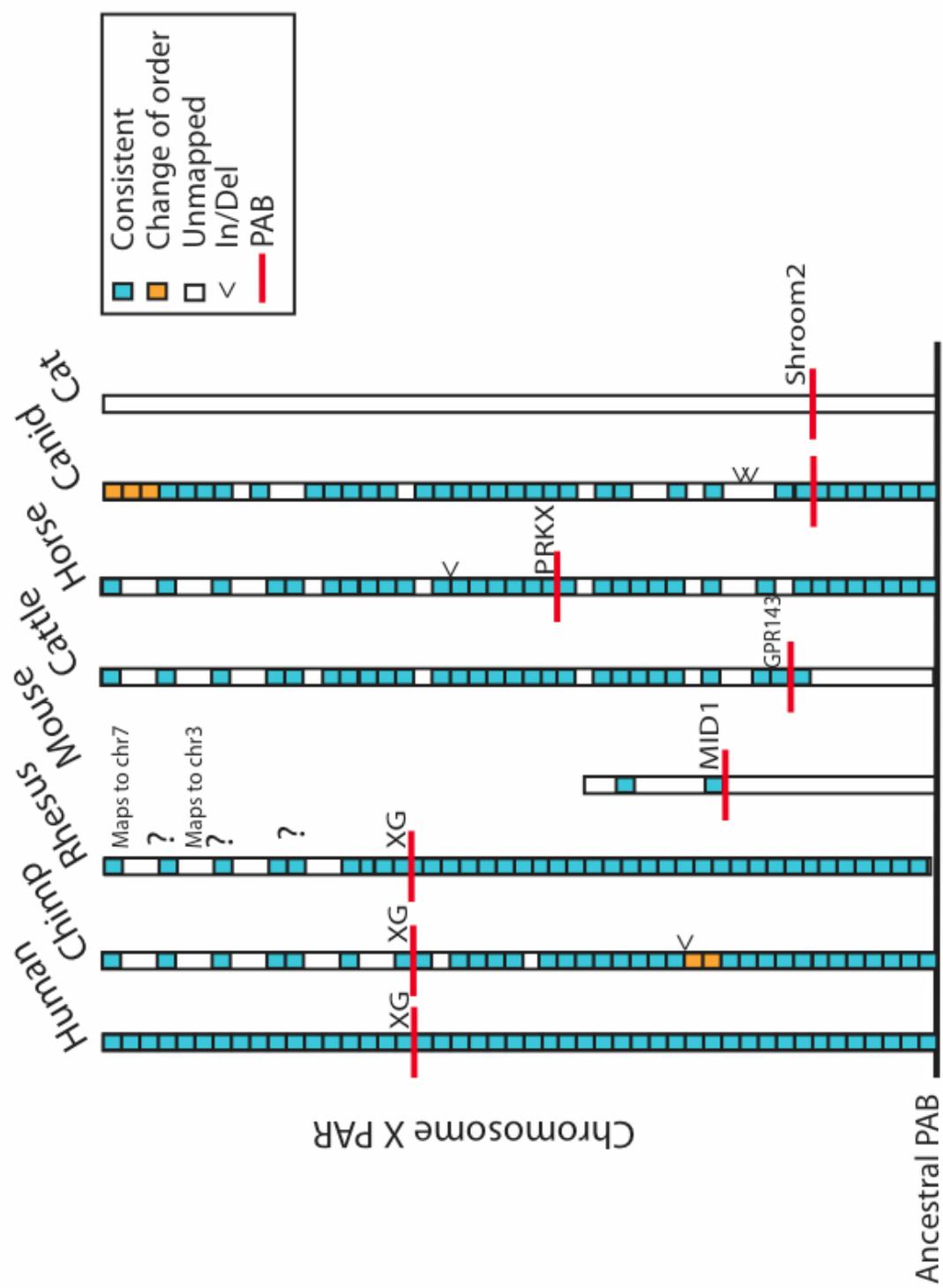


Figure 3.5: Evolution of PAR genic content across the Mammalian groups. An ideogram of the genes contained in the PAR of all groups/organisms mapped to date. The blue boxes represent genes with conserved order and orientation, the orange boxes show genes that have been rearranged in either order or orientation, side carrots represent insertion and deletion events, and white regions represent either missing genes or regions that are unmapped. This figure also shows the regions on the rhesus genome that map to the rhesus chromosome 3 and 7 as well as the unmapped contigs.

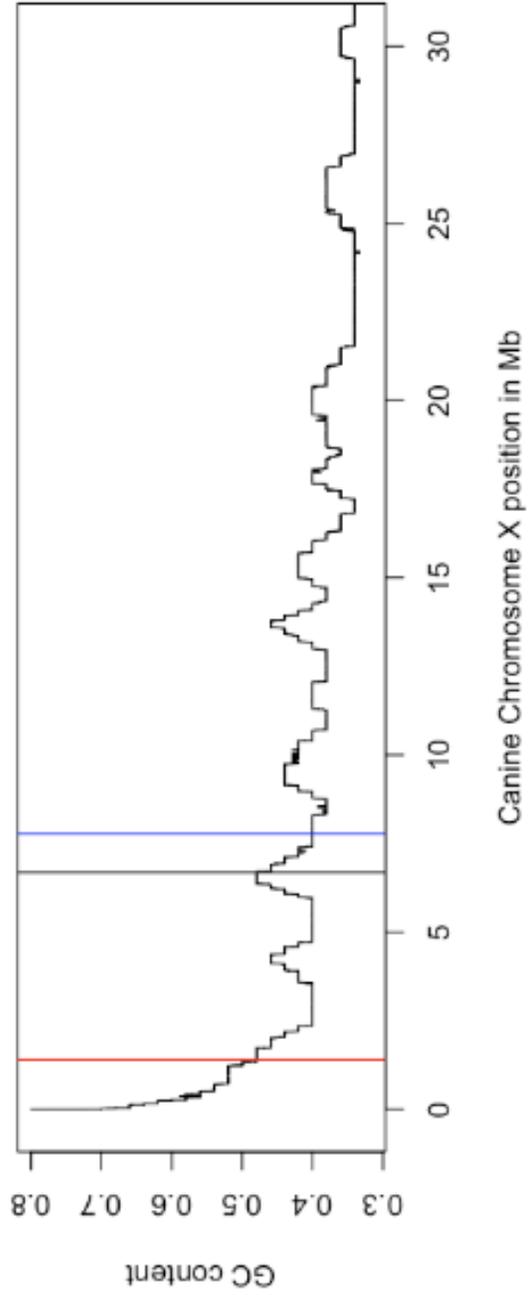
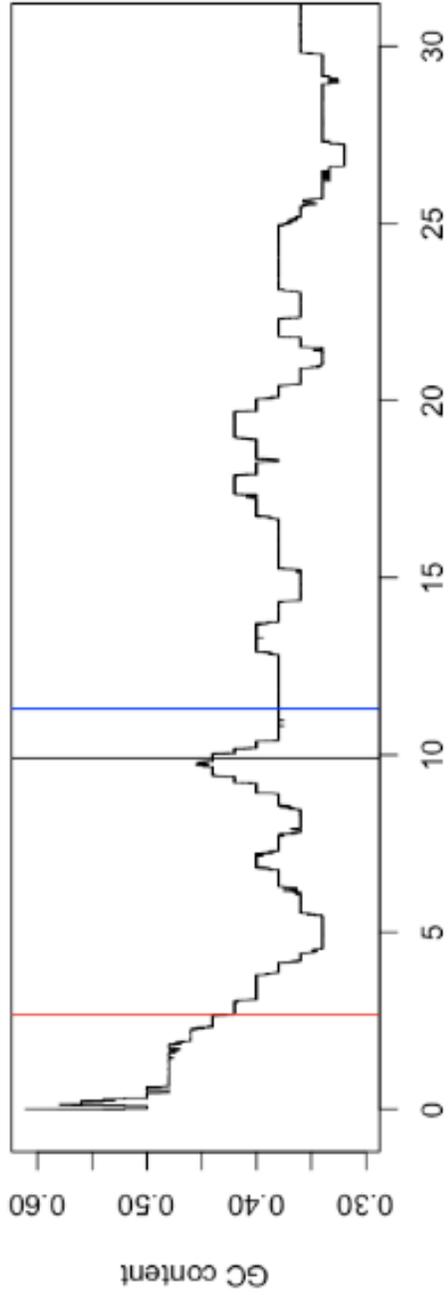


Upon comparison of the canine and human X-chromosomes, however, we find a striking similarity in the GC content. While the GC content is slightly higher in the canine PAR than in the homologous region of the human X-chromosome, the pattern of GC content change along the ancestral PAR has remained remarkably constant over evolutionary time. We therefore suggest that a factor other than recombination may be responsible for the pattern of GC content observed in the distal portion of the X-chromosome.

Estimation of the rhesus macaque PAR:

Using NimbleGen 385k aCGH data collected as part of a previous rhesus macaque copy number variation analysis ^[17], we applied our new method to map the PAR in rhesus macaque. These data were collected from 4 male macaques using a single female as the reference ^[17]. We find that the rhesus PAR extends proximally from the chromosome X p-arm telomere to ~400 kb at the location of the XG gene. This further supports previous work suggesting that the first PAR (PAR1) boundary exists in the same location in all old-world monkeys and great-apes despite the lack of the great-ape specific Alu insertion in OWM ^[18]. The above result indicates, however, that the rhesus PAR1 is only 400 kb in length as compared to the 2.68 Mb human PAR1. Further analysis of the rhesus reference build (rheMac2) showed that the distal portion of the p-arm of the X ends ~10 kb after the gene ZBED1. The portion of the X-PAR from ZBED1 proximal to the XG gene, showed nearly perfect co-linearity with the human PAR region. To investigate the reason for the missing section of the X-chromosome from ZBED1 to the telomere, we analyzed the primate alignments available on the UCSC genome browser.

Figure 3.6: GC content of human and canine PAR. Sliding window median of GC content in 10 kb windows along the first 30 Mb of the human and dog X-chromosomes. The red line shows the current boundary position of the human PAR, the black line shows the current boundary position of the dog PAR and the blue line shows the hypothesized position of the ancestral PAR.



We found that two regions of the human X-chromosome distal to ZBED1 showed alignments with rheMac2. The first of these regions maps to a single, unanchored contig on rhesus chromosome 7 (chr7: 1,146,358 – 1,211,637). This contig contains four genes, which map to the most distal portion of the human PAR1. These genes (PLCXD1, GTPBP6, NCRNA00107, and PPP2R3B) map to the rhesus contig in the same order as seen on the human PAR1. The second region maps to four, small, unanchored contigs on rhesus chromosome 3 spanning the region from 38,092,603 to 38,186,563. This region contains no genes in either the rhesus or the human genome builds. Based on the primate alignments, we find that the rhesus build is still missing two large regions of the rhesus PAR. The first spans the region from human chromosome X position 346,251 to human chromosome X position 844,335 and contains the gene SHOX, a known disease gene in humans. The second, much larger region spans the region from human chromosome X 930,577 to 2,352,116 and contains genes CRLF2, CSF2RA, IL3RA, SLC25A6, ASMTL, P2RY8, SFRS17A, and ASMT. Searching the rhesus BAC and shotgun libraries on NCBI resulted in hits for all of these genes (see methods). Therefore, all of the genes missing from the genome build are present in the rhesus genome but as unassembled contigs, which have not been placed in the rheMac2 build.

Additionally we used our method to test for the existence of a second PAR (PAR2) on the q-arm telomere as seen in humans. Consistent with previous research we find no evidence for the existence of a second PAR on the q-arm of the rhesus X-chromosome. This further supports the hypothesis that the generation of the PAR2 region is due to a human-specific

translocation of these X-chromosomal regions on to the Y-chromosome. Again, based on the reference genome build we find that most genes present in the human PAR2 are found in the same order and orientation on the rhesus X. The one exception of this is the most distal gene WASH1, which is found in an unanchored contig on rhesus chromosome 13, similar to the missing regions of PAR1 above. Therefore, the initial translocation from the autosomes onto the X-chromosome leading to the genic content of PAR2 and the subsequent rearrangements of these genes ^[19] occurred before the OWM/great-ape split. Further analysis of the canine genome build also shows that again, with the exception of WASH1, these genes exist on the X-chromosome and not on the autosomes, pushing back the date of this initial translocation event to ~87 million years ^[20].

Gene Content and Evolution of the PAR:

Combining our results from domestic dogs, wild canids and rhesus macaque with the information from previous studies of the PAR boundary and genic content we have constructed the most complete analysis of the evolution and genic content of the PAR to date.

The results of our analysis show that the genic content of the ancestral PAR region from the AMELX gene distal to the X telomere shows a high level of conservation consistent with Ohno's law ^[1]. The one exception to this observation is the laboratory mouse ^[6]. However, many of the genes that resided on the PAR in other species are found on the autosomes in the mouse. This suggests that these genes may require a diploid expression. Based on our analysis of rhesus macaque, we also suggest that the genic content is largely conserved in this species even though the mapping is incomplete in the current build. A previous study had also suggested that two

genes from the PAR in chimpanzees had moved to the non-PAR portion of the X-chromosome ^[8]. A closer inspection of these genes in the browser shows that one is in the unmapped portion of the X-chromosome and the other is again likely a mapping artifact. We do find evidence for a single inversion of the genes PNPLA4 and VCX as well single insertion of gene RPL28 in the chimp genome. These genes are found in the ancestral PAR region however, which is now on the non-PAR X-chromosome in the OWM and great-apes. Within the canine PAR region we find a single inversion in the most distal three genes, changing the order of these genes with respect to all other species analyzed (Figure 3.5).

Based on the combined results we see that the PAR boundary has moved several times during the evolution of mammals, however, we find no evidence for recent movement. Analysis of the Bovidea and the *Canis* data, combined with the evidence for the PAR boundary position in domestic cat, show that primary movement of the PAR seems to have occurred early in the radiation of mammals.

3.4 Discussion

Here we have developed a simple, precise, and efficient method for mapping the PAR boundary. Our method can be employed to map the PAR for additional species as new genomes become available. The method also makes it far easier to map the PAR in multiple individuals within a species and within closely related species without requiring the development of additional molecular resources such as cell lines. We have applied this method to the mapping the PAR in 75 breeds of domestic dog as well as four species from the genus *Canis*. We have also used this method to map the PAR in the rhesus macaque. Our results provide insight into several aspects of PAR

evolution. We are able to show for the first time that the PAR appears to be stable across individuals within a species. Our results are consistent with reduced recombination near the boundary of the PAR leading to the existence of divergent haplotypes within this region. This reduction in recombination leads to extended “PAR attrition” in the canid species. The level of attrition seen in the canid species extends much further than that previously seen in the bovids ^[7]. This taken with the suggestion that that bovid PAR boundary is a recent movement, and the position of the canid PAR near the human X-chromosome strata-3 boundary lead us to suggest that the Carnivora PAR boundary is located at the most ancestral position so far observed in any organism and may represent the location of the first reduction in size from the ancestral eutherian position thought to be located in the gene AMELX ^[21].

Our results taken together with previous work show that this moving boundary is remarkably stable within more recent evolutionary time. The current results show that the PAR is not only fixed within domestic dogs, but also fixed likely within the genus *Canis* and possibly across the entire order Carnivora. We also find that the position of the PAR boundary in rhesus macaques supports the hypothesis that the primate PAR is fixed across all old-world monkeys and great apes. Previous work has also shown that the bovid PAR boundary is shared across all members of that family that have been tested. Therefore, it appears that the PAR boundary moved several times early in the radiation of mammalian species, but the position of the boundary has been relatively stable since that time. Given the increasing availability of next-generation sequencing methods, this assertion can be tested in coming years with the completion of genomes from additional mammalian species.

The current study also calls into question the prevailing notion of higher recombination leading to higher GC content in the PAR. The current results show that the entire ancestral PAR spanning distally from the AMELX gene shows the same pattern of GC content variation in the domestic dog and human genome even though these species show drastic differences in the position of the PAR boundary. More work is needed to disentangle the mystery of GC content variation within this region.

Finally we have examined the gene content of the X chromosome spanning the ancestral PAR and find that the genes contained on the X have been remarkably stable over the evolution of mammals (Figure 3.5). With the exception of mouse, we find very few duplications, deletions, translocations, or inversion within this region. Given the extensive rearrangements seen in the rest of the genome when comparing these organisms, even in the number of chromosomes, we find this relative quiescence intriguing.

3.5 Methods

Data collection:

Dog and wild canid samples were collected as previously described in ^[15] (Boyko et al. accepted). The rhesus aCHG array data were collected as described in ^[22] and downloaded via the Gene Expression Omnibus site. Array intensity data for the Canine Affymetrix chips were calculated as described in (Degenhardt et al in submission).

Novel method for mapping of PAR:

Previous methods for mapping the PAR involved the construction of FISH probes from mapped BAC libraries. Additionally, these methods then require that cell-lines are available in order to localize these probes to the prophase chromosome spread. These methods are therefore prohibitive in the

number of samples one can analyze as well as the rate at which these novel PAR regions can be analyzed. Our novel method works by using the reference build to construct an oligonucleotide array. We then only require simple DNA extractions from a minimum of a single male and female to determine the location of the PAR. Increasing the number of male and female samples analyzed will increase our confidence in the localization the PAR boundary.

In order to determine the most likely position of the PAR boundary from array data we have implemented a modified copy number HMM. Our HMM has been modified from its original version as follows. The equation used to calculate the transition probabilities was optimized to identify larger events, i.e. the distance at which a transition becomes likely was extended to ~ 1 Mb. We also removed the step to transform the intensities to $N(0,1)$ and then removed the chromosomal wave correction. These modifications improve the detection of larger fragments.

To test our novel method for mapping the PAR boundary we first applied the method to human data for which the exact location of the PAR is known from multiple individuals. For this we used the Affymatrix 500k data collected on HapMap individuals. We mapped the PAR in all male HapMap individuals using the females as the reference population. The SNP intensity data were quantile normalized and extracted using the tools available from Affymetrix. The median intensity for each SNP position of all female individuals was calculated and then used as the reference to calculate the \log_2 intensity ratio. These data were then run through a modified version of Spot_CNV, termed Spot_PAR, to determine the most likely boundary location. The modifications of Spot_CNV included removing the additional normalization step, removing the genomic-wave correction step and reducing the transition

probabilities to optimize for the detection of large events.

Next we applied this method to the previously collected dog and wild canid data. The data were again quantile normalized using the standard Affymetrix software, however, the normalized intensity data for the Canine V2 custom SNP chips was extracted and summarized using a previously developed custom method. Analysis of the data proceeded similarly to the human data. The median SNP intensity was calculated across all female samples for each SNP position and this value was used to calculate \log_2 intensity ratio. We then used Spot_PAR to determine the most likely boundary position for each male dog.

Data from the four male rhesus macaques used in the study by Lee et al. ^[17] were obtained from the GEO website. These data were analyzed using spot_PAR we analyzed the normalized \log_2 intensity data directly in the form it was obtained.

To determine the genic content of the PAR from additional species, we used the information available on the UCSC genome browser for humans, chimpanzees, mouse, rhesus and domestic dog. For the bovid and equine PAR, we used the genic content and orientation presented here ^[8] and here ^[7].

In order to determine the existence of the missing section of the rhesus macaque build, we first visually inspected the primate alignments from the human genome browser. The identified regions were determined to be unanchored by examining previous publications regarding the linkage map ^[23] and the radiation-hybrid map ^[24]. The BACs containing these sections were not placed in either of the above studies. Further, both of these fragments were placed within large gaps with no connection to the assembly. To further determine the existence of the additional PAR genes not discovered above,

we obtained the exon sequence from the human UCSC browser table for the unmapped regions and then used BLAST to look for alignments in the NCBI nucleotide database restricted to search only for macaque specific hits. All genes were found to exist as gene predictions based on GNOMON and to be located in small shotgun sequence contigs that had not been placed in the reference build.

REFERENCES

1. Ohno S. (1967) Sex chromosomes and sex-linked genes. : Springer Berlin..
2. Burgoyne PS. (1982) Genetic homology and crossing over in the X and Y chromosomes of mammals. *Hum Genet* 61(2): 85-90.
3. Ellis N, Goodfellow PN. (1989) The mammalian pseudoautosomal region. *Trends in Genetics* 5: 406-410.
4. Das PJ, Chowdhary BP, Raudsepp T. (2009) Characterization of the bovine pseudoautosomal region and comparison with sheep, goat, and other mammalian pseudoautosomal regions. *Cytogenet Genome Res* 126(1-2): 139-147.
5. Petit C, Levilliers J, Weissenbach J. (1988) Physical mapping of the human pseudo-autosomal region; comparison with genetic linkage map. *EMBO J* 7(8): 2369-2376.
6. Perry J, Palmer S, Gabriel A, Ashworth A. (2001) A short pseudoautosomal region in laboratory mice. *Genome Res* 11(11): 1826-1832.
7. Van Laere AS, Coppieters W, Georges M. (2008) Characterization of the bovine pseudoautosomal boundary: Documenting the evolutionary history of mammalian sex chromosomes. *Genome Res* 18(12): 1884-1895.
8. Raudsepp T, Chowdhary BP. (2008) The horse pseudoautosomal region (PAR): Characterization and comparison with the human, chimp and mouse PARs. *Cytogenet Genome Res* 121(2): 102-109.
9. Young AC, Kirkness EF, Breen M. (2008) Tackling the characterization of canine chromosomal breakpoints with an integrated in-situ/in-silico approach: The canine PAR and PAB. *Chromosome Res* 16(8): 1193-

1202.

10. Das PJ, Chowdhary BP, Raudsepp T. (2009) Characterization of the bovine pseudoautosomal region and comparison with sheep, goat, and other mammalian pseudoautosomal regions. *Cytogenet Genome Res* 126(1-2): 139-147.
11. Lahn BT, Page DC. (1999) Four evolutionary strata on the human X chromosome. *Science* 286(5441): 964-967.
12. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, et al. (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423(6942): 825-837.
13. Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, et al. (2005) The DNA sequence of the human X chromosome. *Nature* 434(7031): 325-337.
14. Wang K, Li M, Hadley D, Liu R, Glessner J, et al. (2007) PennCNV: An integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 17(11): 1665-1674.
15. Vonholdt BM, Pollinger JP, Lohmueller KE, Han E, Parker HG, et al. (2010) Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* 464(7290): 898-902.
16. Murphy WJ, Davis B, David VA, Agarwala R, Schaffer AA, et al. (2007) A 1.5-mb-resolution radiation hybrid map of the cat genome and comparative analysis with the canine and human genomes. *Genomics* 89(2): 189-196.
17. Lee AS, Gutierrez-Arcelus M, Perry GH, Vallender EJ, Johnson WE, et al. (2008) Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. *Hum*

- Mol Genet 17(8): 1127-1136.
18. Ellis N, Yen P, Neiswanger K, Shapiro LJ, Goodfellow PN. (1990) Evolution of the pseudoautosomal boundary in old world monkeys and great apes. *Cell* 63(5): 977-986.
 19. Charchar FJ, Svartman M, El-Mogharbel N, Ventura M, Kirby P, et al. (2003) Complex events in the evolution of the human pseudoautosomal region 2 (PAR2). *Genome Res* 13(2): 281-286.
 20. Rose KD, Archibald JD. (2005) The rise of placental mammals: Origins and relationships of the major extant clades. : Johns Hopkins Univ Pr.
 21. Iwase M, Satta Y, Hirai Y, Hirai H, Imai H, et al. (2003) The amelogenin loci span an ancient pseudoautosomal boundary in diverse mammalian species. *Proc Natl Acad Sci U S A* 100(9): 5258-5263.
 22. Lee AS, Gutierrez-Arcelus M, Perry GH, Vallender EJ, Johnson WE, et al. (2008) Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. *Hum Mol Genet* 17(8): 1127-1136.
 23. Rogers J, Garcia R, Shelledy W, Kaplan J, Arya A, et al. (2006) An initial genetic linkage map of the rhesus macaque (*Macaca mulatta*) genome using human microsatellite loci. *Genomics* 87(1): 30-38.
 24. Murphy WJ, Agarwala R, Schaffer AA, Stephens R, Smith C, Jr, et al. (2005) A rhesus macaque radiation hybrid map and comparative analysis with the human genome. *Genomics* 86(4): 383-395

APPENDIX

Supplementary Table 1 for Chapter 1: Chromosome and positions of the 1220 high-confidence copy number variable regions detected in the discovery set.

Chromosome	Start Position	End Position
chr1	3000644	3235169
chr1	3822566	4117090
chr1	4582240	5114110
chr1	9055335	9521822
chr1	10335181	10565505
chr1	11255981	11579983
chr1	12685135	12967027
chr1	13051442	14736488
chr1	18539185	18634296
chr1	19062642	19302352
chr1	20336886	20472524
chr1	24809430	24980186
chr1	25416626	25490435
chr1	25718943	25849056
chr1	26593657	26785145
chr1	27630825	27956581
chr1	35097243	35437160
chr1	38660324	41082723
chr1	43215422	43285026
chr1	43901557	44071542
chr1	46810015	47109976

chr1	48089171	48384096
chr1	53444214	53674909
chr1	56681034	57103154
chr1	58178872	58318485
chr1	58577248	59061050
chr1	59417101	59699942
chr1	62283192	63052178
chr1	63531780	64169748
chr1	71164290	72550763
chr1	78046272	78238933
chr1	79119307	79250188
chr1	81296736	81477440
chr1	81770617	81867536
chr1	84521577	84719370
chr1	85464893	85549555
chr1	85867249	86160389
chr1	90420255	90957023
chr1	93218295	93591235
chr1	98158169	98187757
chr1	98197528	98274888
chr1	98815836	99027880
chr1	99720622	100222059
chr1	101109172	101835118
chr1	106813764	107437468
chr1	108115387	108421887

chr1	108811454	109048520
chr1	111115818	111183766
chr1	113369949	113505178
chr1	114777142	114873770
chr1	115727500	115765952
chr1	119904509	120123427
chr1	123723529	123901347
chr1	124136750	124994470
chr1	125125146	125405633
chr2	4345719	4538505
chr2	4879531	5068364
chr2	7219945	7266621
chr2	7496646	8465313
chr2	8708478	8932343
chr2	10138443	10684556
chr2	10971993	11302136
chr2	11900421	12004288
chr2	12483868	12929725
chr2	13097632	13520359
chr2	15122913	15500210
chr2	18514845	18849820
chr2	19360520	19613399
chr2	20381004	20453714
chr2	20943172	21258104
chr2	22171160	22343879

chr2	25927754	26198302
chr2	29128025	30860021
chr2	30880380	31150729
chr2	33154763	33306189
chr2	34003612	34174007
chr2	34944575	35417212
chr2	35935762	35982254
chr2	36257252	37571694
chr2	44588424	44818616
chr2	44928619	45162871
chr2	49814148	49974670
chr2	52732458	53233259
chr2	55904816	56279680
chr2	58888672	59175508
chr2	59539647	60100157
chr2	60793488	60880200
chr2	61544849	61760556
chr2	64384122	64408558
chr2	64540057	64597572
chr2	65487595	66399935
chr2	66616059	66953407
chr2	68140245	68248458
chr2	68929221	69213920
chr2	69723209	70753387
chr2	73201002	73385681

chr2	73539888	73591450
chr2	75760425	75801139
chr2	76829469	76868303
chr2	77487304	77553387
chr2	77717393	77754138
chr2	77975348	78079394
chr2	78497487	78757732
chr2	79069666	79393724
chr2	80070134	80362849
chr2	81296808	81389461
chr2	82031360	82176654
chr2	83232947	83288664
chr2	83444920	83515711
chr2	85763679	85855229
chr2	86673963	87043658
chr3	3023852	3620587
chr3	3931620	4112009
chr3	7157590	7480730
chr3	13800744	13879127
chr3	17865317	18054602
chr3	26911616	26952327
chr3	27492245	28354775
chr3	29675862	29728566
chr3	32940802	33836297
chr3	34166254	34403807

chr3	35490383	35608297
chr3	36656068	37138163
chr3	38435577	38607091
chr3	40168757	40254259
chr3	40464652	40634873
chr3	42294033	42348458
chr3	45782517	46147426
chr3	49538829	49619879
chr3	53131637	53285522
chr3	53737530	53882908
chr3	56028905	56189807
chr3	56730322	56804659
chr3	56965210	57063862
chr3	59570345	59644944
chr3	60604161	60654451
chr3	60863080	60931147
chr3	62661742	62983791
chr3	63065087	63094611
chr3	63224083	63449859
chr3	63533946	63693565
chr3	65120405	65438469
chr3	65450536	65521032
chr3	65782820	65936595
chr3	66726168	66863717
chr3	67854875	67914760

chr3	68288751	68398777
chr3	69843667	70917031
chr3	71145245	71215207
chr3	72404255	72666187
chr3	72998116	73071262
chr3	74265439	74351297
chr3	76840426	76905021
chr3	79887712	80034036
chr3	81028322	81754165
chr3	85371914	85864893
chr3	87024213	87215341
chr3	87776878	87941501
chr3	88069395	88110452
chr3	88205396	88226776
chr3	88290384	88566638
chr3	89116878	89343330
chr3	90757469	91594342
chr3	91909053	92083794
chr3	92157207	92304010
chr3	93653614	93699325
chr3	93771750	94025739
chr3	94139044	94685652
chr4	9064600	9152159
chr4	14285295	14636627
chr4	18583997	18809746

chr4	19673915	21620551
chr4	24330206	24437613
chr4	24604129	25123613
chr4	31559414	31640033
chr4	31888189	32325768
chr4	34353574	34987294
chr4	36871741	38288119
chr4	38345881	38392573
chr4	39905095	39970972
chr4	45570693	45712965
chr4	48614924	50254106
chr4	53258548	53595471
chr4	59347450	59734154
chr4	61226050	61548405
chr4	61764704	62204616
chr4	65445199	65766045
chr4	69910536	70103134
chr4	73796943	73909518
chr4	75883664	76320446
chr4	76733488	76760770
chr4	76875474	76952068
chr4	79482450	79606267
chr4	81439894	82200385
chr4	83327208	83750155
chr4	84425089	84757832

chr4	86044772	86511725
chr4	86633786	89201748
chr4	89222231	89393801
chr4	89454151	90474184
chr4	90507106	91475954
chr5	3468575	4388522
chr5	5945613	6468991
chr5	9331695	10456935
chr5	12716435	13309793
chr5	16355428	16419533
chr5	17534273	17963578
chr5	18065537	18100817
chr5	21588775	21785799
chr5	23273627	23392169
chr5	25927327	26142906
chr5	26451489	26627787
chr5	28321275	29265468
chr5	29836736	30368364
chr5	30900232	31011976
chr5	31984377	32159024
chr5	32757807	33075380
chr5	33469065	34042443
chr5	34066265	34350049
chr5	34631197	34949383
chr5	35101386	35755815

chr5	36862791	37152828
chr5	38713660	38808874
chr5	39579110	39883763
chr5	40079424	40358804
chr5	40411376	40584518
chr5	41625388	41696553
chr5	43843394	44435870
chr5	44585680	44935045
chr5	48522899	48608611
chr5	48999644	49100246
chr5	50102227	50141027
chr5	50924216	51151633
chr5	51730260	51977855
chr5	52870642	52921744
chr5	54040702	54130772
chr5	54689288	54991421
chr5	56232363	56616662
chr5	57634903	57740366
chr5	59086591	59914675
chr5	60117252	60616169
chr5	60774189	61253954
chr5	62091337	62258672
chr5	64035047	64346516
chr5	65029997	65406128
chr5	67766628	68063545

chr5	69337700	69512795
chr5	70003357	70369654
chr5	70646504	70675907
chr5	71720408	72554750
chr5	74661344	74763803
chr5	77223684	77417808
chr5	78211465	78432176
chr5	78446966	78535173
chr5	78737784	79005008
chr5	80761018	81583586
chr5	84497378	85454699
chr5	85909515	86465541
chr5	86650971	86727703
chr5	87447870	87822280
chr5	88805115	88853335
chr5	88968141	89159442
chr5	90283538	90412656
chr5	90865436	91168960
chr5	91302417	91434888
chr6	6578790	7200739
chr6	8465758	8492113
chr6	11678147	11794401
chr6	12730935	12875650
chr6	14870302	15077655
chr6	15435658	15581600

chr6	16137398	16216416
chr6	17039225	17228477
chr6	17605853	17705773
chr6	17923386	18007207
chr6	18193815	18621019
chr6	19252440	19287294
chr6	19749873	19934438
chr6	28640185	28730260
chr6	30219253	30355315
chr6	33256894	33427083
chr6	34819558	34874368
chr6	36719987	36780134
chr6	36880969	37125345
chr6	39041602	39399054
chr6	39663065	40060595
chr6	41109622	43452950
chr6	43630337	43996361
chr6	44337390	44556451
chr6	45012873	45646399
chr6	47585210	47798202
chr6	48195689	51646466
chr6	51884582	52015962
chr6	57638148	57724276
chr6	62056263	62193782
chr6	63313277	63493091

chr6	71206938	71382935
chr6	73130793	73249458
chr6	73463863	73508897
chr6	74886317	75036629
chr6	75187434	75530157
chr6	77878219	78031153
chr6	78106812	78598270
chr6	78894075	79035462
chr6	79942418	80637976
chr7	3234221	5190915
chr7	7404323	8190927
chr7	9941989	10061984
chr7	10113985	10329928
chr7	13578657	13732713
chr7	19458976	19566789
chr7	19877197	19896293
chr7	24114119	24202086
chr7	25380139	25527411
chr7	30084102	30200964
chr7	32879001	33065885
chr7	34593071	34856610
chr7	36515432	36785317
chr7	41321758	41811734
chr7	43670987	43771664
chr7	44014865	45600240

chr7	46819369	46872762
chr7	47591455	47668582
chr7	48638021	49152204
chr7	50489489	50765181
chr7	54438613	54617243
chr7	55686901	55995111
chr7	58618709	59052596
chr7	60396286	60810252
chr7	61786431	62081640
chr7	63317978	63511377
chr7	64382233	64433572
chr7	65273397	65399568
chr7	67224831	67403314
chr7	67785249	67865713
chr7	68042626	68152598
chr7	68502176	68601581
chr7	70920767	71238273
chr7	74289675	74340715
chr7	74632322	74731706
chr7	75082865	75116299
chr7	76870373	76935078
chr7	78403055	78744113
chr7	78851268	79099414
chr7	81866114	82020583
chr7	82194162	82305899

chr7	82691705	83002961
chr7	83151495	83288257
chr7	83367088	83846193
chr8	4587821	5006421
chr8	5245970	5415041
chr8	5472446	5907038
chr8	7075651	7295783
chr8	9546275	9756055
chr8	10486541	11520067
chr8	12420448	12578824
chr8	15967305	17110634
chr8	41064633	41357303
chr8	46106804	46320703
chr8	47762152	47813006
chr8	52185662	52252934
chr8	52317520	52458098
chr8	52571217	52603594
chr8	52667932	52947532
chr8	57768147	58903939
chr8	59501997	59680839
chr8	60022980	60471828
chr8	60541338	60693901
chr8	61241529	61559774
chr8	63847870	63915234
chr8	64466345	64592702

chr8	64826969	64890401
chr8	66531996	66640167
chr8	67399382	67641964
chr8	67694837	67823037
chr8	68339740	68378146
chr8	68470275	68498964
chr8	69055780	70347569
chr8	72308176	72536887
chr8	74801229	75871568
chr8	76115875	77179381
chr9	3403829	3769127
chr9	7890288	7970066
chr9	8890902	8924577
chr9	9395979	9467284
chr9	10030103	10280555
chr9	10383390	11202860
chr9	11605229	12140376
chr9	12872633	12957459
chr9	13348440	13493109
chr9	19013467	19164278
chr9	19289821	21696907
chr9	22387333	22577905
chr9	24107899	24339588
chr9	25357693	25422653
chr9	26821548	27500212

chr9	27532090	27615764
chr9	28113224	28297474
chr9	28841199	28972155
chr9	30583003	31045958
chr9	31431367	32517379
chr9	32666562	33255837
chr9	33720856	33973513
chr9	36882814	37222555
chr9	38477859	38579140
chr9	40950922	41129838
chr9	41213075	41265799
chr9	41851387	42066728
chr9	42103549	42135595
chr9	42199538	42333775
chr9	42606201	42916473
chr9	44075057	44444463
chr9	45451780	45651555
chr9	45846878	45972782
chr9	49621588	50053647
chr9	52529818	52669100
chr9	52998402	53187535
chr9	53567729	53983166
chr9	54309323	54484543
chr9	55350084	55398078
chr9	55678977	55805008

chr9	57256010	57478638
chr9	59483554	59771940
chr9	62274375	62554585
chr9	63190450	63559662
chr9	63706691	63742629
chr10	4034278	4805900
chr10	7446537	7949324
chr10	11429197	11616330
chr10	14304061	14587412
chr10	17627841	18428031
chr10	18645753	19348875
chr10	20663647	21891719
chr10	22576099	22639172
chr10	22803745	22863981
chr10	24239936	24523450
chr10	26362762	26758522
chr10	31685561	31848826
chr10	34227814	34314854
chr10	35630644	35765777
chr10	36441326	36640641
chr10	37422188	37671433
chr10	39756079	39952607
chr10	41676851	42483781
chr10	42538568	43329282
chr10	44627065	44808625

chr10	48492055	48582562
chr10	51556084	51691502
chr10	55060601	55284700
chr10	56306181	56706726
chr10	56975575	57204668
chr10	57394807	58235098
chr10	62061877	63145599
chr10	64413902	64952171
chr10	68290145	68619832
chr10	70310076	70559032
chr10	70696871	70836821
chr10	71141511	71178344
chr10	71386792	71451123
chr10	71558934	72274570
chr11	4869391	5848248
chr11	10463124	10574865
chr11	11026039	11071686
chr11	13238762	13449905
chr11	13935027	14460353
chr11	17325124	17366477
chr11	19391418	19435915
chr11	20014381	20184365
chr11	24457728	24487096
chr11	28844990	29131672
chr11	31356456	31569508

chr11	32799092	33942981
chr11	43544043	43791254
chr11	44098185	44312390
chr11	45191870	46683861
chr11	54361352	54479155
chr11	55691754	56185476
chr11	61384335	61414660
chr11	61986249	62250855
chr11	62379796	62604726
chr11	64612701	65362136
chr11	66353785	66458564
chr11	67655362	67705667
chr11	67950671	68083541
chr11	68401242	68739069
chr11	70091862	70164165
chr11	70267031	70562584
chr11	71358593	71763790
chr11	72256650	72605864
chr11	73591010	73641171
chr11	74542838	75456823
chr11	76244615	76498881
chr11	76513511	76805702
chr11	76832164	77004831
chr11	77182077	77309663
chr12	4427845	4483565

chr12	4771730	5286965
chr12	5542391	5680266
chr12	8675263	8835332
chr12	9442538	9565684
chr12	10132113	10198426
chr12	10980250	11011094
chr12	11625599	11726142
chr12	12090452	12176093
chr12	15907563	16000090
chr12	19949922	20318634
chr12	20584770	20659690
chr12	22025056	22125848
chr12	22178187	22396023
chr12	25352866	25654976
chr12	27181547	27478809
chr12	29209630	29389108
chr12	33239283	33361548
chr12	48072780	48256701
chr12	53769330	54046255
chr12	64060897	64353933
chr12	65218929	65454302
chr12	69652483	69750236
chr12	71160841	71211432
chr12	71266002	71329195
chr12	71493221	71597020

chr12	73058170	73150004
chr12	73582621	73690737
chr12	74476646	75449056
chr13	5079942	5172605
chr13	12785564	12897466
chr13	20882866	20988744
chr13	22966236	23078420
chr13	23155124	23243268
chr13	24265991	24363484
chr13	25358958	25428455
chr13	26840412	27097125
chr13	28321310	28460869
chr13	31331029	31549297
chr13	32643673	32813925
chr13	34809758	35196146
chr13	36436834	36495894
chr13	36718259	36822948
chr13	37125080	37231765
chr13	37489795	37713679
chr13	38635582	41816616
chr13	42226360	42367326
chr13	42980628	43297868
chr13	43489965	43791586
chr13	45440098	46270847
chr13	47100630	47245262

chr13	47831836	48448218
chr13	53056275	53536779
chr13	57038985	57515204
chr13	61644605	61751394
chr13	61819418	62270554
chr13	62430297	62728386
chr13	63823184	64528065
chr13	66095024	66178334
chr14	3144248	3301085
chr14	3494546	3930999
chr14	5142960	5713875
chr14	8053401	8218137
chr14	11924605	12254637
chr14	14958807	15106712
chr14	23469834	23764014
chr14	26608327	27051433
chr14	32341115	32652580
chr14	33813564	34102409
chr14	36505280	37042781
chr14	38810771	38917460
chr14	41192324	41378621
chr14	42018633	42194142
chr14	42644318	42709577
chr14	43227178	43380011
chr14	46134301	46712792

chr14	48972369	49116370
chr14	49533793	49986611
chr14	54644252	55046216
chr14	60647544	61365255
chr14	61823376	61982591
chr14	62105205	62261259
chr14	63580097	63769223
chr15	3079243	3657068
chr15	8618476	8759893
chr15	10154900	10270544
chr15	10956686	11172695
chr15	13802897	13927439
chr15	14586511	15163461
chr15	15736532	15873825
chr15	18670306	18721377
chr15	19267025	19913761
chr15	20502372	20563168
chr15	20722627	21156022
chr15	23142619	23568256
chr15	25492215	26106159
chr15	31252676	31880765
chr15	44213906	44423366
chr15	45245996	45371709
chr15	49467939	49577811
chr15	50070694	50288828

chr15	50772826	51014720
chr15	52224908	52581083
chr15	56854967	56962090
chr15	60371681	61064409
chr15	61892709	62045994
chr15	62961106	63090954
chr15	64825380	65579029
chr15	65966729	66365102
chr15	66836248	67035761
chr16	3010385	4474550
chr16	6729495	6970962
chr16	7181329	7693708
chr16	13454483	13539271
chr16	13973653	14754806
chr16	17961331	18436022
chr16	19981249	20138058
chr16	20897045	21004414
chr16	22793375	22917738
chr16	23014207	23446224
chr16	23763353	23903536
chr16	23998999	24100075
chr16	29815699	29883777
chr16	30984102	31179611
chr16	33045456	33141434
chr16	35149464	35293090

chr16	42952493	43633599
chr16	44949538	45284458
chr16	48994966	49189525
chr16	49737419	49907775
chr16	50288505	50420259
chr16	50546763	50730566
chr16	51235057	51436071
chr16	52418453	53313751
chr16	53633661	53808477
chr16	54965802	55192895
chr16	56447972	56842362
chr16	56909522	57579772
chr16	57940869	57968035
chr16	58591701	58663614
chr16	58878840	59403991
chr16	60518774	60820685
chr16	61154532	61451600
chr16	61890403	62535335
chr17	3006082	4595475
chr17	5102371	5339198
chr17	5952540	6201910
chr17	8786683	8920959
chr17	9551083	10054083
chr17	11206391	11237903
chr17	11615459	12128346

chr17	12466949	13917094
chr17	13952302	14065908
chr17	15350239	15674716
chr17	16216354	16307077
chr17	19692589	20417642
chr17	20701072	20763767
chr17	23493179	23771558
chr17	24041364	24828238
chr17	26220342	26280758
chr17	26573617	26701508
chr17	27726170	27792496
chr17	29192456	29214913
chr17	29695767	30217205
chr17	30535374	31753533
chr17	33813474	33938828
chr17	34043282	34162862
chr17	35399367	36077896
chr17	37725738	37935260
chr17	39186183	39230391
chr17	40381865	40477673
chr17	40724266	40810873
chr17	42311931	42542607
chr17	42675286	42886776
chr17	42968946	43176930
chr17	43536511	43679382

chr17	43773261	44221758
chr17	44395173	46321291
chr17	47385578	47845576
chr17	48079454	49413580
chr17	49937172	50166373
chr17	51213728	51233876
chr17	51347335	51397533
chr17	51655632	51783390
chr17	52724269	52829766
chr17	53886849	53938261
chr17	54500607	54582217
chr17	57602697	57692629
chr17	58463702	58817853
chr17	59730061	59969037
chr17	60346873	60588002
chr17	61936296	62000878
chr17	64240785	64359050
chr17	65428874	65523384
chr17	66241234	66325982
chr17	66434463	66489631
chr18	3041220	3536574
chr18	4119729	5167047
chr18	8729874	8837053
chr18	14250419	14418201
chr18	15592233	15725602

chr18	21351233	21825695
chr18	26701254	27474848
chr18	28275530	29219324
chr18	31979143	33127128
chr18	33431982	33899796
chr18	36198095	36323150
chr18	36505206	36594981
chr18	37101919	37215127
chr18	40327387	40502056
chr18	41846158	41896479
chr18	42219411	44550745
chr18	45669370	45846916
chr18	47115362	47185554
chr18	47303444	47368951
chr18	48132654	48231017
chr18	48461344	48714503
chr18	50404393	50480293
chr18	51115068	51209791
chr18	52088945	52174831
chr18	57132811	57500389
chr18	57666408	57713392
chr19	13816712	15199637
chr19	22386909	23375848
chr19	24432602	24959138
chr19	25835096	26059497

chr19	26713368	26751798
chr19	28863762	29450617
chr19	30564758	30800228
chr19	31416016	31633362
chr19	32148228	32813261
chr19	33809694	34154733
chr19	34885355	35045851
chr19	35164543	35206886
chr19	40105911	40371535
chr19	45633642	46154179
chr19	49595112	52124337
chr19	52233976	52295433
chr19	53274588	53627450
chr19	55033987	56663424
chr20	9117281	9135611
chr20	9588030	9844816
chr20	12460442	12728179
chr20	13097576	13454081
chr20	14264114	14711649
chr20	16414140	16544458
chr20	17510141	17805357
chr20	20133461	20535667
chr20	20576861	20691669
chr20	21772025	21858866
chr20	22773044	22952484

chr20	23198699	23342163
chr20	26383295	26606689
chr20	27686029	27804991
chr20	29788562	29949982
chr20	32375963	32560750
chr20	32729318	33010368
chr20	33924893	34031783
chr20	34855369	35016721
chr20	36173504	36548262
chr20	38881718	39098878
chr20	41795286	42619804
chr20	44386192	45126661
chr20	48416452	48465827
chr20	48684847	48732228
chr20	50050134	50776895
chr20	54275140	54886255
chr20	56194891	56448871
chr20	58657202	58734967
chr20	60102096	60334177
chr20	60407183	60992022
chr21	9335078	9525727
chr21	9597945	9648484
chr21	10514541	10937893
chr21	13790189	14033132
chr21	14367607	14449097

chr21	19519558	19673655
chr21	20567084	20951123
chr21	24514251	25488490
chr21	27822072	28016268
chr21	28410626	28619224
chr21	29670892	29965007
chr21	30094925	30345952
chr21	30995562	31177420
chr21	31509010	31605659
chr21	32196237	32417622
chr21	32493836	32775465
chr21	32891383	33164315
chr21	33246356	33339495
chr21	33542346	33832163
chr21	37277703	37389004
chr21	37822708	37866999
chr21	38220736	38313095
chr21	43255656	43533585
chr21	43775877	43960588
chr21	45207219	45242345
chr21	46087809	46176576
chr21	46431887	46593878
chr21	47891248	47938140
chr21	48211191	48625978
chr21	48768253	49755933

chr21	49785531	49910399
chr21	50807054	51084252
chr21	51366910	51633164
chr21	52967863	53578657
chr22	3007077	3943906
chr22	6598327	7344209
chr22	12166642	12257921
chr22	21721798	22166956
chr22	24427983	24616279
chr22	25896264	26163532
chr22	30180727	30276438
chr22	31416285	31700486
chr22	34127451	34280556
chr22	35065170	35157371
chr22	37657464	37753351
chr22	44422646	44611661
chr22	44942608	45309128
chr22	45592424	47538931
chr22	48487967	48592059
chr22	51373565	51414255
chr22	51428999	51576878
chr22	53795768	54002853
chr22	54052311	54136865
chr22	54241605	54774470
chr22	55363473	56850487

chr22	57424536	57674307
chr22	59076279	59224305
chr22	59752793	60100808
chr22	60332320	60418956
chr22	62042653	62137911
chr22	62793783	62898222
chr22	63256401	63567010
chr22	63629928	63753437
chr22	63783519	64105985
chr23	3053724	3851707
chr23	8250800	8526000
chr23	10427223	10468509
chr23	10884861	10930235
chr23	11034030	11134333
chr23	11275977	11382522
chr23	11911957	11969844
chr23	14275103	14455134
chr23	14665485	14905565
chr23	15109284	15201523
chr23	20154471	20198195
chr23	20317000	20374104
chr23	21261572	21291300
chr23	23346717	23843293
chr23	24913716	25053591
chr23	26292491	26441914

chr23	26879951	27116594
chr23	30691406	30937688
chr23	31603688	31702999
chr23	34366849	34563072
chr23	34624020	34892160
chr23	37237505	37463476
chr23	37580346	38317673
chr23	40030864	40178428
chr23	40784024	41181267
chr23	42064417	42342246
chr23	43054683	43466705
chr23	44521389	45246248
chr23	45879822	46436890
chr23	47281946	47362337
chr23	49393271	49765702
chr23	50013413	50085005
chr23	51288686	51357618
chr23	52117081	52145930
chr23	53163124	53441335
chr23	53584407	53779333
chr23	53996619	54533076
chr23	54777003	55386667
chr24	4760099	4909128
chr24	10589856	10845319
chr24	12285989	12650091

chr24	14371336	14454435
chr24	15133761	15194350
chr24	17453989	17964959
chr24	20707028	20904853
chr24	21091208	21485690
chr24	21702488	21819458
chr24	21960235	22126602
chr24	22165076	22385237
chr24	22612874	22683407
chr24	24256309	24333725
chr24	25258729	25336080
chr24	27123137	27429720
chr24	27993274	28045538
chr24	30661952	31376760
chr24	32678886	32920295
chr24	32988019	33096924
chr24	33357603	33517149
chr24	35029893	35295309
chr24	35471281	35541432
chr24	35794220	36052080
chr24	37969518	38036682
chr24	40393647	40665178
chr24	41069649	42114688
chr24	42191990	42283285
chr24	42915348	44363967

chr24	47302169	47432478
chr24	48473436	48603453
chr24	49006291	49163902
chr24	49277789	50496971
chr24	50538346	50674002
chr25	4937679	5436749
chr25	17664610	17697294
chr25	21539901	21870333
chr25	28276084	28773051
chr25	30055838	30301220
chr25	30420023	30554676
chr25	34275277	34351692
chr25	34411231	34587072
chr25	35485283	35609749
chr25	36097274	36388076
chr25	36945537	37059490
chr25	37350850	37445148
chr25	38006468	38190406
chr25	38379607	38482233
chr25	38930907	39388461
chr25	40563792	41030415
chr25	41490212	41608630
chr25	42539683	42637035
chr25	42814837	42934659
chr25	43618739	43790614

chr25	44179368	45450095
chr25	47549482	47594105
chr25	47650961	47758699
chr25	48007219	48192507
chr25	48279406	48650636
chr25	49009930	49079193
chr25	49754355	49802933
chr25	50027499	50166273
chr25	50837076	50999025
chr25	51161344	51243852
chr25	52327051	52938600
chr25	52966693	53284533
chr25	53323252	54521746
chr26	3166735	3351443
chr26	3537143	3817400
chr26	4129486	4306961
chr26	6773880	6834962
chr26	11399815	11558197
chr26	13423798	13449503
chr26	13896757	13960129
chr26	14637670	14689731
chr26	16528428	16663647
chr26	17265056	17421330
chr26	17716455	18231102
chr26	18450224	18492211

chr26	20322553	20416119
chr26	20882355	21011060
chr26	23120496	23278353
chr26	23561600	23700196
chr26	24147765	24199720
chr26	28050875	29625230
chr26	30140710	30563549
chr26	30781465	30902113
chr26	31892952	31986623
chr26	32103093	32649177
chr26	33973090	34022146
chr26	34273598	35301250
chr26	36248963	36774426
chr26	36827011	38077301
chr26	38590936	38738016
chr26	39188777	39226014
chr26	39515488	39718698
chr26	39761211	39954040
chr26	40097691	40151706
chr26	40462488	40635972
chr26	40659910	41533632
chr26	41601609	42028847
chr27	3086744	3684579
chr27	3859098	4010312
chr27	4199565	4250017

chr27	5025775	5062532
chr27	5591240	6001370
chr27	6492668	6540877
chr27	8872961	9125602
chr27	10343730	11196529
chr27	12224742	13119085
chr27	14740636	14854832
chr27	16130851	16198146
chr27	16766852	16841117
chr27	20201757	20676216
chr27	20878003	21009117
chr27	21167562	21238258
chr27	23056973	23271715
chr27	25514211	25809306
chr27	28652158	29096851
chr27	29631928	30173995
chr27	31951133	32778283
chr27	34336319	34679285
chr27	36109992	36291138
chr27	38824009	39129733
chr27	39318470	39479352
chr27	41105122	41195227
chr27	42332019	42540534
chr27	42699806	42791283
chr27	43465072	43536319

chr27	44210145	44331052
chr27	44502846	44736829
chr28	7044809	7151158
chr28	13629655	13923622
chr28	16010091	16117743
chr28	18892729	19064898
chr28	19339287	19376692
chr28	20785974	20866099
chr28	21995853	22561751
chr28	23001743	23705808
chr28	25547359	25620884
chr28	26240910	26349544
chr28	27268202	27310165
chr28	28338587	28486945
chr28	30860993	30919260
chr28	32487119	32572151
chr28	32706241	32776839
chr28	33121346	33177040
chr28	33913997	34261735
chr28	34377820	34705148
chr28	35168097	35281443
chr28	36088039	36135740
chr28	37975525	38104148
chr28	38571632	38860154
chr28	39903758	39962017

chr28	40103280	40178642
chr28	40577023	40646186
chr28	41461296	41590174
chr28	41791991	41911238
chr28	42165934	44033718
chr29	3958430	4295720
chr29	5185841	5283441
chr29	12237042	12420976
chr29	13423069	13746822
chr29	14743948	14886217
chr29	15949774	16059500
chr29	33050738	33576030
chr29	34768806	35701646
chr29	36352549	36582590
chr29	37204875	38201243
chr29	38922706	39071373
chr29	39119213	39242840
chr29	40637462	40808128
chr29	41919539	42090632
chr29	42465194	42877623
chr29	43177857	43271484
chr29	43588053	44129870
chr29	44193774	44452829
chr29	44478203	44800834
chr30	3039134	3418567

chr30	10413042	10619011
chr30	11850900	11916565
chr30	12589617	13107473
chr30	14474944	14599955
chr30	20845096	20909246
chr30	22817609	23059168
chr30	23636058	23738504
chr30	26351795	26466548
chr30	26623223	26820667
chr30	29060448	29370860
chr30	29721068	29953421
chr30	32432717	32758169
chr30	32927897	33226084
chr30	33300458	33342985
chr30	34375560	34821353
chr30	34917249	35056824
chr30	35209400	35267567
chr30	36037141	36286571
chr30	37858940	38029193
chr30	38714870	38738805
chr30	39808327	39885678
chr30	40460317	40508425
chr30	41175193	41818921
chr30	42184951	42290735
chr30	42465862	42976413

chr31	3660095	3930319
chr31	7930334	8142729
chr31	13980620	14267780
chr31	25185661	25311197
chr31	25796839	26192171
chr31	31947118	31989288
chr31	32456899	33483405
chr31	35971237	36081536
chr31	36316907	36658658
chr31	36718277	37587958
chr31	38903908	39050938
chr31	39614935	39970415
chr31	40006624	40695677
chr31	40728571	41335815
chr31	41450471	42049353
chr32	4281517	4406614
chr32	5577416	5693112
chr32	7627931	8046491
chr32	8980689	9048180
chr32	11651272	11945028
chr32	15399825	15512879
chr32	22615744	22742778
chr32	25188094	25375410
chr32	25921407	26109744
chr32	37768701	37939968

chr33	4486060	4649454
chr33	5596268	5925334
chr33	7772702	7970376
chr33	12743985	13127338
chr33	15694668	15771306
chr33	17142952	17184148
chr33	19681116	19776517
chr33	20226334	20300948
chr33	24392969	25513695
chr33	29180973	29266494
chr33	29425341	29529980
chr33	30039766	30268764
chr33	31072023	31190288
chr33	33769855	34102599
chr33	34363342	34409279
chr34	3476963	3687765
chr34	4260696	4447848
chr34	5032377	5352711
chr34	5499476	5956971
chr34	6547298	6605742
chr34	8401649	8548456
chr34	9593487	9715137
chr34	11403205	11464021
chr34	12683127	12730289
chr34	12867591	13137372

chr34	13837055	15005162
chr34	15230146	15312872
chr34	15324529	15471706
chr34	17307459	18100683
chr34	20401537	20501964
chr34	23309642	23387687
chr34	23727055	23839963
chr34	23984161	24237687
chr34	24562414	24610384
chr34	25330686	25516293
chr34	29788673	29893432
chr34	31514110	31841296
chr34	40388041	40644703
chr34	41286862	42712484
chr34	42745871	43221465
chr34	43508716	44182239
chr34	44356864	44598810
chr34	44724184	45121405
chr35	3548671	3846573
chr35	5042242	5184062
chr35	6232867	6688975
chr35	6898081	6935733
chr35	9110905	9865645
chr35	11302403	11808894
chr35	12153431	12318775

chr35	14254740	14289086
chr35	15080226	15215486
chr35	15481247	15571438
chr35	15939113	16009178
chr35	18546715	18658713
chr35	21038330	21179912
chr35	22906154	23028919
chr35	23645553	23971947
chr35	24210838	24580369
chr35	25218866	25309630
chr35	26054030	26095369
chr35	26995373	27222895
chr35	27325035	27863993
chr35	27982088	28137852
chr35	28669587	28968604
chr35	29187147	29461891
chr36	5686317	5846159
chr36	7849659	8034757
chr36	11947852	11998373
chr36	12406317	12502872
chr36	14040409	14224083
chr36	14320729	14414092
chr36	15245087	15578850
chr36	20891035	21152297
chr36	21660209	21747306

chr36	26677350	26775203
chr36	28858905	29119226
chr36	29549446	30191218
chr36	31658855	32062735
chr36	32967193	33227006
chr37	4894511	5041965
chr37	5168289	5473110
chr37	5866368	6500622
chr37	9787512	9870623
chr37	10298282	10560008
chr37	17283032	17368107
chr37	18574707	18625007
chr37	18769135	19063906
chr37	24476577	24871454
chr37	26272451	26331094
chr37	26685718	26771681
chr37	27534629	28002709
chr37	28047122	28077540
chr37	28728159	29150616
chr37	29709322	30336170
chr37	30582674	30844014
chr37	31492477	31715228
chr37	32093487	32281477
chr37	32784527	32880476
chr37	32911512	33912299

chr38	3679415	3746841
chr38	5385644	5715013
chr38	7008358	7141343
chr38	8788104	9356728
chr38	11813014	12280212
chr38	14153647	14450739
chr38	15002328	15282886
chr38	16298810	16481923
chr38	17191712	17438025
chr38	18347613	18386084
chr38	19239552	19476185
chr38	20020613	20226729
chr38	21052956	21192923
chr38	21606263	21672589
chr38	21939266	22403817
chr38	22466974	22501435
chr38	24114476	24225358
chr38	24367427	24416393
chr38	24467415	24673851
chr38	24803437	24856192
chr38	24980096	25385494
chr38	25695180	25873154
chr38	26174694	26243192
chr38	26276476	26369854

Supplementary Table 2 for Chapter 1: Comparison of CNVRs found in this study to those found in [16]

Found in Nicholas et al.	
chr1 106813764 107437468:	chr1 107164760 107437965
chr1 108115387 108421887:	chr1 108112502 108121036
chr1 108811454 109048520:	chr1 108886700 108914537
chr1 78046272 78238933:	chr1 78070108 78136431
chr2 20381004 20453714:	chr2 20398148 20435022
chr2 20943172 21258104:	chr2 21069053 21221596
chr2 25927754 26198302:	chr2 26027172 26094460
chr2 68140245 68248458:	chr2 68153519 68181928
chr2 7219945 7266621:	chr2 7237277 7265972
chr2 7496646 8465313:	chr2 8366495 8390861
	chr2 8393772 8617358
chr2 85763679 85855229:	chr2 85849001 858666057
chr2 86673963 87043658:	chr2 86817008 86921689
chr2 8708478 8932343:	chr2 8695906 8892258
chr3 3023852 3620587:	chr3 3040595 3049185
chr3 34166254 34403807:	chr3 34152139 34196885
chr3 38435577 38607091:	chr3 38575956 38595036
chr4 61226050 61548405:	chr4 61446171 61458512
chr5 32757807 33075380:	chr5 33066557 33078734
chr6 12730935 12875650:	chr6 12792094 12850099
chr6 14870302 15077655:	chr6 15060023 15088875
chr6 18749873 19934438:	chr6 19839036 19851104
chr6 43630337 43996361:	chr6 43744090 43797557
chr6 48195689 51646466:	chr6 48887041 48905088
	chr6 49227049 49243137
chr7 43670987 43771664:	chr7 43721579 43769166
chr8 5245970 5415041:	chr8 5244689 5434539
chr8 76115875 77179381:	chr8 76367315 76467801
	chr8 76094443 76151150
chr9 10383390 11202860:	chr9 10480818 10544330
chr9 11605229 12140376:	chr9 11621390 11942236
chr9 13348440 13493109:	chr9 13350206 13421806
chr9 19289821 21696907:	chr9 19778694 21163725
chr10 18645753 19348875:	chr10 19064933 19091725
chr11 13935027 14460353:	chr11 14069249 14422490
chr11 43544043 43791254:	chr11 43768087 43798409
chr13 42980628 43297868:	chr13 43294192 43308680
chr13 61819418 62270554:	chr13 61852461 61965024
chr13 62430297 62728386:	chr13 62433325 62464200
chr13 66095024 66178334:	chr13 66073148 66139454
chr14 3144248 3301085:	chr14 3150487 3322175
chr14 3494546 3930999:	chr14 3674038 3694049
chr14 5142960 5713875:	chr14 5505315 5525495
chr14 60647544 61365255:	chr14 61274056 61317155
chr15 3070243 3657068:	chr15 3125718 3138191
chr15 64825380 65579029:	chr15 65430993 65453339
chr16 13454483 13539271:	chr16 13529829 13575105
chr16 3010385 4474550:	chr16 3006293 3066927
chr16 42952493 43633599:	chr16 43007825 43024039
	chr1 78160227 78171123
	chr1 78219715 78227981
	chr1 78236424 78246568
	chr2 21246857 21257108
	chr2 7555428 7625158
	chr2 7482453 7553252
	chr2 86694359 86711448
	chr3 34214040 34304448
	chr3 34315592 34327763
	chr3 34358825 34396403
	chr2 7627163 7953160
	chr2 8349554 8357828
	chr2 7964911 7996934
	chr2 8073467 8337211
	chr6 48259041 48274087
	chr6 48548088 48577093
	chr6 49903020 49927329
	chr6 49996040 50017058
	chr6 48814001 48832083
	chr6 50125017 50147061
	chr8 76770775 76859468
	chr8 76892039 77122974
	chr8 76581052 76644640
	chr8 76715004 76759616
	chr9 10414074 10429161
	chr9 11954868 12057245
	chr9 13464877 13511710
	chr9 21234054 21356447
	chr10 19144391 19154541
	chr11 13935449 13972387
	chr13 61826098 61841475
	chr15 3185301 3194098
	chr15 3213083 3227669
	chr16 4305763 4355649
	chr16 4127072 4137266
	chr16 3171749 3190481
	chr16 4223224 4297478

Found this study
chr16 53633661 53808477: chr16 53623086 53803291
chr16 56909522 57579772: chr16 56995747 57102014
chr16 61890403 62533335: chr16 61961098 62098430
chr17 24041364 24828238: chr17 24435613 24794917
chr17 40724266 40810873: chr17 40744453 40752749
chr17 42311931 42542607: chr17 42389425 42422753
chr17 59730061 59969037: chr17 59787027 59917864
chr17 60346873 60588002: chr17 60456246 60485652
chr18 14250419 14418201: chr18 14303066 14430292
chr18 21351233 21825695: chr18 21343918 21415061
chr18 28275530 28219324: chr18 28306681 28319889
chr18 40327387 40502056: chr18 40410081 40477496
chr18 42219411 44550745: chr18 43366575 4387347
chr18 52088945 52174831: chr18 43436181 43490874
chr18 57132811 57500389: chr18 44271503 442888316
chr18 57666408 57713392: chr18 57303991 57366216
chr19 22386909 23375848: chr19 22405293 22418508
chr19 24432602 24959138: chr19 23239124 23267756
chr19 55033987 56663424: chr19 24465564 24477427
chr20 56194891 56448871: chr20 56255024 56299307
chr21 30094925 30345952: chr21 30191025 30318483
chr21 30995562 31177420: chr21 31086051 31157768
chr21 31509010 31605659: chr21 31554665 31604430
chr21 32196237 32417622: chr21 32299229 32347238
chr21 33246356 33339495: chr21 33272908 33293267
chr21 43775877 43960588: chr21 43621051 43837413
chr21 49785531 49910399: chr21 49786030 49821631
chr22 55635473 56850487: chr22 56695076 56734245
chr23 37237505 37463476: chr23 37344596 37352850
chr24 22165076 22385237: chr24 22230078 22379453
chr25 48007219 48192507: chr25 48056062 48113154
chr25 52966693 53284533: chr25 53005035 53015342
chr26 28050875 29625230: chr26 28067007 28466000
chr26 29452147 29486818: chr26 29148411 29282576
chr26 30140710 30563549: chr26 29452147 29486818
chr26 31892952 31986623: chr26 30374046 30603765
chr26 34273598 35301250: chr26 31895098 31981553
chr27 16130851 16198146: chr27 16141865 16207653
chr27 28652158 29096851: chr27 28699444 28708618
chr27 8872961 9125602: chr27 9032592 9046764
chr28 16010091 16117743: chr28 16085013 16110991
chr28 35168097 35281443: chr28 35223069 35268034
chr29 44478203 44800834: chr29 44768817 44783743
chr30 40460317 40508425: chr30 40469105 40487810
chr31 32458699 33483405: chr31 33437272 33475555
chr31 40006624 40695677: chr31 40623054 40657249
chr34 44724184 45121405: chr34 45099216 45109344
chr38 24114476 24225358: chr38 24089004 24116103
chr38 26174694 26243192: chr38 26215871 26251030

Found in Nicholas et al.
chr16 62196518 62380994 chr16 62407316 62569296
chr17 40778397 40796995
chr17 59722864 59763864
chr17 60456246 60485652
chr18 14246032 14260858
chr18 21453014 21566771
chr18 28447219 28457365
chr18 43497817 43524058
chr18 43743063 43768874
chr18 44366215 44396872
chr19 22889981 23016866 chr19 22440119 22505098 chr19 22840008 22858355 chr19 23108981 23148128
chr21 30165320 30175764 chr21 30342946 30353280
chr21 43858119 43894969
chr25 53119029 53135059 chr25 53157213 53196829 chr25 53266470 53288387
chr26 29355078 29377975 chr26 28479061 28822424 chr26 28832700 28849656 chr26 28857934 28984356
chr26 28994395 29027572 chr26 29031916 29091517 chr26 29301987 29314460 chr26 29319685 29338395
chr26 29516268 29532788 chr26 28915510 28926842 chr27 8993491 9003754
chr29 44785954 44822901