1.0 Introduction

Genetic variation is a central topic in Evolutionary Biology. The most hotly discussed form at present is the vast amount of naturally occurring structural (i.e. over 1kb) variation. Copy number variations (CNVs) are the most abundant, diverse, and well-studied class of structural variation. Over the past five years, facilitated largely by the establishment of new resources and technologies, CNVs have come under a great deal of scrutiny. Despite many descriptive and functional studies in primates and mice, their significance to macro-evolution is only now being understood; and their impact on micro-evolutionary processes has not been addressed.

One of the most well studied mammalian models in micro-evolution are the various subspecies of the common house mouse, *Mus musculus*. This model system lends itself well to the study of genetic incompatibilities underlying reproductive isolation between genetically similar subspecies. Reproductive isolation figures prominently in Evolutionary Biology for its role in the process of speciation. This thesis makes an examination of CNVs in hybrids of two partially reproductively isolated *Mus musculus ssps.* What emerges is a unique and unexpected finding relevant to both Evolutionary Biology and our growing knowledge of CNVs. Here, I begin with an introduction to the *Mus musculus* model system and proceed to review the relevant literature regarding CNVs before focusing on the specific items addressed in this thesis.

1.1 *Mus musculus ssps.*: A Model for Evolutionary Genetics

1.1.1 The Origins of *Mus musculus ssps.*

Mus musculus is familiar to most biologists as a model organism in biomedical research. Most laboratory strains are actually hybrid compositions of three naturally-occurring and distinct subspecies: *Mus musculus domesticus*, *M. m. musculus* and *M. m. castaneus*. (Frazer et al., 2007; Yang et al., 2007). Their origins have been traced to modern-day Northern India, having diverged approximately 1 million years ago (MYA) (Guénet and Bonhomme, 2003) with *M. m. domesticus* and *M. m. musculus* as recent as <500, 000 years ago (Salcedo et al., 2007). Distinct geographic ranges have been described: *M. m. domesticus* in Western Europe, Northern Africa and the near East; *M. m. musculus* in Eastern Europe and Northern Asia; and *M. m. castaneus* throughout

South-East Asia (Fig. 1). Several points of secondary contact, or hybrid zones, have been described, the most well studied are between *M. m. domesticus* and *M. m. musculus* in Europe and between *M. m. musculus* and *M. m. castaneus* in Japan, where a stable hybrid subspecies, *M. m. molossinus*, persists (Yonekawa et al., 1988).

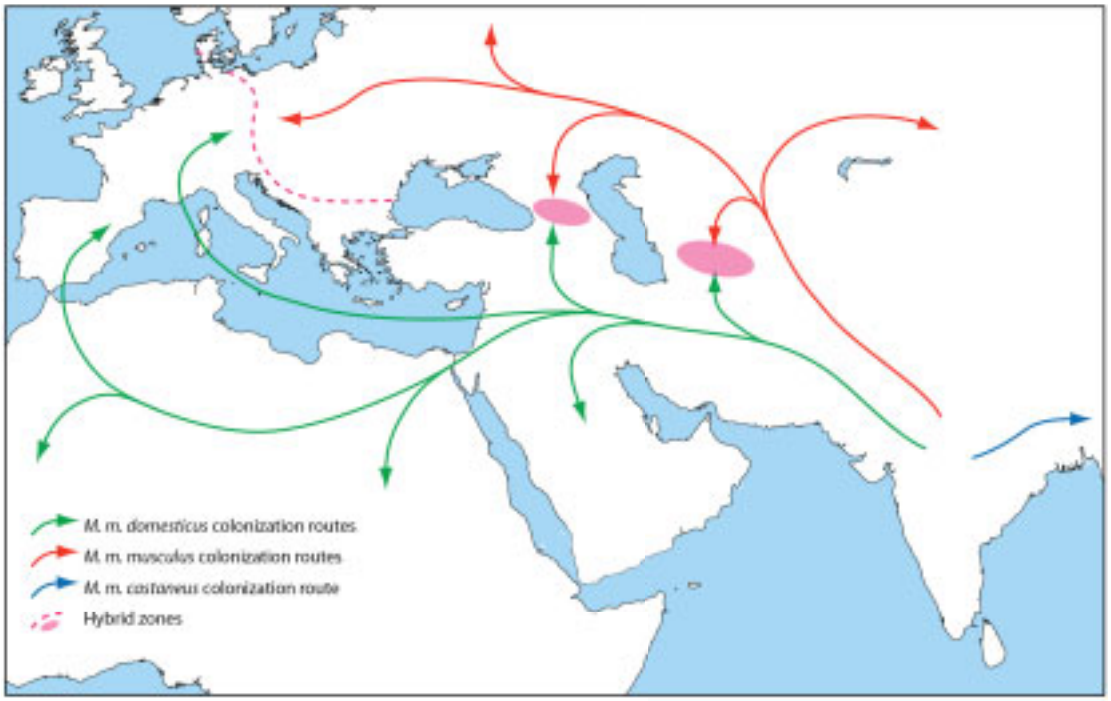

Figure 1. Geographic Distribution of *M. m. musculus ssps.*
Mus musculus subspecies originated in Northern India, diverging about 1MYA. *M. m. domesticus* traveled westward through the Fertile Crescent and the Mediterranean Basin into Western Europe and Northern Africa. *M. m. musculus* traveled northward, migrating to Northern Asia and Eastern Europe. *M. m. castaneus* traveled eastward and can be found in South-East Asia. Magenta areas highlight hybrid zones, points of secondary contact between the two sub-species. The most well studied hybrid zone runs from the Jutland peninsula in Denmark through Germany and onto the Black Sea. Although several transects have been well studied, the exact border of the entire hybrid zone is still not entirely resolved. (Figure based on Guénet and Bonhomme 2003).

Given the drive for genetic homogeneity in inbred laboratory mouse strains, the value of genetically diverse wild-derived populations of *Mus musculus* cannot be understated. Outbred stocks have already proven themselves useful in refined QTL analysis and evolutionary studies (Chia et al., 2005; Guénet and Bonhomme, 2003). It is clear that the growing interest in genetic variation (including CNVs) will also benefit by taking advantage of wild mouse resources.

1.2 A Portrait of Copy Number Variation

1.2.1 The Genomic Landscape of Copy Number Variation

In the past five years, analyses of genetic variation in humans and mouse have identified extensive, naturally occurring CNVs as a common form of structural genetic variation (Conrad et al., 2006; Cutler et al., 2007; Graubert et al., 2007; Iafrate et al., 2004; Kidd et al., 2008; Li et al., 2004; McCarroll et al., 2006; Perry et al., 2008b; Redon et al., 2006; Sebat et al., 2004; She et al., 2008; Snijders et al., 2005; Tuzun et al., 2005; Watkins-Chow and Pavan, 2008). CNVs are genetic loci 1Kb or greater that are present as a variable copy number compared to a reference genome, possibly encompassing genes or influencing surrounding gene expression (Freeman et al., 2006; Stranger et al., 2007). The most important discoveries to come from these studies are: i) CNVs are remarkably abundant, even in presumably healthy individuals; ii) CNV loci range in size from 1kb to more than 1Mb and can overlap; iii) Mutation rates at some CNV loci can be quite high; iv) CNVs can distinguish species and populations; v) CNVs can encompass genes or influence gene expression of surrounding genes; vi) Genes broadly defined as acting at the molecular-environment interface are overrepresented in CNVs; and vii) Most CNVs arise as byproducts of ineffective recombination. The major studies that have lead to this current portrait of CNVs are described below.

The first two comprehensive reports of human CNVs appeared in 2004 (Iafrate et al., 2004; Sebat et al., 2004). These were the first studies to analyze genomic DNA of presumably healthy humans by array comparative genome hybridization (aCGH). This method involves differentially labeling reference and experimental genomic DNA with fluorescent dyes. The DNA samples are pooled together and hybridized to a microarray chip containing any variety of DNA probes (Pinkel and Albertson, 2005a; Pinkel and Albertson, 2005b). Amplifications and deletions are then represented as the log2 ratio of experimental signal intensity to the reference signal intensity. Both studies identified dozens of CNV loci, having an enriched association with segmental duplications (SDs, duplicated loci > 1kb with over 90% sequence similarity).

Other studies focusing on deletions (Conrad et al., 2006; McCarroll et al., 2006) discovered that genic markers are strongly underrepresented in deletions. However, of genes encompassed by deletions, those involved in immunity and defense, sensory perception, cell adhesion and signal transduction were overrepresented. These are among

the first reports which suggest that CNVs have a functional impact and are under some form of selection.

Large-scale population-based CNV detection studies have also been undertaken (Redon et al 2006). Using 270 individuals from the International HapMap Project (The International Consortium, 2003), a staggering 1447 CNV loci, covering 12% of the genome, were discovered. Over half of these loci overlap with RefSeq genes. Overrepresented gene classes include cell adhesion, sensory perception of smell and chemical stimulus and neurophysiological processes. Genes associated with cell signaling, proliferation, kinases and other phosphorylation-related categories were under-represented. This study also showed that individuals within a population cluster on the basis of diallelic CNVs.

Paired-end sequencing is the most sensitive CNV detection technique. In this approach, both ends of a fosmid (genomic DNA clone of approximately 40kb) are sequenced and mapped to a reference genome. Consistent discrepancies in the expected versus mapped clone size reveals insertions and deletions in the test sample. Studies using this technique reveal that individuals can have several hundered CNVs, mostly between 10-50kb (Kidd et al., 2008; Perry et al., 2008a; Tuzun et al., 2005). More than half of these CNV loci map to segmental duplications, which only represent 5% of the genome (Tuzun et al 2005; She 2004, Bailey et al 2002). Of the genes encompassed by CNVs, a general trend of molecular-environmental interaction is observed: including drug detoxification, innate immune response and inflammation, surface integrity, and surface antigens (Tuzun et al., 2005). Large gene families are also overrepresented in CNV loci (Kidd et al., 2008). Once again, this suggests an important functional aspect of CNV loci and hints at an involvement in adaptive evolution.

CNVs have also been well characterized in inbred mouse strains. Similar to Human studies, CNVs are both abundant, and associated with SDs (Adams et al., 2005; Graubert et al., 2007; Li et al., 2004; Snijders et al., 2005). Compared to the reference sequence (C57Bl/6 strain), mice strains contain an average of 51 CNV loci, accounting for 10Mb of DNA (Cutler et al., 2007). The evolutionary divergence of laboratory strains likely accounts for the greater number (over 2000 loci) and larger average size (over 180kb) of CNVs in mice compared to humans (Cutler et al., 2007; Graubert et al., 2007; She et al., 2008; Yang et al., 2007). Like humans, SDs represent approximately

5% of the mouse genome (She et al., 2008). This most recent figure is a two- to three-fold increase over previous estimates, suggesting that associations between CNVs and SDs, although already significant, may have even been previously underestimated.

There are several indicators of the functional importance of CNVs in mice. For instance, intergenic regions are overrepresented in deletions and stable genomic regions are enriched for genes with no or few paralogs, in contrast to large multigene families strongly enriched in CNVs (Cutler et al., 2007). Once again, this links functional redundancy to dynamic regions of the genome. Furthermore, similar types of genes appear to be enriched in mouse CNVs as in humans: pheromone binding, antigen binding, antigen presentation by MHC class I receptors, defense response and steroid processing genes, receptor activity, signal transduction, carbohydrate binding, resonse to stimulus and G protein-coupled receptors (Cutler et al., 2007; Graubert et al., 2007). Those genes enriched in stable genomic regions are more likely to be involved in basic cellular processes such as nucleotide binding, protein folding and cell cycle regulation, also similar to what has been observed in humans (Cutler et al., 2007).

These thorough descriptive studies in humans and mice have ignited a new appreciation for CNVs as a major source of genetic variation. It is with this solid foundation that studies can move into the functional arena.

1.2.2 Consequences of Copy Number Variation

Structural variation is clearly abundant, however it also has a significant functional aspect. Consequences of CNVs have been studied in relation to their contribution to disease, adaptive evolution and effects on gene expression.

In humans, the most noteworthy outcome of CNV research has come in the identification of rare and *de novo* CNVs. These typically large deletions often encompass only a single gene and are associated with autism, schizophrenia and mental retardation (de Vries et al., 2005; Jacquemont et al., 2006; Marshall et al., 2008; Sebat et al., 2007; Walsh et al., 2008). This offers a new perspective on the etiology of these complex trait diseases that contrasts with the widely accepted "common disease-common allele" model where disease is the result of the modest contribution from combinations of several common alleles.