Genomic Destabilization at Copy Number Variable Loci in Intersubspecific Hybrids of *Mus musculus ssps.*

Inaugural - Dissertation

Zur
Erlangung des Doktorgrades
Der Mathematisch-Naturwissenschaftlichen Fakultät
Der Universität zu Köln

Vorgelegt von
Rick J. Scavetta
aus Mississauga, Ontario, Kanada
Köln, 2009

Berichterstatter: **Prof. Dr. Diethard Tautz**

Prof. Dr. Maria Leptin

Tag der letzten mündlichen Prüfung: **--**

This work is dedicated to my parents,
who, knowing they could not follow it,
helped me tread my own path in this world.

Iñaki Echevarne, Bar Giardinetto, Calle Granada del Penedés, Barcelona, July 1994. For a while, Criticism travels side by side with the Work, then Criticism vanishes and it's the Readers who keep pace. The journey may be long or short. Then the Readers die one by one and the Work continues on alone, although a new Criticism and new Readers gradually fall into step with it along its path. Then Criticism dies again and the Readers die again and the Work passes over a trail of bones on its journey toward solitude. To come near the work, to sail in her wake, is a sign of certain death, but new Criticism and new Readers approach her tirelessly and relentlessly and are devoured by time and speed. Finally the Work journeys irremediably alone in the Great Vastness. And one day the Work dies, as all things must die and come to an end: the Sun and the Earth and the Solar System and the Galaxy and the farthest reaches of man's memory. Everything that begins as comedy ends as tragedy.

-Roberto Belaño, The Savage Detectives

Table of Contents

List of Tables and Figures

Tables

Figures

List of Abbreviations

BER	Base-excision repair
CNV	Copy number variation (or variant)
C_t	Threshold cycle
DSB	Double-stranded break
E	Embryonic day
EC	Endogenous control
HR	Homologous recombination
MMEJ	Mismatch repair end-joining
NAHR	Non-allelic homologous recombination
NHEJ	Non-homoogous end-joining
PSV	Paralogous Sequence variation (of variant)
RC-SSBR	Recombination-coupled single-stranded break repair
ROS	Reactive oxygen species
SCE	Sister-chromatid exchange
SD	Segmental duplication
SSB	Single-stranded break

Acknowledgments

Many people have helped me to complete my PhD over the past three years:

First and formost, I must acknowledge the support and insight of my supervisor Prof. Dr. Diethard Tautz. I know that this work would never have been possible in another laboratory. Diethard has consistently supported me in this project; allowed me to express my choices in direction; and supervised me by giving me enough freedom to explore.

At the graduate school in Cologne, I extend a deep-felt gratitude to Dr. Brigitte v.Wilcken-Bergmann. Brigitte has gone far beyond her administrative duties and has consistently been helpful in dealing with all things beauacracy related- of which there are scores. She has, from my first day in Germany, made transitioning to a new culture, city, language, school and laboratory much more plesant.

I also thank my parents, who, despite seldom seeing me, have given me the opportunity to pursue this path in my life and fulfill my scientific curiosity.

Thank go to my friends, inside and outside the laboratory, who made everyday life more enjoyable. They always reminded me of the beautiful things that exist outside the laboratory, and in so doing gave me the energy to continue with this project even when I sometimes had doubt in it myself.

I thank Barbara Kleinhenz, who helped greatly in the transition to Plön, and was a large technical help in the laboratory. Members of the Milinski group who also aided me in transitioning to Plön were Gisela Schmiedeskamp and later on Helga Luttmann.

My colleagues, many of whom cycled between supporter and adversary with an almost Circadian-like rhythm, also deserve thanks. I benefited from the discussions that helped to propel my work forward.

Brigitte Lechner, the librarian here in Plön who helped me by quickly obtaining even the most obscure articles also deserves special mention.

I also acknowledge Arne Nolte and Jarek Bryk, who added thoughtful comments to a draft version of this text.

Acknowledgement and thanks goto the International Graduate School in Genetics and Functional Genomics in Cologne and the Max Planck Gesselschaft.

Zusammenfassung

Variation der Kopienzahl von Genen ist eine wichtige Quelle genetischer Variation innerhalb und zwischen Populationen. Die Mutationsmechanismen die zur Variation der Kopienzahl führen, sowie die Prozesse die die Grösse der betreffenden Regionen regulieren sind wenig untersucht. Diese Arbeit behandelt Variation der Kopienzahl in X und Y chromosomalen Mitgliedern einer grossen Genfamilie in *Mus musculus ssps.* Eine dramatisch erhöhte Amplifikation der Kopienzahl in intersubspezifischen Hybriden zwischen *M. m. domesticus* and *M. m. musculus* wird beschrieben. Dieses Phänomen wird sowohl in natürlichen als auch bei im Labor gezüchteten Hybriden beobachtet. Eine extreme Amplifikation der Kopienzahl, die in Hybriden aus der Natur nicht nachgewiesen wird, kann unter Laborbedingungen generiert werden. Dies legt nahe, dass extreme Destabilisierung der Kopienzahl in der Natur durch Selektion verhindert wird. Spezifische Analysen in Hybridmännchen zeigen das weder meiotische Rekombination oder interchromosomale Austauschprozesse benötigt werden, um Variation in der Kopienzahl zu erzeugen. Damit scheinen besonders Intrachromosomale- (Schwesterchromatid-) Austausche in intersubspezifischen Kreuzungen aufzutreten. Belegt wird dies durch eine grössere Anzahl somatischer Variationen in der Kopienzahl in verschiedenen Organen von Hybriden im Vergleich zu reinerbigen Mäusen. In Hybriden korreliert dies mit Fehlregulation der DNA Reparaturprozesse die Schwesterchromatid Austausche regulieren. Es scheint, das die Stabilität der Kopienenzahl von Genen in reinerbigen Populationen durch Kreuzungen mit Tieren aus anderen Populationen herabgesetzt werden kann, und dass dieser Prozess mit Mutationsprozessen zusammen hängt, die während der Entwicklung ablaufen. Dieses Ergebnis eröffnet eine neue Perspektive auf reproduktive Isolation und könnte für den Aufbau genetischer Inkompatibilität zwischen Unterarten von Mäusen eine Rolle spielen.

Abstract

Copy number variation (CNV) contributes significantly to natural genetic variation within and between populations. However, the mutational mechanisms leading to copy number variation, as well as the processes that control the size of CNV regions are so far not well understood. This thesis deals with CNVs containing X- and Y-linked members of a large gene family in *Mus musculus ssps*. The phenomenon that CNV regions show dramatic copy number amplifications in intersubspecific hybrids of *M. m. domesticus* and *M. m. musculus* is described. This is observed in natural and laboratory-bred hybrids. Extreme copy number amplification, not found in wild-caught hybrids, can be generated under laboratory conditions, suggesting that there is a selection against this CNV destabilization phenomenon in the wild. Specific analysis of hybrid males indicates that neither meiotic recombination nor inter-chromosomal exchange is required for this to occur, suggesting intrachromosomal (i.e. sister chromatid) exchange that can occur at an elevated frequency in intersubspecific crosses. As confirmation, I can detect a greater number of somatic CNVs between organs in hybrid individuals than pure-breds and disruptions in DNA repair pathways known to regulate sister chromatid exchange also appear to be misregulated in some hybrids. It appears that the relative stability of CNV loci in pure-breeding populations can be disrupted in crosses with animals from another population, and this relies on mutational mechanisms acting during development. This finding offers a unique perspective on reproductive isolation and may be important for understanding the build-up of genetic incompatibilities between these subspecies.

Declaration

- Meike Thomas provided the genomic DNA samples for the wild mice pure-bred individuals.
- Chris Voolstra provided raw microarray data and made the original suggestion to use qPCR to assess genomic copy number for *Slx*.
- Ruth Rottscheidt provided DNA and tissue samples of wild hybrid mice and their laboratory-born offspring which she caught and bred as part of her thesis.
- Tina Harr provided useful genotype information of the wild hybrids.
- Kerstin Musolf of the Konrad Lorenz-Institute for Ethnology in Vienna very graciously provided *M. m. musculus* mice and tissue samples.
- Birgit Schmitz helped in picking colonies for Slx cDNA sequencing.
- Barbara Kleinhenz extracted DNA and RNA samples from laboratory-bred hybrids and greatly facilitated my move to the new laboratory facilities in Plön. Barbara also assisted with dissection on occasion.
- Heike Harre was of great assistance in aiding in most of the embryo dissections and also dissections of the remaining laboratory-bred hybrid mice not used in this study. Alexandra Müller also aided with embryo dissections on occasion.
- Maren Volquardsen, under the guidance of Christine Pfeifle cared for my mice after to move to Plön. In Cologne several student helpers had that responsibility.
- Christine Pfeifle was also of great help in obtainin and maintaining mice. She also provided necessary guidance on mouse care techniques.
- Bernhard Haubolt provided an awk script to perform dot-plots on large sequences.
- Arne Nolte translated the abstract.

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).

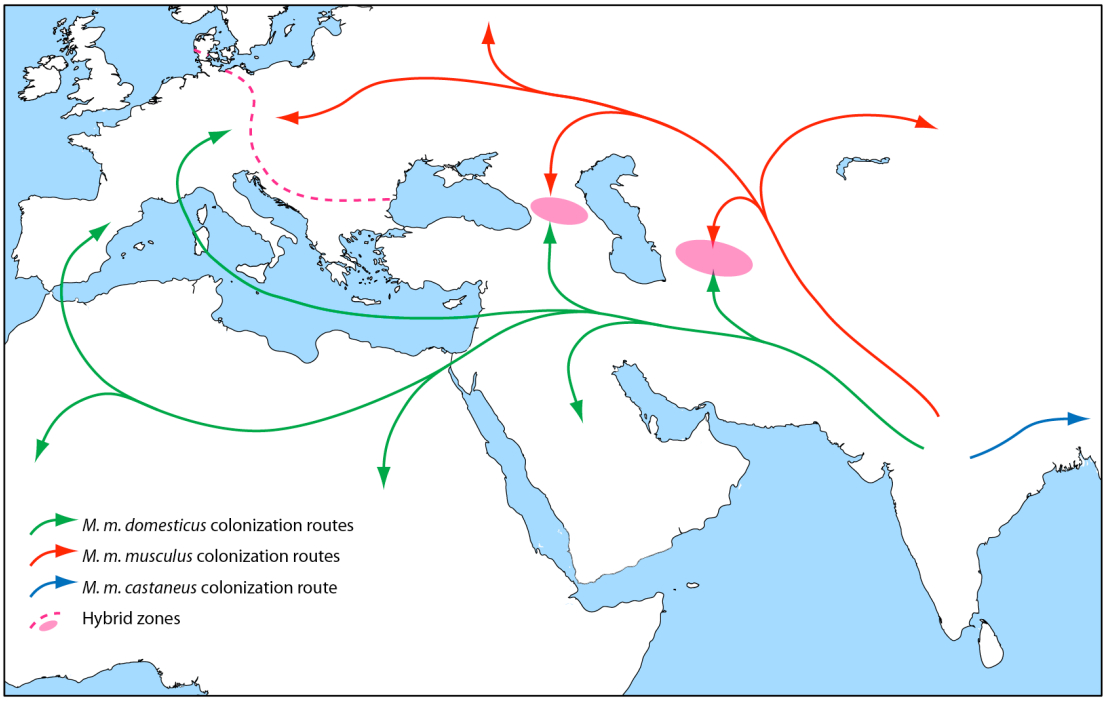


Figure 1. Geographic Distribution of *M. m. musculus ssps.*
Mus musculus subspecies originated in Northern India, diverging about 1MYA. *M. m. domesticus* traveled westward through the Fertile Crescent and the Mediterranean Basin into Western Europe and Northern Africa. *M. m. musculus* traveled northward, 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 has 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 incredibly 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). Amplification 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 affects 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.

Common polymorphisms also have functional significance, with the most dramatic being the association of HIV-resistance with higher copy number of the cytokine CCL3L1 (Gonzalez et al., 2005). CCL3L1 is the most potent ligand known for the CC chemokine receptor 5 (CCR5), the major coreceptor for HIV. Individuals with low CCL3L1 copy number are overrepresented in HIV-positive versus HIV-negative patients (Gonzalez et al., 2005), suggesting competition between the chemokine and HIV for the CCR5 receptor. Further, rhesus macaques experimentally infected with simian-AIDS showed a negative correlation between higher copy number and progression rate of the disease (Degenhardt et al., 2009). Interestingly, CCR5 is a pro-inflammatory chemokine receptor and higher copy number of the CCL3L1 chemokine is associated with the autoimmune diseases type I diabetes and rheumatoid arthritis in Caucasians (McKinney et al., 2007). The same study showed that the adverse effects of high CCL3L1 copy number are offset in those patients who have a dysfunctional CCR5 allele.

There are also examples of adaptive evolution at CNV loci. A correlation between higher copy number of the salivary amylase gene *AMY1* with populations having high starch content diets (Perry et al., 2007) has been observed. Although *AMY1* copy number was the most common polymorphism identified in one of the earliest genome-wide CNV assays (Iafrate et al., 2004), this is largely regarded as the first example of an adaptive CNV. A more elegant study was recently published examining drug resistance in the malaria parasite *Plasmodium falciparum* (Nair et al., 2008). Antimalarial drugs targeting downstream effectors of the folate biosysthesis pathway, used as a first line of defense in Thailand, are associated with a higher (i.e. compensating) copy number of the upstream activator *gch1*. In Laos, where antifolate drugs are the second line of defense, if used at all, lower copy number of *gch1* persists. This study is notable in that adaptive copy number changes in response to strong selective pressure was observed to spread throughout the population quite quickly – antifolate drugs were used in Thailand for only 10 years prior to the study.

These studies also signal a subtle but significant shift in thinking among micro-evolutionary biologists. Previously, a duplicated locus was of interest primarily as a source of genetic redundancy, leading to neo- or sub-functionalization, the classic scenario first proposed by Ohno in 1970 (Reviewed in Cañestro et al. 2007). However, with the realization that the copy number of a gene is itself an allele with phenotypic consequences subject to selection, a new dimension of complexity in duplicated regions is appreciated.

A testament to the functional significance of CNVs is their contribution to variation in gene expression. A comprehensive study assessd the contribution of SNPs and CNVs to the expression of almost 15,000 ESTs from the HapMap dataset (Stranger et al., 2007). Although most (over 80%) of expression variation can be attributed to SNPs, almost 18% is associated with CNVs, with little overlap between the two.

1.2.3 The Origin of Copy Number Variations

Several attempts have been made to describe the origin of CNVs in humans and mice, implementing three distinct approaches: i) Characterizing known or *de novo* CNV loci in well-defined pedigrees to determine mutation rates at specific loci; ii) Sequencing CNV breakpoints in an attempt to uncover footprints of known DNA rearrangement mechanisms; and iii) Comparing primate genomes, with the goal of identifying lineage specific rearrangements and hotspots of CNV formation. I will discuss each of these approaches in turn, as they will help to understand what is currently known about CNV mutation dynamics.

In mice, the genealogy of the C57Bl/6 strain is well documented, with representative substrains spanning ~967 generations of divergence (Egan et al., 2007). This unique resource allows the mutation rate of recurrent CNV loci to be estimated. Of 38 newly arisen CNVs, the mutation rate varies by four orders of magnitude, with some loci being as high as 10^{-2} or roughly 1 mutation event in every 100 newborns. Significantly, three recurrent CNVs were large (2-4Mb) loci with tandemly arrayed genes. In a confirmation that some loci can have unusually high mutation rates, another study surveyed several inbreed mice obtained from Jackson Laboratories. Within these

mice, two CNV loci were detected, suggesting that even isogenic mouse strains maintain segregating variation (Watkins-Chow and Pavan, 2008). Putting this information together with genome wide surveys of CNVs, a clear picture can be drawn of how this form of genetic variation relates to other well-studied varieties (Fig. 2).

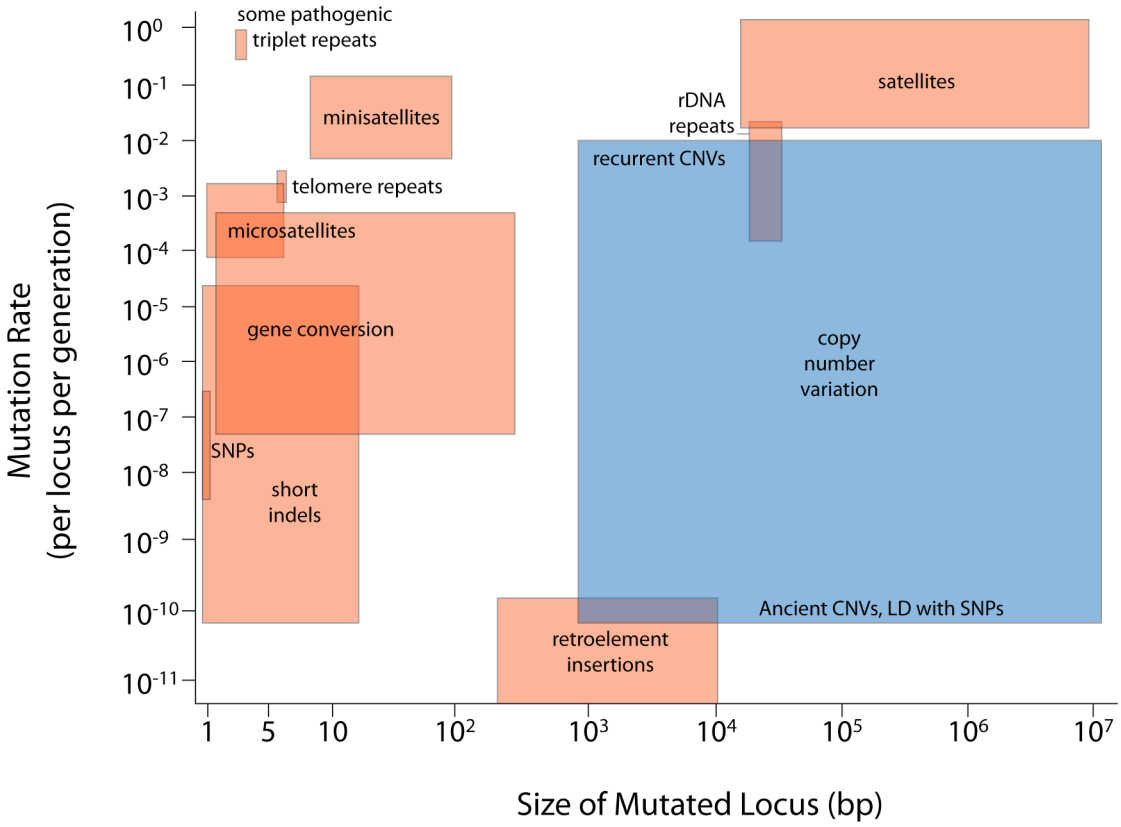


Figure 2. The Landscape of Genetic Variation

Any locus larger than 1kb with a different copy number measured against a reference sample is considered a CNV (blue block). This broad definition means CNVs cover a substantial size range although the mean size of a CNV is approximately 40Kb. Note that there can be population differences in mean sizes, even among humans (Conrad et al., 2006). CNV loci with the highest mutation rates are often tandemly duplicated repeats and mutation is likely facilitated by NAHR (see text). CNVs with very low mutation rate, represent evolutionarily ancestral structural variation and have high linkage disequilibrium with surrounding SNPs. For example, fixed segmental duplications between primate species fall into this category, essentially bringing the mutation rate down to zero. Inversions in general and indels between 100 and 1000bp are, due to technological limitations, the most poorly characterized form of genetic variation and are not depicted in this figure. Figure adapted from Freeman et al. (2006).

Two mechanistically distinct pathways are known to result in the generation of structural variation. The first, non-allelic homologous recombination (NAHR), is similar to homologous recombination (HR), except the invading strand does not insert at the matching (allelic) site. When NAHR occurs during meiosis, copy number remains unchanged but one chromosome carries a deletion and the other, a duplication (Inoue

and Lupski, 2002). The second, non-homologous end joining (NHEJ), is a process well studied for its role in v(d)j locus rearrangement as part of the immune response. Another end-joining sub-pathway, microhomology-mediated end joining (MMEJ), which relies on very short regions of homology (typically 2-8nt), also seems to be involved (McVey and Lee, 2008). All three are DNA repair pathways, but only NAHR relies on matching broken DNA ends with an homologous template and, importantly, has the potential to increase copy number; both end-joining pathways lead to deletions or copy-neutral rearrangements. Each pathway leaves a specific signature at the sequence level. Repetitive elements (e.g. *Alu* elements, Long and Short Interspersed Repeats) and long stretches of homology between breakpoints (e.g. SDs) are the hallmarks of NAHR. Of the two end-joining pathways NHEJ is distinguished by the presence of usually short (1-4nt) indels at the breakpoint, although they can occasionally be much longer, compared to the microhomology (2-8nt) of MMEJ. The informative nucleotide signatures of each mechanism have been used to determine their contribution to CNV formation by sequence analysis. Unfortunately, results are heavily biased on the data set used and can contradict each other. For instance, deletions are easier to detect, and the lack of probes representing deletions in the reference genome means that many amplifications in test subjects are not observed unless fosmid paired-end sequencing is performed (Perry et al., 2008a). Therefore, there is a general a bias towards detection of deletion loci which could be generated by NHEJ and MMEJ. Despite this bias an association between SDs and CNVs is a recurring theme.

Two studies have specifically addressed the association between SDs and CNVs in humans and mice (Sharp et al., 2005; She et al., 2008). To do this, custom SD-enriched aCGH chips were designed. In both studies a several-fold enrichment of CNVs is observed in SDs. This implies that CNVs areise by NAHR in these regions, although it is clear there are CNVs not associated with SDs.

If segmental duplications are truly CNV hotspots, one would expect them to cluster together. In humans, at least, this appears to be the case; with SD distribution following a power law (Kim et al., 2008). The age of an SD can be determined by the sequence similarity between its paralogs, older SDs having more time to accumulate variation. *Alu* elements are most closely associated with older SDs and their presence

drops off sharply with younger SDs (Kim et al., 2008). Small, common (i.e. ancestral) deletions, also have a strong association with *Alu* elements (de Smith et al., 2008). Kim et al. (2008) also found that SDs of a similar age clustered together. Alltogether, this suggests that there is a certain group of SDs likely to have been formed by an *Alu*-mediated mechanism, separate from younger SDs, and that *Alu*-mediated NAHR was most common ~40mya, during a burst of *Alu* activity and has since declined.

Kim et al. (2008) also attribute most CNV formation to end-joining mechanisms, but their arguments are not entirely convincing. They admit that their approach is biased against NAHR events because they bias against repeat rich sequences, and conclude that 40% of breakpoints show microhomology (i.e. MMEJ) and 14% show microinsertions (i.e. NHEJ). However, a detailed description of microhomology nor its statistical significance, is provided. Therefore, given the repeated finding of an association between CNVs and SDs, and that only NAHR would result in amplification, it can be concluded that NAHR plays a pivotal role in CNV formation and that end-joining processes are involved, but to a lesser extent. This also indirectly implies that what we observe as CNVs are mostly tandem arrays of repeating loci and not scattered repeats across the genome, two scenarios not distinguishable by aCGH. Notably, loci which have been confirmed by fibre FISH, where individual strands of DNA are hybridized to labeled probes, confirms that many CNVs are tandem repeats (Perry et al., 2007; Redon et al., 2006).

Comparing structural variation profiles between divergent primate species also aids in understanding the origin of CNVs. One study used a custom human cDNA aCGH platform to survey copy number changes in homologous genes in 9 other primate species encompassing ~60 million years of divergence (Dumas et al., 2007). The most striking finding is that gene duplications permeate the primate lineage. This result diretly addresses the concern that sequence divergence can bias results in multi-species hybridization-based assays, where an overabundance of deletions would be a predicted artifact. CNV loci seemed to cluster together into what have been described as gene nurseries, or hot spots of structural variation (e.g. humans genes amplifications concentrated in pericentromeric regions, which were previously identified as dynamic areas of the genome). Interstingly, genes which vary in copy number between primates

overlap considerably with disease-causing genes that are prone to genomic destabilization (Dumas et al., 2007). This finding suggests that even some evolutionarily ancient CNVs maintain a high mutation rate. Additionally, chimpanzees, our most closely-related sister species, maintain intraspecific CNV loci that are orthologous to human intraspecific CNV loci and also associate strongly with intrachromosomal SDs (Perry et al., 2008b). This correlation in CNV hotspots strongly implicates inherent sequence cues that permit certain loci to be especially plastic, implying that many CNVs are the result of a specific mechanism of genome rearrangement. Additionally, CNV loci likely to be under positive or purifying selection contain genes involved in inflammatory response and cell proliferation (Perry et al., 2006).

In summary, the current portrait of CNVs clearly positions this form of genetic variation in a pivotal role in many areas of Biology. Evolutionary Biology is one discipline that concerns itself heavily with genetic variation, not only allelic distribution but increasingly in functional aspects. One consequence of genetic variation in particular is the evolution of reproductive isolation, long considered a hallmark of true species. The role of CNVs in reproductive isolation has so far not been considered and it is to that topic, as it concerns *Mus musculus ssps.*, that I now turn.

1.3 Genetic Divergence and Reproductive Isolation

1.3.1 Speciation Genetics of *Mus musculus ssps.*

The recent divergence of the three *Mus musculus* subspecies offers the opportunity to investigate early stages of speciation. Incompatibilities between certain genes, which prevent complete admixture between these subspecies, are likely to represent causes, rather than effects, of speciation. These incompatibilities essentially establish barriers to reproduction and are termed speciation genes. Using this model system, it is possible to identify speciation genes and the biological processes affected during the early stages of speciation.

The most productive results in this line of research come from an intensive research program carried out over thirty years by the Forejt laboratory in Prague. Two projects aimed to identify QTL loci associated with hybrid sterility on the autosomes

(Forejt and Iványi, 1974) and the X-chromosome (Storchová et al., 2004) have focused attention on two loci, *Hst1* and *Hstx1,* respectively. The causative gene at *Hst1*, *Prdm9*, was recently reported; it is the first example of a mammalian speciation gene (Mihola et al., 2009). *Prdm9* is a histone 3 lysine 4 trimethyltransferase expressed in ovaries and testes (Hayashi et al., 2005). In both *Prdm9* null mutants and sterile hybrids, pachytene stage spermatocytes lack sex bodies (X-Y bivalents) and exhibit patches of γ-H2AX, the phosphorylated form of histone H2AX, over the synaptonemal complexes. The resulting sperm cells are malformed and the mouse is sterile. The role of *Prdm9* in epigenetics is also of note, as this form of genetic regulation has been largely overlooked in this context.

Another approach to uncovering putative speciation genes has made extensive use of the naturally occurring *musculus-domesticus* hybrid zone in Europe. Genes which underscore reproductive isolation display characteristic geographic gradients of allele frequency (termed clines) across the hybrid zone (Harrison, 1993). For example, consider a fixed polymorphism between *M. m. domesticus* (e.g. always A) and *M. m. musculus* (e.g. always C). Neutral loci are permitted a large amount of introgression and display shallow clines with long tails. However, incompatible loci (i.e. under selection) cannot introgress as much, resulting in a relatively steep cline. Steep clines for X-linked markers have been observed at several transects in the European hybrid zone (Dod et al., 1993; Macholán et al., 2007; Payseur et al., 2004; Tucker et al., 1992). The X-chromosome is of particular interest: it has as a higher rate of evolution compared to autosomes (Stevenson et al., 2007) and is enriched for genes involved in murine spermatogenesis (Wang et al., 2001). In a related approach, a study conducted in our laboratory computationally assayed highly differentiated regions of the genome between the two subspecies (Harr, 2006). Using this approach, the most strongly differentiated region of the X-chromosome mapped in the same area as previous studies using wild mice. Significantly, several of the above studies draw attention to the *Hstx1* locus, but the causative gene remains unidentified.

Divergent gene expression, an indicator of underlying genetic differentiation, is also useful in identifying possible genetic incompatibilities. In another study conducted in our laboratory (Voolstra et al., 2007), expression levels of genes expressed in the brain,

liver/kidney and testes of *M. m. domesticus*, *M. m. musculus*, *M. m. castaneus*, and an unresolved *Mus musculus ssp* population were surveyed by microarrays. Most divergent expression was encompassed in the metabolic organs, suggesting that sexual selection does not drive the early stages of speciation, but rather, is an important aspect of latter stages. There are a surprisingly small number of genes, 23, showing some form of divergent gonadal expression among the four groups surveyed.

Despite the small number of divergently expressed gonadal genes, my interests in reproductive isolation led me to consider them further. Of these genes, only two, *Sycp3-like, X-linked* (*Slx*) and the very closely related gene 4930*27E24Rik* (a paralog of *AK015913*, aka *Slx-like*), are both X-linked and differentially expressed between *M. m. domesticus* and *M. m. musculus*. These two qualities strongly implicate these genes as putative genetic incompatibilities in reproduction. As outlined below, the few studies of *Slx* and *Slx-like* have been complicated by the fact that these genes are part of a large multi-copy, tandemly-duplicated gene family. The high-copy number of these genes posed the enticing possibility that they could be an intersubspecific CNV and prompted a detailed course of study.

1.3.2 The *Xlr* Superfamily

Slx is a member of a the large multicopy *Xlr* superfamily (Escalier et al., 1999). *Xlr*, the founding X-linked member, expresses the pM1 transcript. Many paralogous copies of *Xlr* were originally recognized, but proposed to be pseudogenes (Cohen et al., 1985a; Garchon et al., 1989), and the true nature of those genes remains unresolved. *Xlr* is specifically expressed during the first wave of T-cell development, prior to T-cell receptor locus rearrangement (embryonic day (E) 14-15), but diminishes rapidly thereafter to be maintained in a small but likely functionally significant number of thymus cells (Escalier et al., 1999). *Slx* was originally discovered in nuclei of primary murine spermatocytes by Northern hybridization using a pM1 probe (Calenda et al., 1994). pM1 and the *Slx* transcript have abundant similarity but maintain gene-specific regions (Fig. 3) (Calenda et al., 1994). Despite sequence similarity to *Xlr*, *Slx* is named after the more closely related autosomal gene *Sycp3*, a well known member of the meiotic

synaptomeal complex which aids in pairing homologous chromosomes and the sex body (Dobson et al. 1994; Reynard et al. 2007). Originally, the temporal expression pattern and sub-cellular localization of SLX also suggested a role in sex chromosome condensation and silencing during spermatogenesis. SLX is concentrated around the asynapsed regions of the sex body and in the closely associated nucleolus (Calenda 1994, Escalier 2000, Escalier and Garchon 2005, Fernandez-Capetillo 2003). However the most recent reports, using newly generated and highly specific antibodies show that nuclear localization was falsely determined and SLX is actually localized to the cytoplasm of spermatids (Reynard et al., 2007). Reynard et al. (2007) also determined that SLX has no nuclear localization signal domains. However, just what was previously detected in the nucleus remains an interesting, unanswered question. One possibility is that it could be an even more distantly related member of the *Xlr* superfamily. The same study surveyed *Slx-like* expression, which except for appearing several days earlier in development, seems to mimic *Slx*. *Slx* is also one of the few high copy-number genes to be expressed after meiotic sex chromosome inactivation (Mueller et al., 2008). Interestingly, *Xlr*, but not *Slx* is expressed in prophase I stage oocytes (Escalier et al., 2002), providing evidence for molecular differentiation during meiosis.

The apparently highly regulated expression pattern of *Slx* strongly suggests a role in spermatogenesis. Unfortunately, due to complexities of working with high copy number genes, no functional descriptions have been reported thus far. However, an interesting observation can be made regarding its role in spermatogenesis of hybrid animals. In studies by the Garchon laboratory in Paris (Escalier and Garchon, 2005; Escalier and Garchon, 2000), SLX deposition on the sex body is disrupted in *H2AX*$^{-/-}$ spermatocytes. H2AX, which follows a similar expression profile as SLX, invades the nucleus after being phosphorylated (γ-H2AX) and is necessary for sex body formation (i.e. sex chromosome condensation). In H2AX$^{-/-}$ mutants, the sex body fails to form and SLX remains unlocalized. Decompartmentalized γ-H2AX is observed in *Prdm9*$^{-/-}$ mutants and sterile hybrids carrying incompatible *Prdm9* alleles (Mihola et al., 2009). Thus, one could imagine a chain of events connecting SLX with the only known speciation gene. However, one must be reminded that reports of SLX localizing to the sex body have been convincingly refuted (Reynard et al., 2007). Nonetheless, it is clear

that a gene product with some amount of sequence similarity to SLX is present there. Obviously, the protein composition of the sex body is complex and unresolved, but the growing number of observations centered around the sex body demands some attention.

The Y-linked multicopy gene *Sly* (*Sycp3 like, Y-linked*) is another member of the *Xlr* superfamily and is also transcribed in the testes (Ellis et al., 2005; Touré et al., 2005). *Sly* has 48% and 46% amino acid identity with *Slx* and *Xlr*, respectively (Touré et al., 2005). It appears that *Slx* and *Slx-like* expression is restricted by *Sly,* as *Slx* transcription increases in the presence of large Y chromosome deletions, encompassing the *Sly* locus (Ellis et al., 2005). A Y-linked member of the gene family is of particular interest given the evidence that the Y chromosome contributes to reproductive isolation (Dod et al., 1993; Tucker et al., 1992; Vanlerberghe et al., 1986) in *Mus musculus ssps.*, and has signatures of positive and recent selection including reduced genetic variation (Boissinot and Boursot, 1997).

Slx, Slx-like and, by extension, *Sly* are of interest in an evolutionary context for several reasons. In hybrid mice, compromised immune function and sperm head deformations associated with sterility are genetically determined (Derothe et al., 2004; Moulia et al., 1993; Moulia et al., 1995; Sage et al., 1986; Storchová et al., 2004). *Slx* is expressed in both cell types involved in these phenotypes and *Xlr* is linked to a compromised immune system phenotype (Cohen et al., 1985a; Cohen et al., 1985b). The presence of a multicopy gene family (Garchon et al., 1989) differentially expressed between subspecies (Voolstra et al., 2007) suggests that differences in copy number can lead to genetic incompatibilities in hybrid animals. This would be the first example that a CNV could contribute to speciation via reproductive isolation and is an interesting and unexplored extension of the observed stable, population-specific CNVs, discussed above.

1.4 The Aim and Relevance of this study

In this study I describe an examination of three sex-linked and one autosomal CNV loci using various methods. I will show that the previously unreported and unexpected phenomenon of CNV destabilization occurs in hybrid animals, even as F1 progeny of intersubspecific crosses. Furthermore, an analysis of DNA repair genes during

development enables me to make the first important steps in elucidating a plausible mechanistic pathway for their destabilization.

This thesis is relevant to the ongoing research of reproductive isolation. It represents a phenomenon, and potential isolating mechanism, never before considered. Additionally, it is of interest to studies in copy-number variation for purposes of understanding mutation dynamics involved in this important form of genetic variation. The relevance of these findings will be appreciated as CNV research moves into the fields of somatic variation and micro-evolutionary biology.

2.0 CNV Destabilization *Mus musculus ssp.* Hybrids

2.1 The *Slx* Gene Family

2.1.1 Genomic Architecture

It was clear from a review of the literature that *Slx* and *Slx-like* are multicopy genes. I began with an examination of the genome architecture surrounding these genes on the X-chromosome, using the mouse reference genome sequence. This genome is based on the C57Bl/6 strain, which contains a *M. m. domesticus* X-chromosome. There are several copies of *Slx* annotated on the X-chromosome and in the same region there is an uncharacterized, unrelated gene, *E330016L19Rik (*hereafter *L19),* which was originally identified as a Riken full-length cDNA clone in adult ovaries (Fig. 3A)(Carninci 2005). The *L19* expression pattern is not well described, but it does appear to be in the same tissue as *Xlr* (NCBI UniGene EST Profile Viewer: Mm.335706; Escalier 2002). Due to its expression pattern and proximity to *Slx*, *L19* was also considered for further analysis. Interestingly, I found several copies of *Slx-like* and *Xlr* at a second region approximately 20Mb downstream from the proximal *Slx*/*L19* region. Besides genomic location, major differences between *Slx* and *Slx-like* include the partial or complete loss of exons I, III, IV and VIII, plus small insertions and sequence divergence.

Due to the complexity at these loci and unreliability of annotations, I downloaded BAC scaffolds surrounding the two loci, to accurately determine the genetic architecture for these genes. I also downloaded the genomic regions associated with the RefSeq entries for *Slx (22.8kb)*, *L19 (49.2kb)*, *Slx-like (16.8kb)*, and *Xlr* (13.9kb) (see Genbank files in attached electronic documents). Dot plots were used to compare each genic region against the proximal and distal loci (summarized in Fig. 3A). By this analysis 43 copies of *Slx* and 12 copies of *L19* were identified in a proximal region spanning 9.2Mb, and in the 2.4Mb distal region 16 copies of *Slx-like* flanked by two copies of *Xlr* are present. *Slx* and *L19* were not present in the distal region and *Slx-like* and *Xlr* were not found in the proximal. There was no discernable duplication pattern for *Slx* and *L19*, with many incomplete copies and inversions present (Fig. 3A). *Slx-like* appears to be a tandemly duplicated gene, with all copies in the same orientation. The *Slx-like* and *Xlr*-containing distal region maps in the vicinity of *Hstx1*, the hybrid sterility QTL locus on

the X-chromosome (Storchová et al., 2004) (Fig. 3A), which only added to my interest in these genes as putative genetic incompatibilities between *M. m. domesticus* and *M. m. musculus*.

Unfortunately due to complications with sequence assembly, the Y chromosome remains unresolved and a detailed map could not be generated for the Y-linked *Sly*. Because this region was implicated in being involved in regulating expression of *Slx* and *Slx-like* (Ellis et al., 2005), and because of the sequence similarity between *Slx* and *Sly* (Fig. 3C), it was also included in downstream analyses.

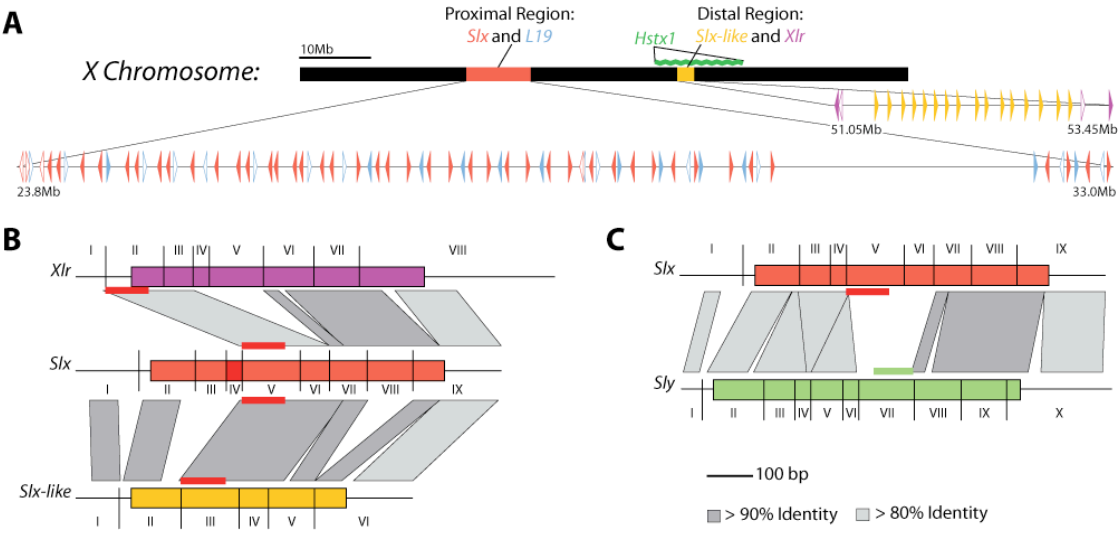


Figure 3. The Genomic Architecture of the *Xlr* Superfamily

A) Chromosome sketch and repeat structure based on dot plot analysis. The proximal locus contains *Slx* and *L19*. It covers 9.2Mb (from 23.8Mb - 33.0Mb) and contains approximately 43 copies of *Slx* (22.8Kb, red triangles) interspersed with 12 copies of *L19* (49.2Kb, blue triangles) in the reference mouse genome build 37.1. Note that the arrangement includes tandem and inverse copies and that some copies are incomplete (open triangles). The distal region lies 21Mb downstream, spanning 2.4Mb, and contains 16 copies of *Slx-like* (16.8Kb, yellow triangles) flanked by two copies of *Xlr* (13.9Kb, purple triangles). The hybrid sterility locus *Hstx1* maps in the general region of the distal cluster (green line), but may not be the cluster itself. Only part of the proximal X-chromosome is diagramed. **B)** Similarity between the *Slx* transcript and its two X-linked paralogs *Xlr* and *Slx-like*. Dark grey boxes represent over 90% sequence identity, light grey boxes over 80%; exons are annotated with roman numerals. The red bar denotes the region that was used for the copy number assays (qPCR and Southern blotting) on genomic DNA. **C)** Comparison of the *Slx* transcript with its Y-linked paralog *Sly*. Exons III and IV of *Slx* are duplicated in *Sly*. The *Sly* qPCR assay (green line), is targeted to *Sly* exon VII and does not share homology with *Slx*.

2.1.2 Variation in Slx cDNA

Given the overlap with *Hstx1* and the divergent expression levels (Voolstra et al., 2007), it was important to understand the qualitative differences in expressed copies of *Slx* and *Slx-like* in the testes of wild mice. I used a 3' rapid amplification of cDNA ends

(3' RACE) to amplify *Slx* and *Slx-like* from an RNA extraction of one *M. m. domesticus* and one *M. m. musculus* individual originally used in the detection of expression divergence (Voolstra et al., 2007). Amplification was conducted with a high fidelity *taq* to reduce the artificial introduction of polymorphisms. (Fig 4).

Ninety-six clones from each animal were sequenced. For *M. m. domesticus*, 22/96 clones matched the *Slx* reference cDNA sequence and 34/96 matched *Slx-like*. In both instances a perfect match to the reference sequence was identified, confirming the accuracy of the technique, as the reference sequence X-chromosome is *M. m. domesticus* derived. In *M. m. musculus* 20/96 clones matched *Slx*, comparable to *M. m. domesticus*, and 11/96 clones matched *Slx-like*, only a third of what was found in *M. m. domesticus*. Although this assay can only be considered as semi-quantitative, it is unexpected that so few *Slx-like* clones were found in *M. m. musculus*, given considering the higher expression in this animal. However, it seems that the assay has not reached saturation, considering that most transcripts are only represented by one clone.

The most surprising feature of the cDNA analysis is the discovery of a unique *M. m. domesticus*-specific *Slx* transcript group. A 2bp deletion is present shortly before the stop codon in many of the *M. m. domesticus Slx* transcripts. This results in a frame shift and a predicted elongation of 14 amino acids (Fig 4). I refer to this *Slx* variant as *Slx2* to distinguish it from the canonical *Slx* transcript, *Slx1*. *Slx1* and *Slx2* contain many paralogous sequence variants (PSVs, i.e. SNPs between paralogously-duplicated loci as opposed to homologous loci on different chromosomes), indicating that the split between the two groups is quite ancient. The retention of many *Slx2* copies suggests a functional role. The absence of *Slx2* in *M. m. musculus* cDNA sequences indicates that either this duplication occurred after the divergence of the two sub-species or was subsequently lost in *M. m. musculus*. *Slx1* transcripts of both sub-species have an abundant amount of non-synonomous PSVs (Fig. 4). Significantly there is not a single *Slx* transcript in common between these two individuals. This could be a result of unsaturation in the assay, or a reflection of divergence between the two subspecies.

Slx Predicted Protein Sequences:

Slx-l Predicted Protein Sequences:

Figure 4. Subspecies specific *Slx* and *Slx-like* alleles. See text page 21

Slx-like is also a highly variable gene in both sub-species. *M. m. domesticus* appears to have a single isoform in high abundance (21 copies), a feature not present in *M. m. musculus*. As with *Slx*, there does not appear to be any overlap between the two sub-species, but the two most similar protein sequences differ by only 1 amino acid substitution.

2.2 CNV Destabilizations in Intersubspecific Hybrids

2.2.1 CNVs in Wild Populations

Animals from natural populations of *M. m. domesticus* and *M. m. musculus* were surveyed for copy number of *Slx*, *L19* and *Sly* using custom designed TaqMan qPCR assays. For the *Xlr* superfamily, one assay targets the least polymorphic genomic region which overlaps with *Slx* exon V, *Slx-like* exon III and *Xlr* exon II (Fig. 3B). I will refer to these targets collectively as *Slx*. Because *L19* is not present in the distal cluster, it is used as a representative of the proximal cluster. The *Sly* gene cluster on the Y-chromosome is targeted with a diagnostic qPCR assay in exon VII (Fig. 3C). qPCR requires gene expression to be measured against an endogenous control (EC). For this purpose, I chose a single copy X-linked gene, *etd*, situated between the two *Xlr* superfamily clusters. Therefore, there is always a 1:1 ratio between a sex chromosome and the EC. Twelve wild-caught individuals each from French (MC), German (D), Czech Republic (CR) and Kazakhastan (AL) populations were assayed for copy number. The French and German mice represent *M. m. domesticus* populations; the Czech and Kasak, *M. m. musculus* (Ihle et al., 2006). I found a significant increase in mean copy number for *Slx* and *Sly* in *M. m. musculus*, but not for *L19* (Table 1 & 2, Fig. 5 & 6A). There is no significant difference in any intrasubspecific population comparisons.

Figure 4. Subspecies specific *Slx* and *Slx-like* alleles.
Gonadal mRNA was amplified by 3' RACE. A unique *M. m. domesticus Slx* variant (*Slx2*) was identified. There is no overlap in alleles between the two subspecies for either gene and both genes containg many non-synonymous PSVs.

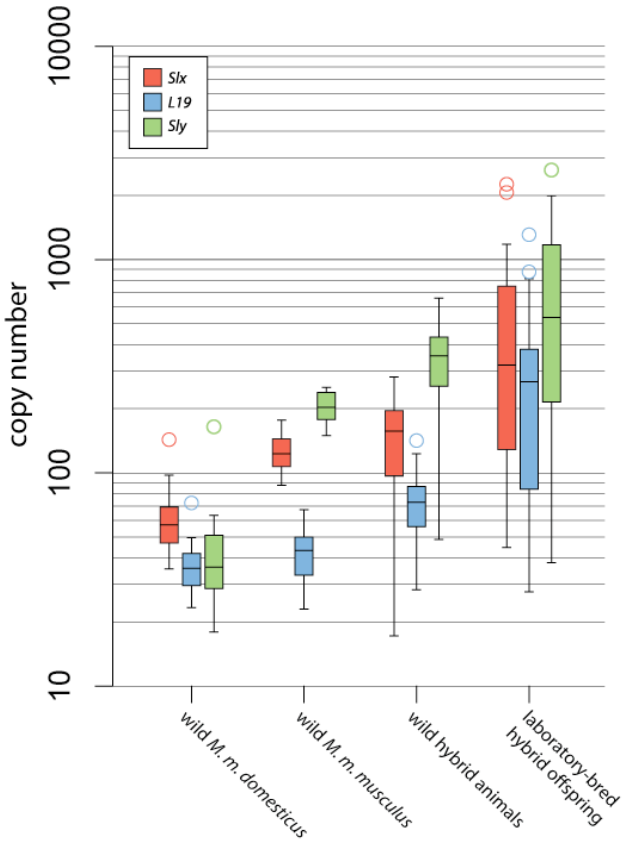


Figure 5. Copy Number Variation in Pure and Hybrid Individuals.
qPCR was used to assay copy number in 24 unrelated wild-caught *M. m. domesticus* and *M. m. musculus* individuals from two populations each (see text). Box plots show the median and the inter-quartile range (IQR), with outliers (>±1.5x IQR) marked as circles. A total of 39 animals that were directly caught in the hybrid zone close to Munich were analysed in the same way. These show significantly higher mean copy numbers and higher variances than the parental populations (see Table 1 for individual fold change values, and Table 2 for significance analyses). A subset of the wild caught animals was used for setting up breeding pairs in the laboratory and 39 offspring from 7 hybrid animal mating pairs were also analysed. These show even higher means and variances than the wild hybrids (see Tables 2 & 3). Note the logarithmic scale of the Y-axis. The results shown in this figure were obtained with a slightly different qPCR protocol in comparison to the results in Fig. 9, which explains the slight differences in average copy numbers in the pure-bred control populations (see Table 2).

Table 1. *Slx*, *L19* and *Sly* copy numbers in wild pure-bred populations

			Copy Number		
			Slx	L19	Sly
M. m. domesticus	Germany (D)	D1	41 ± 3	41 ± 3	
		D2	48 ± 4	37 ± 3	
		D3	46 ± 6	37 ± 5	
		D5	42 ± 3	32 ± 2	
		D10	46 ± 5	34 ± 4	
		D15	73 ± 12	50 ± 8	
		D17	57 ± 7	43 ± 6	
		D9	36 ± 10	29 ± 9	25 ± 7
		D12	58 ± 4	41 ± 3	33 ± 3
		D13	58 ± 8	42 ± 6	34 ± 4
		D19	52 ± 2	31 ± 2	63 ± 6
		D25	97 ± 7	23 ± 2	23 ± 2
	France (MC)	MC07	66 ± 9	36 ± 5	
		MC13	51 ± 4	42 ± 3	
		MC15	54 ± 4	34 ± 3	
		MC22	61 ± 4	29 ± 1	
		MC5	45 ± 6	25 ± 3	
		MC6	52 ± 3	35 ± 2	
		MC2	74 ± 35	49 ± 23	57 ± 27
		MC3	62 ± 2	25 ± 2	45 ± 2
		MC4	73 ± 5	28 ± 2	42 ± 1
		MC8	57 ± 4	44 ± 3	36 ± 3
		MC10	88 ± 3	30 ± 2	18 ± 2
		MC18	143 ± 13	72 ± 6	165 ± 13
Average			58 ± 15	36 ± 8	38 ± 15
M. m. musculus	Kazakhstan (AL)	AL01	158 ± 7	56 ± 2	
		Al02	145 ± 13	47 ± 4	
		AL07	136 ± 14	45 ± 5	
		AL09	116 ± 12	42 ± 4	
		AL11	88 ± 7	33 ± 3	
		AL03	177 ± 16	67 ± 6	238 ± 19
		AL04	117 ± 13	37 ± 4	178 ± 18
		AL06	127 ± 17	50 ± 7	150 ± 22
		AL13	112 ± 13	53 ± 6	208 ± 27
		AL14	150 ± 17	55 ± 6	216 ± 18
		AL15	148 ± 14	50 ± 5	246 ± 11
		AL16	122 ± 6	62 ± 3	252 ± 18
	Czech Republic (CR)	CR01	107 ± 5	33 ± 2	
		CR05	118 ± 15	35 ± 5	
		CR06	143 ± 11	46 ± 4	
		CR12	160 ± 16	47 ± 5	
		CR13	99 ± 5	30 ± 1	
		CR14	91 ± 3	23 ± 3	
		CR15	103 ± 10	27 ± 0	190 ± 3
		CR16	108 ± 5	23 ± 2	160 ± 12
		CR17	128 ± 15	47 ± 5	239 ± 17
		CR10	97 ± 6	32 ± 2	172 ± 21
		CR02	123 ± 4	35 ± 1	203 ± 22
		CR03	144 ± 6	40 ± 2	181 ± 8
Average			126 ± 24	42 ± 12	203 ± 34

Table 2. Mean copy numbers and independent t-test comparisons

			Six			L19			Sly		
n	groups	Comparisons	CN mean[1]	SD[2]	p-value[3]	CN mean[1]	SD[2]	p-value[3]	CN mean[1]	SD[2]	p-value[3]
24[a]	2[e]	Wild *M. m. domesticus* vs.	62	23		37	11		49	41	
24[a]	2[e]	Wild *M. m. musculus*	126	24(=)	**<< 0.005**	42	12(=)	**0.11**	203	34(=)	**<< 0.005**
11[a,b]	4[e]	Western hybrid zone animals	63	24(=)	**0.88**	47	12(=)	**0.014**	232	119(≠)	**0.012**
11[a,b]	2[e]	offspring from Western hybrid zone animals	115	44(≠)	**<< 0.005**	65	30(≠)	**0.011**	113	101(≠)	**0.190**
24[a,b]	2[e]	Wild *M. m. musculus* vs.	126	24		42	12		203	34	
28[a,b]	7[e]	Eastern hybrid zone animals	182	41(≠)	**<< 0.005**	84	21(≠)	**<< 0.005**	427	123(≠)	**0.01**
28[a,b]	5[e]	offspring from Eastern hybrid zone animals	667	52(≠)	**<< 0.005**	414	285(≠)	**<< 0.005**	1055	722(≠)	**0.01**
26[a,c]	7[f]	Male *M. m. musculus* vs.	163	69		35	14				
25[d], 8[a]	3[f]	F1 hybrid males	237	123(≠)	**0.012**	44	21(≠)	**0.027**			
16[d], 4[a]	1[f]	Backcross males from F1 hybrid male x *M. m. musculus* female	353	125(≠)	**<< 0.005**	79	29(≠)	**<< 0.005**			
83[d], 24[a]	6[f]	Male *M. m. domesticus* vs.	111	40		31	7				
28[d], 8[a]	3[f]	F1 hybrid males	124	52(=)	**0.193**	52	34(≠)	**0.004**			
26[a,c]	7[f]	Male *M. m. musculus* vs.							99	28	
27[d], 8[a]	3[f]	F1 hyrid males							133	34(=)	**<< 0.005**
16[d], 4[a]	1[f]	Backcross males from F1 hybrid male x *M. m. musculus* female							197	82(≠)	**<< 0.005**
82[d], 24[a]	6[f]	Male *M. m. domesticus* vs.							33	10	
26[d], 8[a]	3[f]	F1 hybrid males							46	28(≠)	**0.043**

[1]Copy number mean, [2]Standard deviation. In brackets: Equal or unequal variances between the two groups being compared. [3]Red=Not significant, Light green: P<0.05, Dark green: P<0.005.
n equals: [a]Individuals, or DNA samples taken from: [b]Heart. [c]Ear. [d]Heart or Liver (up to two meaurements per individual).
Groups equal: [e]Geographic locations, see text, or [f]Sibships.

2.2.2 CNVs in Wild Hybrid Populations

To determine the significance of these CNV loci to reproductive isolation, I assayed copy numbers from 39 animals that were caught across a natural hybrid zone in Bavaria. If we take copy number as an allele, then patterns of introgression across a natural hybrid zone can inform on genetic incompatabilities between these alleles, that is a steep versus shallow cline of introgression. However, instead of a typical gradient of allele frequencies, I discovered a striking increase in mean copy number in all three assays. (Tables 2 & 3, Figs. 5 & 6). Plotting the values for each individual onto the geographic location where the animal was caught does show indications of an East-West cline (Fig. 6), however compared to wild populations all values are inflated. All three assays show the highest copy numbers in the Eastern part of the hybrid zone, where the hybrids are mostly *M. m. musculus* with some introgression of autosomal *M. m. domesticus* alleles (B. Harr, personal communication). This correlates to a higher copy number of *Slx* and *Sly* in the wild *M. m. musculus* populations, described above, although the absolute values measured in many of the hybrid zone animals exceeds those of the pure subspecies.

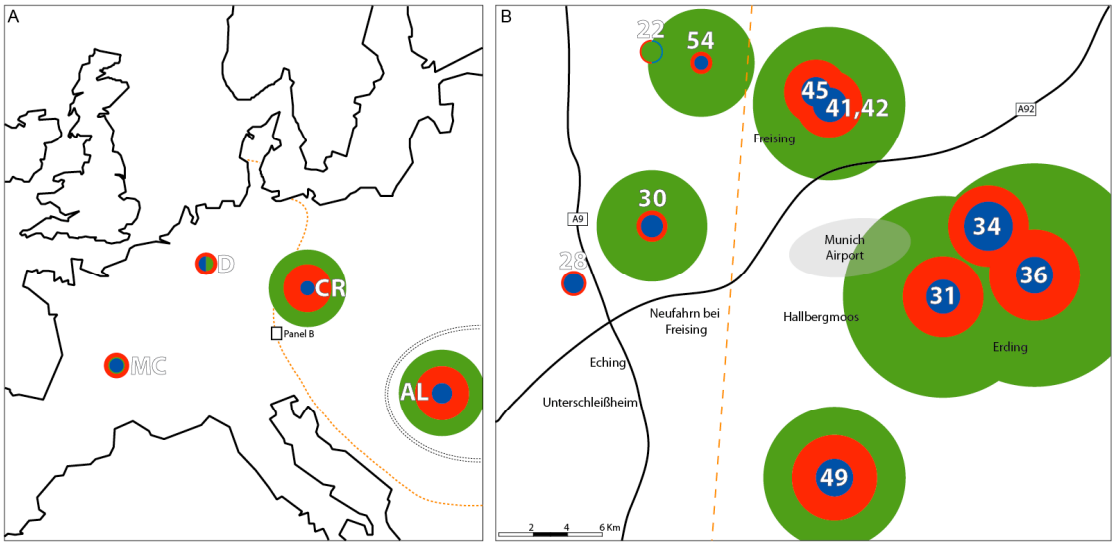


Figure 6. Copy Number Distribution in Wild Hybrids.
The size of the circles represent average copy numbers for the respective sites (red = Slx, blue = L19, green = *Sly*, see tables 1 and 3 for exact values). For comparison purposes all circles are drawn on the same scale. The orange line represents the inferred midpoint of the hybridzone, based on the analysis of diagnostic SNPs in hybrid animals (data provided by R. Rottscheidt and B. Harr). **A)** Four pure-bred populations of 12 individuals each, and **B)** 39 hybrid mice from 11 capture sites at a known Bavarian transect of the hybrid zone. The circles are centered around capture sites.

Table 3. *Slx*, *L19* and *Sly* copy numbers in wild hybrid populations

		Copy Number		
		Slx	L19	Sly
Western Hybrid Zone	**22.1 (f)**	48 ± 7	51 ± 4	
	22.3 (m)	69 ± 10	60 ± 5	49 ± 4
	28.1 (f)	67 ± 8	47 ± 3	
	28.2 (f)	69 ± 14	61 ± 4	
	28.4 (f)	51 ± 16	38 ± 5	
	30.1 (m)	61 ± 16	52 ± 8	255 ± 40
	30.2 (m)	104 ± 17	66 ± 6	336 ± 27
	30.3 (m)	64 ± 11	43 ± 3	228 ± 24
	54.1 (m)	44 ± 3	37 ± 1	372 ± 13
	54.2 (m)	99 ± 9	28 ± 2	154 ± 18
	54.3 (f)	17 ± 2	38 ± 2	
	West Avg.	63 ± 24	47 ± 12	232 ± 119
Eastern Hybrid Zone	**31.1 (f)**	181 ± 36	85 ± 9	
	31.5 (f)	169 ± 39	73 ± 6	
	31.6 (m)	218 ± 37	87 ± 9	494 ± 64
	31.7 (f)	228 ± 40	91 ± 9	
	34.1 (f)	199 ± 31	120 ± 7	
	36.1 (f)	166 ± 16	67 ± 6	
	36.2 (m)	282 ± 32	108 ± 9	660 ± 54
	36.3 (m)	223 ± 20	97 ± 10	433 ± 40
	41.1 (f)	145 ± 12	78 ± 9	
	41.2 (m)	197 ± 8	123 ± 9	464 ± 38
	41.3 (m)	227 ± 25	112 ± 14	
	41.4 (f)	125 ± 13	46 ± 5	
	42.1 (f)	203 ± 19	81 ± 9	
	42.2 (m)	95 ± 6	71 ± 5	373 ± 25
	42.3 (f)	169 ± 18	78 ± 6	
	42.4 (f)	136 ± 14	87 ± 4	
	42.5 (f)	183 ± 7	85 ± 8	
	42.6 (m)	174 ± 27	77 ± 9	290 ± 37
	45.1 (f)	154 ± 14	80 ± 2	
	45.2 (f)	156 ± 8	61 ± 2	
	45.4 (f)	139 ± 13	68 ± 6	
	45.5 (f)	154 ± 24	79 ± 8	
	45.6 (f)	178 ± 20	72 ± 6	
	45.7 (f)	161 ± 21	71 ± 7	
	49.2 (f)	195 ± 24	142 ± 6	
	49.2 (f)	252 ± 12	92 ± 3	
	49.3 (m)	235 ± 9	86 ± 5	428 ± 25
	49.4 (m)	157 ± 14	49 ± 3	274 ± 14
	East Avg.	182 ± 41	84 ± 21	427 ± 123
	All Hybrids	149 ± 66	74 ± 25	344 ± 153

To make the most appropriate use of statistical measurements, I divided the hybrid individuals into two groups based on their SNP profiles provided to me by B. Harr. This allowed me to compare the Eastern *M. m. musculus*-like hybrids to the pure *M. m. musculus* populations and the Western *M. m. domesticus*-like hybrids to the pure *M. m. domesticus* populations. Using independent t-test and accounting for unequal variances where necessary, all assays, with the exception of *Slx* in the Western hybrids,

have significantly higher means in the hybrids compared to the respective pure-bred population.

Studies on a variety of genes have shown that there can be strong selection against incompatible hybrid genotypes in the hybrid zone (Teeter et al., 2008). Therefore, I tested whether offspring of animals from the hybrid zone show even greater copy number increases under protected laboratory conditions, which would suggest that there is selection against the most extreme variants in the wild. Among the 39 offspring from seven crosses of hybrid zone animals, there is a huge variance in copy number with up to 40-fold differences in copy number at a given locus (Table 4, Figs. 5 & 7). Breaking this down to the individual families shows that this variance can occur among the offspring of the same parents (Fig. 7). Since I don't find such extreme values among animals directly caught in the hybrid zone, I can infer that there is a selection against animals with extreme copy number differences, i.e. that they do not survive long after birth under natural conditions. Once again, for statistical tests I separated the offspring from Eastern versus Western hybrid animals and compared means against appropriate pure-bred populations (Table 2). All assays show a significant increase in mean copy number in the laboratory-bred offspring of wild-caught hybrid animals.

Table 4. *Slx*, *L19* and *Sly* copy numbers in hybrid offspring families

				Copy Number		
				Slx	L19	Sly
Cross 1	P		22.3 (m	69 ± 10	60 ± 5	49 ± 4
			22.1 (f)	48 ± 7	51 ± 4	
	F1	1	1cm4	94 ± 7	70 ± 7	55 ± 5
		2	1cm2	45 ± 5	68 ± 6	43 ± 3
		3	1cm1	87 ± 10	146 ± 21	57 ± 8
		4	1cf1	138 ± 21	65 ± 6	
		5	1bf1	80 ± 7	59 ± 6	
		6	1am1	72 ± 5	49 ± 4	38 ± 3
		7	1af1	129 ± 21	28 ± 2	
Cross 2	P		30.1 (m	61 ± 16	52 ± 8	255 ± 40
			28.1 (f)	67 ± 8	47 ± 3	
	F1	1	2m2	128 ± 14	51 ± 4	225 ± 20
		2	2m1	155 ± 16	54 ± 6	261 ± 29
		3	2f2	140 ± 3	55 ± 6	
		4	2f1	200 ± 27	73 ± 9	
Cross 3	P		36.3 (m	223 ± 20	97 ± 10	433 ± 40
			34.1 (f)	199 ± 31	120 ± 7	
	F1	1	3bm3	2065 ± 302	444 ± 57	2631 ± 342
		2	3bm2	741 ± 63	379 ± 65	1993 ± 181
		3	3bm1	617 ± 148	286 ± 36	1723 ± 284
		4	3bm1	376 ± 45	144 ± 17	823 ± 94
		5	3am1	990 ± 108	342 ± 46	1476 ± 214
		6	3af3	592 ± 57	311 ± 48	
		7	3af2	982 ± 93	350 ± 47	
		8	3af1	901 ± 75	320 ± 30	
Cross 4	P		36.2 (m	282 ± 32	108 ± 9	660 ± 54
			36.1 (f)	166 ± 16	67 ± 6	
	F1	1	4cm1	736 ± 62	382 ± 38	868 ± 62
		2	4cf1	345 ± 55	809 ± 77	
		3	4bm2	320 ± 33	147 ± 5	743 ± 64
		4	4bm1	1180 ± 143	345 ± 44	1620 ± 47
		5	4bf3	104 ± 13	95 ± 15	
		6	4bf2	549 ± 57	279 ± 18	
		7	4bf1	880 ± 85	191 ± 27	
		8	4af1	761 ± 7	246 ± 19	
Cross 5	P		42.2 (m	95 ± 6	71 ± 5	373 ± 25
			42.4 (f)	136 ± 14	87 ± 4	
	F1	1	6m2	160 ± 10	262 ± 19	536 ± 37
		2	6m1	973 ± 65	813 ± 67	
		3	6f1	129 ± 13	121 ± 12	217 ± 25
Cross 6	P		42.6 (m	174 ± 27	77 ± 9	290 ± 37
			42.5 (f)	183 ± 7	85 ± 8	
	F1	1	7m2	112 ± 7	445 ± 46	871 ± 46
		2	7m1	114 ± 9	104 ± 7	213 ± 14
		3	7f2	575 ± 38	644 ± 22	
		4	7f1	114 ± 6	783 ± 73	
Cross 7	P		49.4 (m	157 ± 14	49 ± 3	274 ± 14
			49.2 (f)	252 ± 12	92 ± 3	
	F1	1	11m2	292 ± 12	282 ± 27	517 ± 48
		2	11m1	290 ± 22	267 ± 8	535 ± 14
		3	11f3	2258 ± 94	1309 ± 163	
		4	11f2	911 ± 103	877 ± 107	
		5	11f1	605 ± 43	613 ± 71	

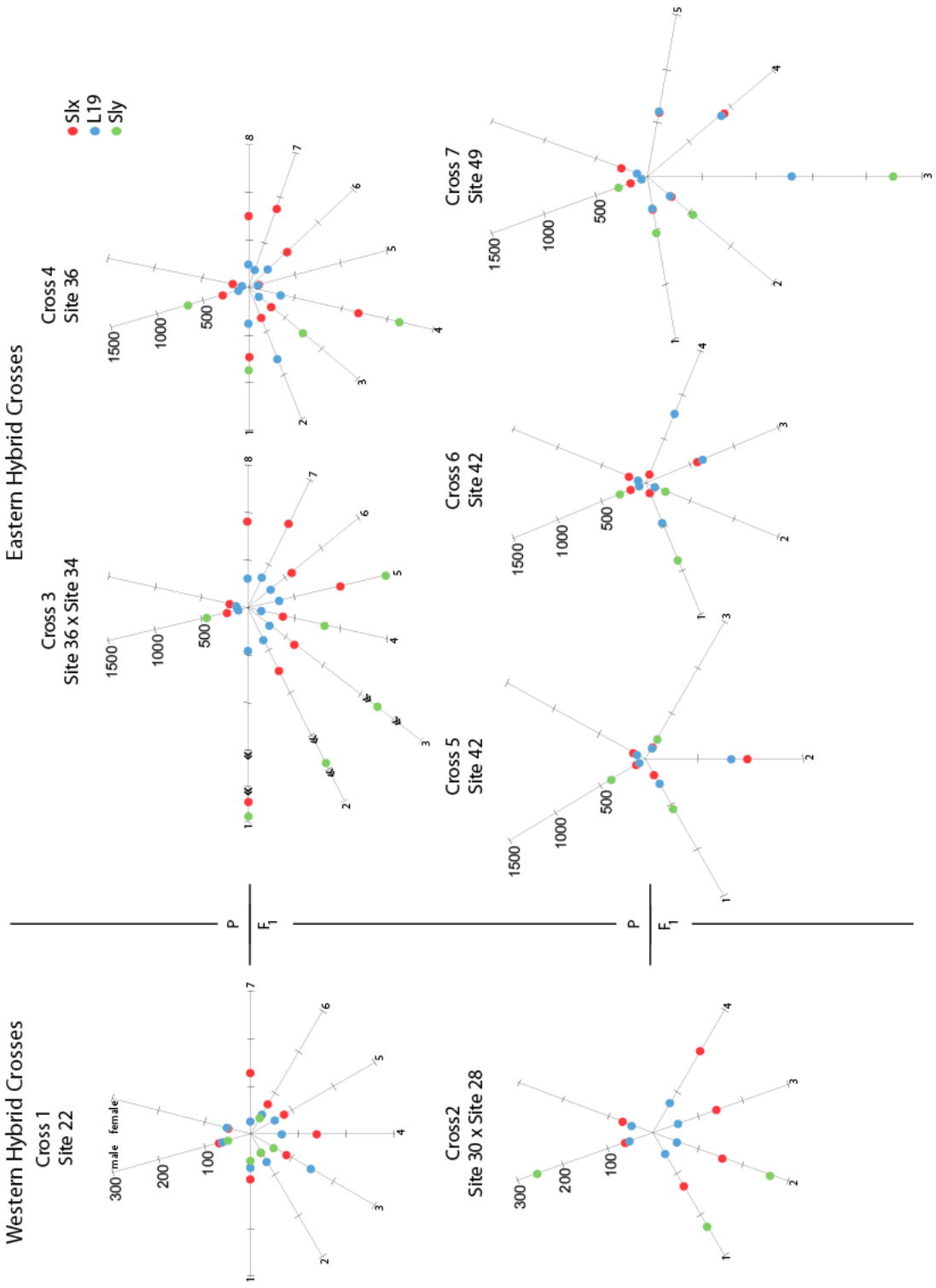


Figure 7. Copy Numbers in Hybrid Families.

Each diagram represents a single cross using individuals from a single or two adjacent sites, as indicated (see Fig. 6B for site numbers). For each plot, the parental copy numbers are represented on the upward facing axes, the offspring individuals are represented on the horizontal and down facing axes. There is a large range of variation in the F1 generation that cannot be accounted for by the parental genotypes.

Although there is a trend towards higher copy number in hybrid individuals, some variation in copy number in the pure-bred individuals is also observed. Additionally, despite a very high reproducibility for technical triplicates on the same assay plate, technical replicates performed on different days did not always yield the same results. This suggests a technical limitation of the PCR assays, caused by two factors. The first is the necessity to use a single copy gene as a reference, which becomes problematic when the copy number of the locus of interest can be several hundred times higher. Still, quantitative PCR is known to work reasonably well even for such extreme differences and I confirmed this in calibration assays (see materials and methods). The second factor concerns the fact that I survey gene families whose copies are not all identical, i.e. there are polymorphisms in the qPCR primer binding sites, despite targeting the least polymorphic region among paralogs (Fig 8). In addition, I could expect additional polymorphisms in the copies that are not expressed. Such polymorphisms can significantly alter the performance of qPCR and this is very sensitive to the exact conditions applied. Hence, for these loci qPCR has an unavoidable inherent technical noise. The same would be true for hybridization based methods on microarray platforms with short oligonucleotide probes.

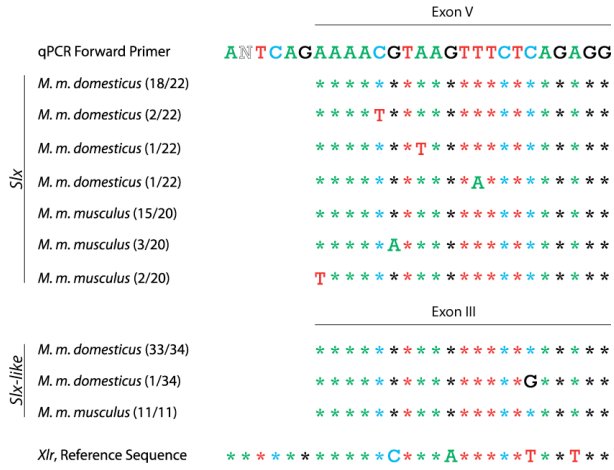


Figure 8. Polymorphisms in the Slx qPCR forward primer binding site.
From the cDNA analysis previously described, Several polymorphisms in the Slx qPCR assay forward primer binding site are noted. Note that this only represents the expressed variation and further polymorphisms may be present on the genomic DNA, which is used for the qPCR.

2.3 CNVs in Laboratory-Bred Hybrids

2.3.1 CNVs in the F1 and F2 Generation

To determine when, genealogically, CNV destabilization occurs, I generated hybrid animals in the laboratory. Given the inherent qPCR noise, I assayed two organs per individual and repeated most experiments. This means that for each individual there can be up to four data points per assay. To eliminate technical noise further, I averaged all meaurements for all individuals of a particular genotypic class, viewing them as a population. For this, I also took advantage of the fact that my loci of interest are on the sex chromosomes. Therefore, I limited my study to only male offspring, where the identity of the sex chromosomes are unambiguous. I define three broad genotypic classes: pure-bred controls, F1 hybrids and an (F2) backcross, and group individuals based on the origin (*M. m. domesticus* or *M. m. musculus*) of each sex chromosome (Fig. 9). These crosses are summarized in Table 5 and stastical values are presented in Table 2.

Surprisingly, mean copy number is significantly increased for all assays in the F1 hybrids, save for *L19* on the *M. m. domesticus* inherited X-chromosome (Tables 2 & 5, Fig. 9). This indicates that, independent of meiotic recombination, simply being in a hybrid genome is already enough to cause CNV destabilization. It appears that the *M. m. musculus*-inherited X-chromosome is most strongly affected, likely because it has more copies of *Slx* and *Sly* than *M. m. domesticus*. Additionally, the most extreme copy number increase is observed in the backcross population, and notably, the X-chromosome in this case in inherited from a pure-bred *M. m. musculus* mother, with the father being an F1 hybrid. Therefore, there was no opportunity for hybrid meiotic recombination in any of the chromosomes surveyed.

Table 5. Individual copy numbers for laboratory bred animals.

		HEART												LIVER									
		Six				*L19*				*Sly*				*Six*				*L19*				*Sly*			
		Run I		Run II		Run I		Run II		Run I		Run II		Run I		Run II		Run I		Run II		Run I		Run II	
		CN	SD	CN	SD	CN	SD	CN	SD	CN	SD	CN	SD	CN	SD	CN	SD	CN	SD	CN	SD	CN	SD	CN	SD
Pure *M m domesticus*	Cross 11 Sibship	100	5	185	24	31	4	29	5	45	2	48	10	80	2	72	6	29	1	24	1	33	2	72	3
				99	10			26	1					76	4	90	6	19	2	33	3				
		129	14	96	8	31	3	31	2	40	3	34	2	112	9	117	9	30	4	32	1	36	2	46	3
		96	10	120	17	33	3	45	4	33	3	36	5	75	9			20	1			31	3		
	Cross 12 Sibship	86	10	91	6	25	2	31	3	35	2	34	2	88	6	78	1	24	1	29	1	28	3	56	3
				76	3			24	2			34	4	87	9	75	7	27	2	29	1	41	5	64	5
		99	5	111	7	31	1	34	3	33	1	34	2	91	4	97	8	32	3	33	1	33	3	49	4
		96	1	106	15	31	1	32	4	31	2	34	3	125	20			34	4	42	6	37	4	55	4
		104	12	130	19	32	2	39	6	35	1	40	5	77	8	178	31			30	5	37	5	29	5
	Cross 16 Sibship	67	3	62	5			17	2	25	2	23	1	69	9	78	7	20	2	29	3	24	2	50	2
		98	11	74	12	32	4	27	3	24	2	25	4	163	8	151	17	47	5	40	5	42	1	37	4
		76	3	105	11	23	2	39	4	26	1	33	3	140	15	130	19	29	3	42	7	34	2	36	5
		68	1	111	11			35	5	21	2	34	3	155	18	75	4	35	2			35	4	28	2
	Cross 4 Sibship	167	15	142	20	34	5			30	3	22	3												
		263	18	78	9	62	6	29	3	48	5	40	4	116	9	106	7	32	2	35	4	25	1	40	2
		159	16	217	33	39	3	23	3	30	3	47	8												
		139	11	145	27	39	5	41	6	27	3	25	4												
		144	21	139	9	40	6	32	3	31	4	21	1	114	7	90	12	28	3	25	3	18	1	30	4
	Cross 7 Sibship	276	16	75	13	31	5	15	2	28	2	21	3	152	22	148	12	40	5	31	4	34	4	21	2
				99	14	33	3	27	2			33	2	142	14	107	5	39	3	29	2	35	3	23	2
		107	4	70	7	25	1	21	0	22	2	22	2	159	15	92	8	31	2	35	3	29	2	34	5
	Cross 8 Sibship	80	9			24	3	44	3	29	3	38	4	67	3	134	18	26	3	21	2	25	2	21	2
		80	10	103	12	22	2	40	1	29	2	42	3	68	10	114	11	25	2			27	1	27	4
		101	15	135	20	28	4	34	3	30	3	23	3	94	10			33	3	34	4	31	4		
Type A F1 Hybrids	Family 2 Sibship	98	5	70	2	32	2	23	2	129	10	152	15												
		83	12	85	10	27	2	34	3	123	5			200	29	104	19	94	6	30	4	137	8	101	19
		79	9	70	9	29	4	25	4	117	18	185	27	185	15	116	19	84	7	32	6	114	10	82	14
	Family 3 Sibship	121	12	101	14	62	8	27	3	121	10	116	17	99	8	117	1	49	6	30	1	108	9	119	17
		126	13	94	13	63	8	29	5	130	12	138	15												
		102	6	92	9	55	3	25	2	122	9	106	16	105	8	210	10	52	2	51	5	107	3	210	19
	Family 4 Sibship	228	31	259	15	91	11	55	7	179	24	198	27	99	12	103	7	44	5	24	1	106	6	125	13
		127	18	210	2	59	9			127	17	209	22	110	12	90	15	43	2	23	2	109	3	141	19
Type B F1 Hybrids	Family 5 Sibship	639	27			179	19	97	16	59	2	139	25	427	35	394	63	78	4	81	8	46	4	61	8
														193	22	188	19	37	2	35	6	26	1		
		292	25	230	28	79	3	44	6	35	4	50	7	180	9	220	24	36	2	45	3	26	2	44	5
												95	10	222	11	193	18	43	3	39	4	25	2	50	7
		156	3	190	21	37	2	53	7	25	1	44	6	162	10	138	21	32	1	31	5	23	1	35	5
	Family 7 Sibship	181	6	257	14	30	2	48	4	34	2	29	2	174	15	155	15	30	0	25	3	25	1	30	4
	Family 8 Sibship	493	62			103	8			110	8			226	20			43	1			45	1		
		184	14	159	13	32	3	30	3	31	2	35	0	148	13	146	30	30	2	31	6	31	1	30	5
Backcross	Family 16 Sibship	549	50	478	36	112	8	128	13	242	2	336	15	265	15	347	37	60	2	71	9	126	3	190	22
		403	12	403	47	95	3	93	11	239	8	307	34	172	10	176	13	43	2	46	4	113	6	138	10
		520	54	488	11	111	8	100	12	270	19	308	22	241	29	219	19	52	3	34	3	114	9	127	4
		408	29	448		95	6	106	14	212	14	233	10	247	16	299	17	59	4	64	5	118	5	85	9
Pure *M m musculus*		233	38			54	8			129	11														
		125	27			25	8			66	14														
		136	42			37	9			67	18														
		120	31			25	3			65	7														
		110	16			29	6			69	6														
		91	29			20	3			67	10														
		78	4			14	3			79	4														
		108	25			22	3			72	9														
		145	17			32	6			106	21														
										106	19														
		109	27			27	3			86	11														
		182	20			50	3			105	13														
		151	14			34	4			108	8														
		99	14			24	3			87	10														
		410	76			75	6			171	36														
		259	62			59	12			127	32														
		182	22			33	6			109	9														
		116	11			25	4			80	6														
		179	38			35	4			114	22														
		134	17			30	3			93	10														
		136	32			21	3			56	10														
		232	54			54	8			157	14														
		150	25			34	4			111	14														
		183	18			37	10			105	15														
		207	42			46	10			117	26														
		224	11			44	4			127	5														
		146	8			32	5			92	12														

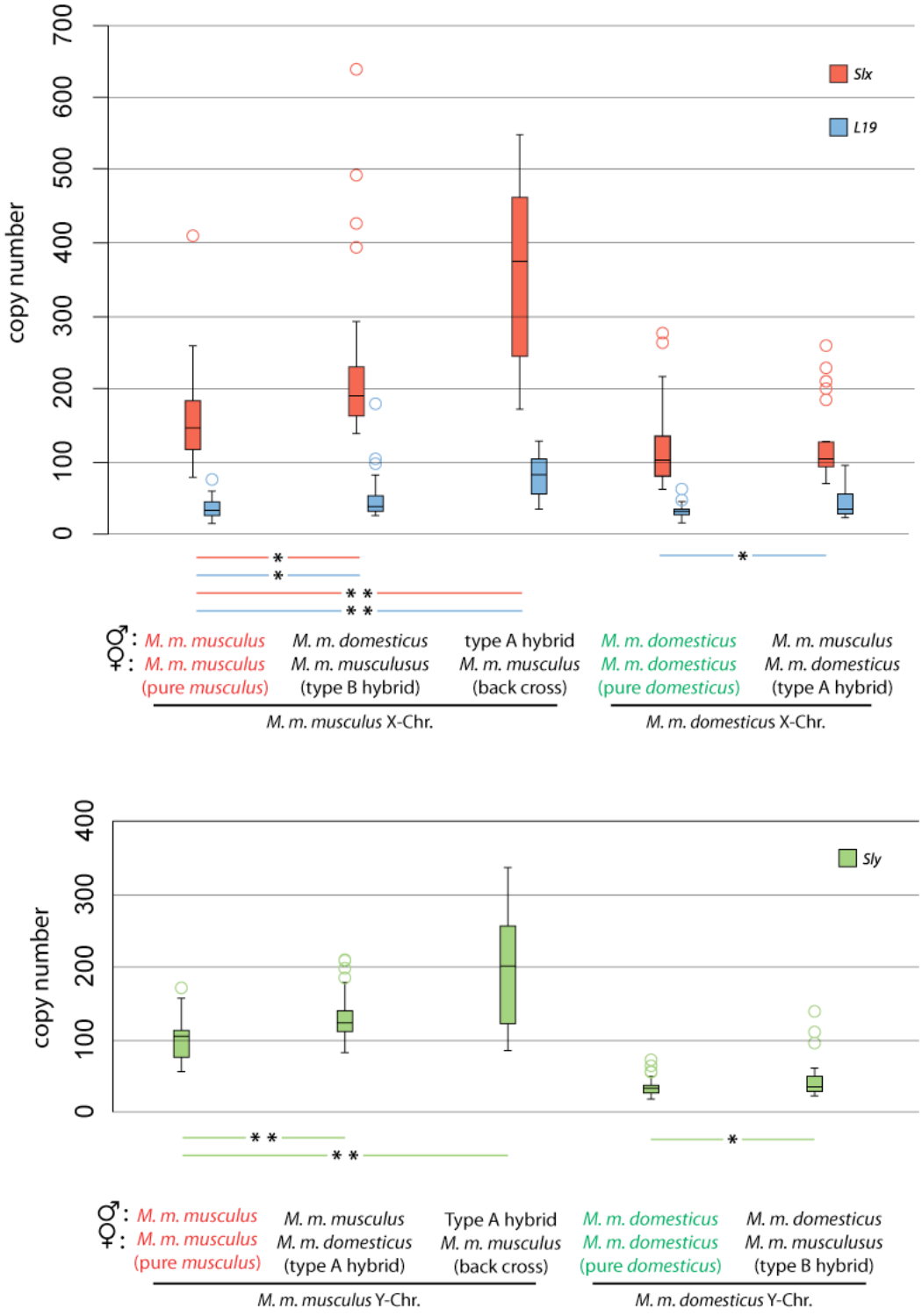


Figure 9. CNV in Hybrid Crosses performed in the laboratory.

Because only males are analysed, the origin of the sex chromosomes is unambiguous and because all mothers are pure-bred there is no opportunity for hybrid meiotic recombination. Each category on the X-axis represents a genotypic class, the parents of which are listed. Data points combine up to four qPCR results of each individual, boxes represent the median and the inter-quartile range (IQR, between the 25th and 75th percentile) outliers (greater or lesser then 1.5x the IQR) are represented as open circles. All loci, with the exception of *Slx* on the *M. m. domesticus* derived X-chromosome have a significant increase in mean copy number over the pure-breeding controls, but also a higher variance (single asterisk, $p < 0.05$; two asterisks, $p < 0.005$; see table 2 for precise p-values and table 5 for all individual measurements).

Given that the destabilization is visible in the F1 hybrid offspring, I reasoned that the mutations must have arisen as mitotic mutations during development. Therefore, one would predict mosaic effects, i.e. different tissues of the same individual may have different copy numbers, essentially a somatic CNV. To detect this, I developed a Southern blot assay. The genomic *Slx* and *Slx-like* repeats have a diagnostic difference in an *EcoRI* site, producing predominant 5.5kb and 8.4kb fragments, respectively (Fig. 10). Given that the proximal *Slx* and the distal *Slx-like* clusters are independent, one would expect the ratio in signal intensity between the two bands to vary between organs if somatic variation were present. I used DNA extracted from the heart (mesoderm) and liver (endoderm) of *M. m. domesticus* and hybrid animals. I chose these organs because they originate from different germs layers, which are defined early in development, giving the greatest amount of time for mosaic clones to arise and therefore be detected. I measured the signal intensity ratios of *Slx-like*:*Slx* in the heart and liver in pure-bred *M. m. domesticus* controls, the laboratory-born offspring from the wild-caught hybrids described in the previous section, and the laboratory-bred hybrids described above (Fig. 10). To respresent all information in a single value, I divided the liver *Slx-like*:*Slx* ratio by the heart ratio. For *M. m. domesticus* individuals, this value ranged from -0.7 to 0.7. Of the 21 laboratory-bred hybrids, 4 showed changes greater or lower than this range, 9 of the 25 offspring from wild hybrids were also beyond this range (Fig. 11). Some hybrid animals show an almost two-fold difference between these tissues, substantiating the notion that they are effectively mosaics for copy numbers at these CNV loci.

Figure 10. *Slx-like* and *Slx* Southern Blots.

Southern blots using a *Slx* probe on genomic DNA from the heart (left) and liver (right) of individuals representing populations of control pure-bred *M. m. domesticus* (17), laboratory-bred hybrids (21), and offspring from wild-caught hybrids (25). The ratio of signal intensity between the upper *Slx-like* and lower *Slx* band are shown below each blot. Hybrids marked by an asterisk have greater difference between liver and heart ratios than observed in the control population (see text and Fig. 11).

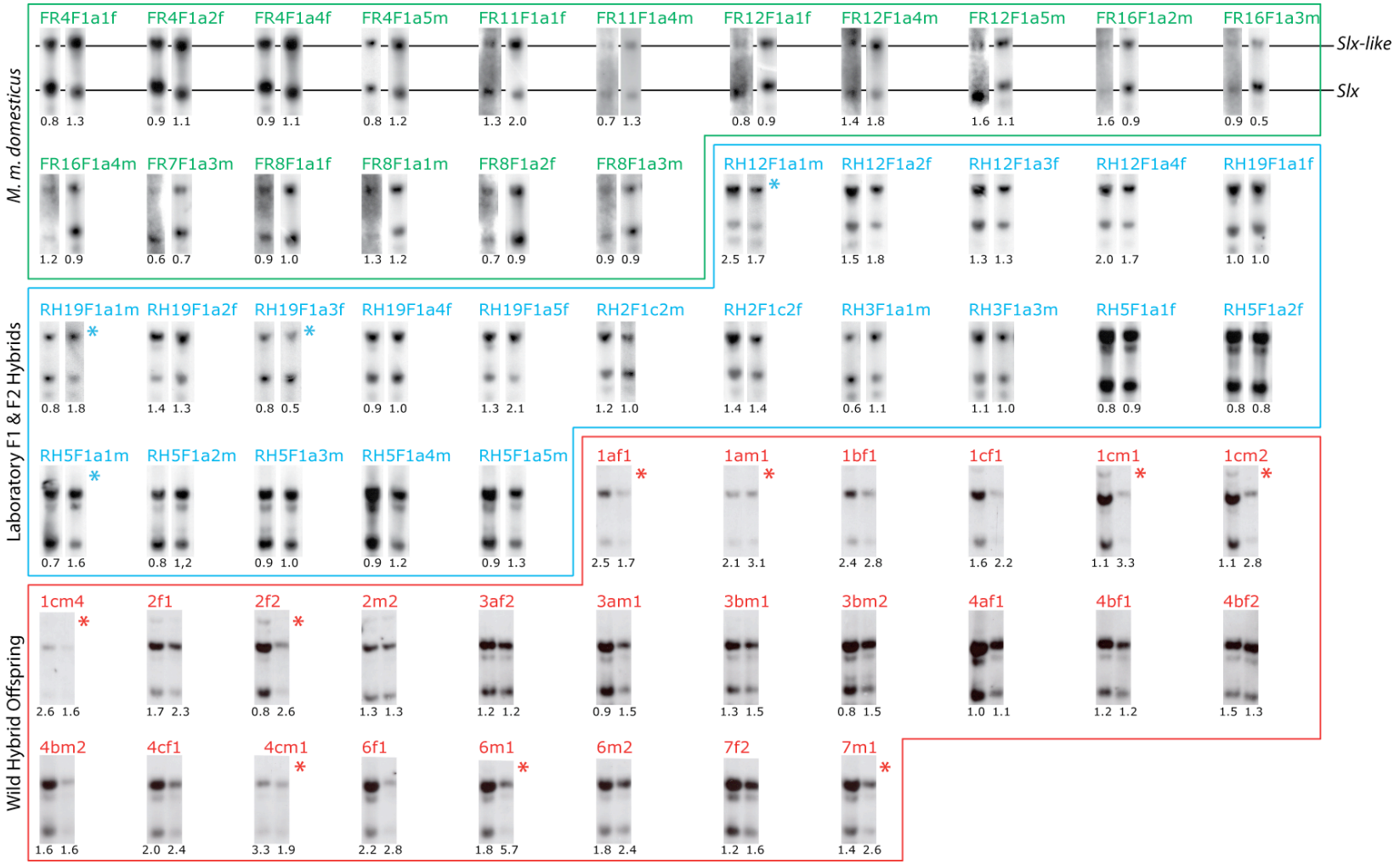


Figure 10. ***Slx-like*** **and** ***Slx*** **Southern Blots.** See text page 34.

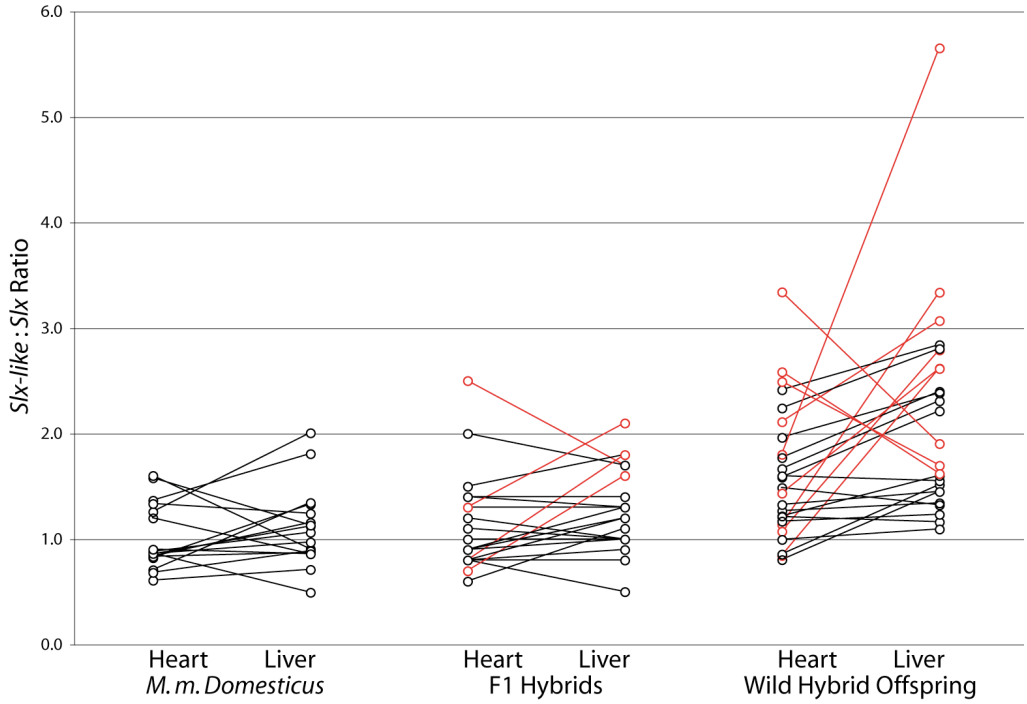


Figure 11. Southern Blot Analysis for Different Tissues.
A comparison of *Slx-like*:*Slx* hybridization intensity ratios was made between heart and liver DNA samples from the same individual. In the control *M. m. domesticus* population, the difference in intensity ratios between organs ranged from -0.7 to 0.7. Many hybrid individuals are outside this range (marked in red) and are indicated by an asterisk in Fig. 10.

2.3.2 Detection of CNV Destabilization by aCGH

In order to assess how many loci are affected by this destabilization and to confirm my results with another method, I preformed aCGH analysis. I used the Agilent 244K pre-designed mouse aCGH platform which contains approximately 244,000 60mer probes covering the entire genome with an average density of 1kb, however density is higher in gene rich regions. For this analysis four samples, one pure-bred *M. m. domesticus*, one pure-bred *M. m. musculus*, one Type A F1 hybrid and one backcross hybrid were analysed. The two hybrid individuals were chosen because of their extreme *Slx* copy number values as determined by qPCR and thus the most likely to show amplification at the *Slx* and *Slx-like* loci, but potentially other loci as well. The four individuals were not related and as such acted only as representatives of their genotypic classes.

aCGH probes in both X-linked loci of interest have higher log2 ratios than either pure-bred control. The log2 ratio represents the signal intensity of the experimental sample over that of the reference (in this case C57Bl/6J). A higher ratio corresponds to a

higher copy number detected by that probe in the experimental DNA. Seven probes covered the proximal *Slx* region, with hybrids showing the highest log2 ratio in all but one. Of the four probes covering the distal *Slx-like* region, hybrids represent again the highest log2 ratio for all but one. This is yet another test which confirms CNV destabilization at these X-linked loci.

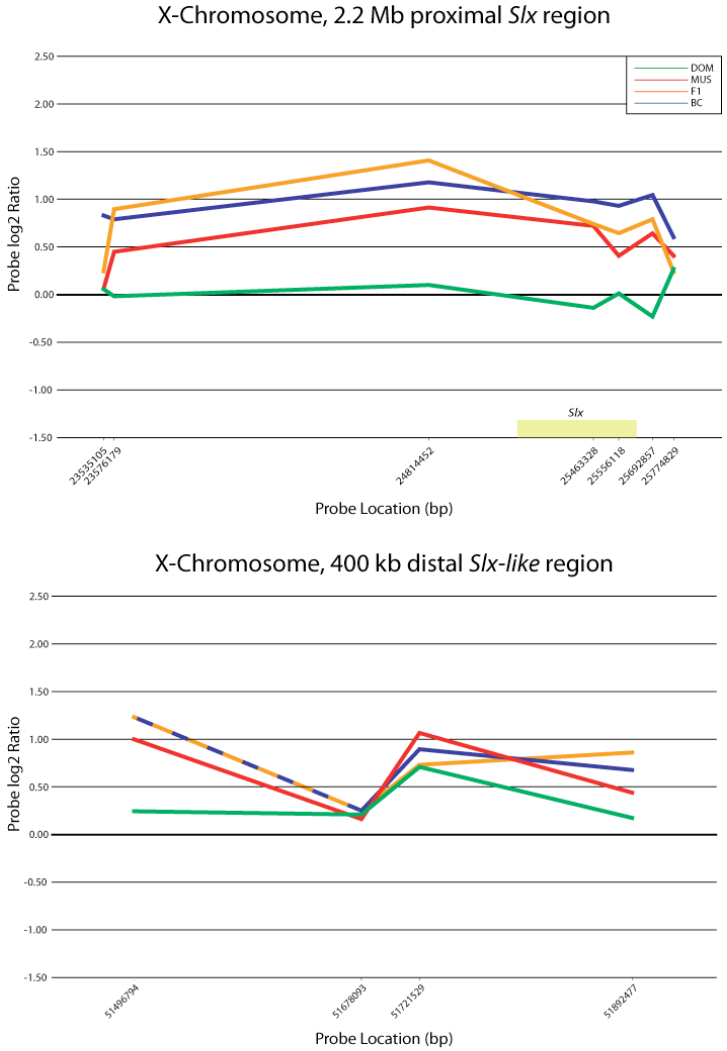


Figure 12. *Slx*- and *Slx-like*-localized aCGH probes.
In both the proximal (upper graph) and distal (lower) X-linked loci, hybrids have the highest log2 ratios for almost all probes.

The same array platform was used in a previous study (Cutler et al., 2007) to survey 42 inbred mouse strains commonly used in the laboratory. Two of these mouse stains are of *M. m. musculus* descent: CZECHII/EiJ (hereafter CZE) and PWK/PhJ (hereafter PWK). Because we both used C57Bl/6J as the reference strain, it is possible to

directly compare my results to those of the previously published CZE and PWK data. Loci that are present in both CZE and PWK are likely to be ancestral and common within *M. m. musculus*, and I should be able to also detect the same loci in my *M. m. musculus* sample. Thus I only consider CNV loci found in both CZE and PWK, which I refer to as high-confidence CNVs.

CZE and PWK have 26 high-confidence CNVs when compared to C57Bl/6J. Of these 26 loci, 21 (80%) were also present as CNVs in the *M. m. musculus* individual I used for my aCGH analysis. For autosomal CNVs, the F1 log2 ratio should be intermediate between *M. m. domesticus* and *M. m. musculus*. log2 ratios of the backcross individual, the progeny of an F1 hybrid male and a *M. m. musculus* female, should fall between the F1 and *M. m. musculus* signals. Two examples of typical high-confidence autosomal CNV loci are shown in Fig. 14.

One of the high-confidence CNV loci displays an unexpected pattern in the F1 and backcross individuals (Fig 13C). In this locus on chromosome 17, both hybrids have log2 ratios that are as high, or higher than *M. m. musculus*. This profile is consistent with CNV destabilization, of the sort I have described on the X-chromosome. It is worth to note that another locus, a deletion on Chr. 6 in *M. m. musculus* appears as a combined deletion with a small amplification in the F1 individual, and is only observed as a deletion in other individuals. However, the probes as this locus are quite noisy and it is difficult to consider this as a destabilization. Therefore from the 21 high-confidence loci I can confirm, one shows clear signs of unexpected amplification.

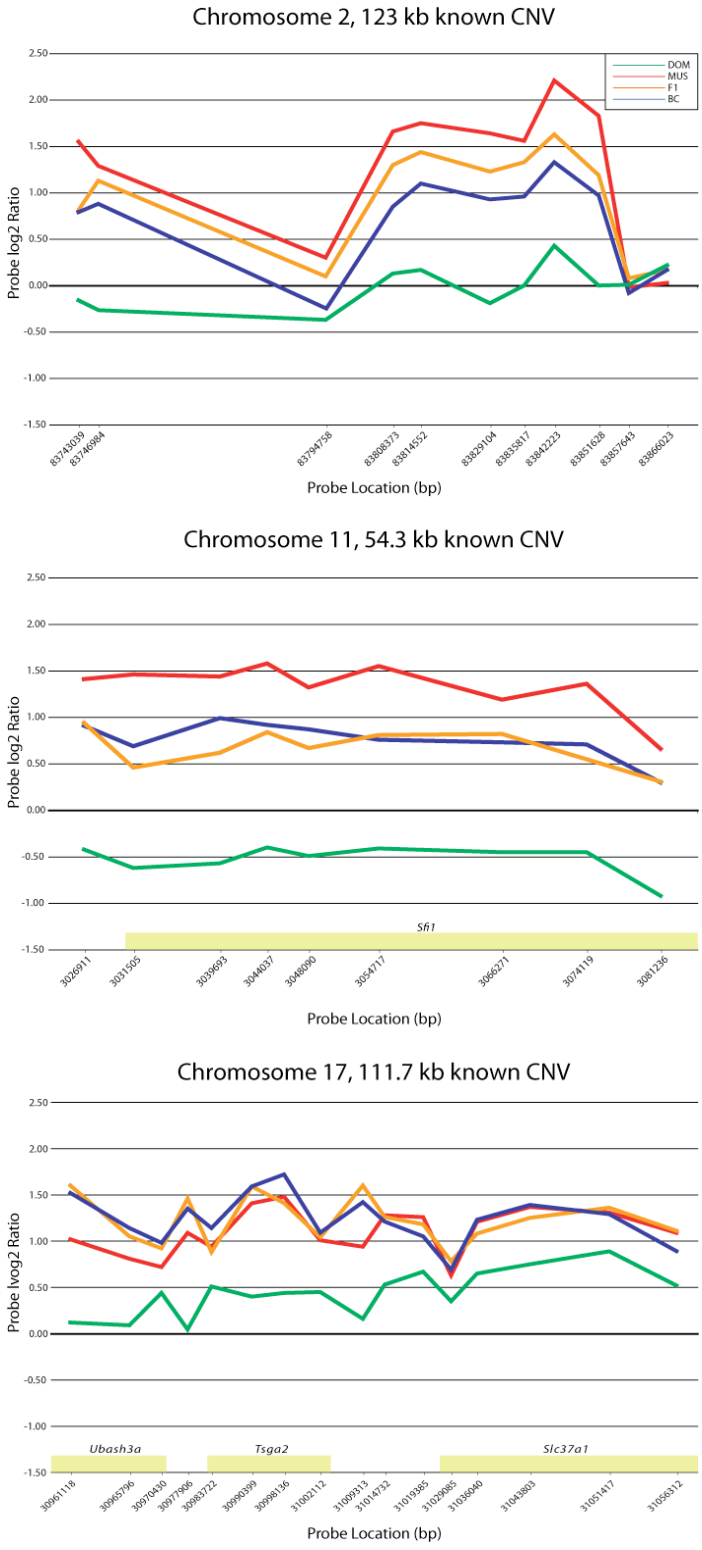


Figure 13. High Confidence CNV Loci in Hybrid Individuals

In these three loci, *M. m. musculus* and *M. m. domesticus* are strongly differentiated. The first two examples on chromosome 2 and 11 are the expected results for a Mendelian inheritance pattern, the F1 and backcross hybrids are clearly intermediate between the two pure-bred animals. However, Chromosome 17 shows a unique and unexpected pattern. The F1 and backcross individuals have log2 ratios as high, and mostly higher, than *M. m. musculus*. This represent 5% of the high-confidence CNVs and is similar to what I describe for the X-chromosome loci. Annotated genes overlapping probe locations are diagramed in yellow.

2.4 Discussion

These results provide evidence that the CNV loci that I have studied can become destabilized in *Mus musculus ssps.* intersubspecific crosses. Furthermore, animals can be mosaic for copy numbers, indicating somatic instability. This is in line with recent reports on copy number variation at CNV loci in embryonic stem cells of the mouse (Liang et al., 2008), among monozygotic human twins (Bruder et al., 2008), and somatic mosaicism within individuals (Piotrowski et al., 2008).

It seems clear that recombination mechanisms are the driving force for this variability. In a study of primary mutation events at four loci with known propensity for duplications and deletions, mutations are specific to meiosis with an excess of interchromosomal recombination (Turner et al., 2008). In contrast, the copy number changes observed here study are clearly independent of meiotic events, since they can already be seen in F1 offspring. Furthermore, the focus on sex-chromosome loci in males excludes the possibility of interchromosomal exchanges. Thus, it is clear that intrachromosomal exchanges can also cause copy number variation. Intrachromosomal exchanges were also invoked as the major factor for the recombination processes leading to concerted evolution in ribosomal genes (Schlötterer and Tautz, 1994) and there is evidence from cell-culture experiments that these processes can be very efficient (Read et al., 2004).

The fact that loci are mostly amplified is particularly noteworthy. Turner et al. (2008) found in their study an excess of deletions caused by meiotic events and postulate that other mechanisms would have to be biased towards expansion to lead to a relative stability of loci. Thus, it seems that the intrachromosomal events observed here provide such a possible expansion mechanism. Due to the technical limitations with the PCR assays, I cannot directly study the effects in individual germ cells for our loci. However, the fact that copy numbers are relatively stable within and between populations of the same subspecies suggest that there is a stable balance between insertion and deletions keeping the average size under control. Also, the fact that there is a gradient of copy number differences across the hybrid zone suggests that there is a heritable component in the copy number expansions and that the effects in the hybrids are not solely somatic.

M. m. domesticus and *M. m. musculus* are not considered true species, since they can produce viable hybrids. On the other hand, they form a rather stable hybrid zone under natural conditions, and hybrids are less fit than their parental populations. The reasons for this are not fully understood and it appears to be linked to several regions on the X-chromosome (Good et al., 2008; Storchová et al., 2004). My observation that animals with extreme copy number changes can only be found in laboratory bred animals from the hybrid zone, but not in directly caught animals, implies that these extreme changes may have adverse effects on viability.

This destabilization effect is reminiscent of hybrid dysgenesis in *Drosophila*. Originally discovered in crosses of different wild-derived strains, dysgenesis leads to multiple mutations and genomic rearrangements in F1 offspring (Kidwell et al., 1977). It is generally thought that it is caused by the mobilization of transposable elements, where a transacting repressor becomes impaired in the hybrid situation, a mechanism that has been studied in detail for P elements (Castro and Carareto, 2004). A hybrid dysgenesis-like effect was also observed in marsupial mammal F1 hybrids, where centromeres of one of the parental chromosome sets are highly expanded (O`Neill et al. 1998; Metcalfe et al. 2007). In this case it seems that retroviral-like elements and global demethylation are involved in the expansion process, although it remains open whether they are caused by transposition events or other global rearrangements (O´Neill et al. 2002). There is also no evidence that global demethylation effects occur in hybrids of placental mammals, including mouse hybrids (Roemer et al. 1999). On the other hand, genomic incompatibilities with respect to imprinted loci have been observed in crosses between closely related species of *Peromyscus* (Vrana et al. 1998, 2000). Interestingly, there are also tissue specific effects with respect to disruption of imprinting at specific loci (Wiley et al. 2008), which seem to be comparable to the effects seen at these CNV loci.

There have been long standing speculations about the incompatibility of genomes in hybrid situations. Barbara McClintock discussed already in her Nobel lecture on transposable elements a possible role of "genome stress" in the formation of new species from hybrids (McClintock 1984). There are also conjectures that mismatch repair pathways may play a role in generating new point mutations and new microsatellite alleles in hybrids (Amos et al. 1996, 2008). Thus, the effects observed in CNV regions

are well in line with these previous observations. However, CNVs may be of particular relevance, since they code for functional RNAs and proteins. Given that copy number differences can build up very quickly in populations without genetic exchange, these might significantly contribute to early reproductive isolation after an initial separation of gene pools. Genes involved in reproductive isolation have so far been mainly recovered from *Drosophila* and have turned out to be fast evolving single copy genes (Orr et al. 2004). But even these fast evolutionary rates are orders of magnitude smaller than the changes at CNV loci. Hence, CNV variation may be a key to understanding early speciation processes.

It is striking to consider that such destabiliztion can occur so rapidly and that viable offspring survive to adulthood, let alone birthed. However, the details of this destabilization are not complete. Although it may occur throughout the genome, it may very well be localized to tissue neutral regions, for instance *Slx*, *L19* and *Sly* in the organs examined, where they are not expressed. To have a functional impact the destabilized loci must contain some functional component (*cis*-regulatory element, gene, etc.) for specific cell types, its interacting partners must be present, it must be expressed at the right time and, it must also surpass the threshold of canalization. A genome-wide analysis of somatic structural variation is now tractable in tumors. The situation described herein would benefit with the application of future advances in this field.

3.0 DNA Repair Pathways in Hybrid Mice

3.1 Genome Maintenance and Instability

The observation of somatic mutation in hybrid animals places attention on genome maintenance in the soma. In the developing and aging organism, somatic genome maintenance, i.e. avoidance and repair of inevitable mutation, is of paramount importance. Somatic mutations are of such importance that aging is often defined in terms of its accumulation (Vijg et al., 2005) and over 30 human diseases with variable expressivity are now attributed to them (Gottlieb et al., 2001). In the past few years, our understanding of genome maintenance has come to be understood as an intersection between three fundamental genetic processes: DNA repair, replication and recombination (West, 2003) and refers to the identification and repair of all variety of DNA alterations, from gross chromosomal rearrangements to point mutations by a number of interconnected DNA repair pathways.

Segmental aneuploidy refers to deletions, amplifications or translocations, and is mostly used in reference to somatic malignant tissues (Geigl et al., 2008). Segmental aneuploidy is essentially what I observe as CNV destabilization. A general feature of segmental aneuploidy is caused by DNA double-stranded breaks (DSBs) that remain unrepaired as a cell enters M phase. Normally, this would lead to cell death, but when it occurs in cells lacking robust checkpoints, gene amplification can ensue producing extra-chromosomal fragments, tandem duplications, or scattered insertions (Albertson, 2006).

A related concept in cancer genetics is that genomic instability – caused by sporadic loss of damage-response mechanisms – is important in cancer initiation and/or progression (Thacker and Zdzienicka, 2004). Although this is a prevalent perspective, debate persists regarding the question of genome instability as a cause or effect of tumorigenesis (Sieber et al., 2003). Nonetheless, this sets the framework for an investigation into the possibility of reduced DNA repair capacity leading to genomic instability in somatic tissues.

Several lines of evidence support the role of DNA repair pathways in CNV destabilization. For instance, it is already clear that some recurrent CNV loci can have remarkably high mutation rates (Egan et al., 2007), which are likely to occur by NAHR.

NAHR is simply a type of homologous recombination (HR) and is considered the source of many CNVs (She et al., 2008). It is important to understand that HR is not only important during meiosis, but plays a crucial role during DNA repair in somatic tissue (West 2003). Some examples will help to illustrate this point. For instance, in the chick DT40 cell line, loss of *Rad51* – a central HR component, involved in strand invasion – results in the accumulation of cells with abundant unrepaired DSBs at the G2/M phase transition (Sonoda et al. 1998). There are several murine paralogs in the *Rad51* gene family which have non-redundant functions (Thacker, 1999). Null mutants of one member, *Xrcc2*, exhibit up to an order-of-magnitude increase in chromosomal alterations and an increased occurrence of homologous recombination in mouse embryonic fibroblast (MEF) cells (Deans et al., 2003). Furthermore, $Xrcc2^{+/-}$ mice exhibit haploinsuffency, indicating dosage sensitivity in the HR repair pathway. Another significant finding is the clear distinction of phenotypes in HR versus NHEJ deficient cell lines. NHEJ specific mutant MEF cells show a ratio of rearrangements:fragments of 3:13 compared to 36:10 for the $Xrcc2^{-/-}$ mutant (Deans et al., 2003; Karanjawala et al., 1999). This also highlights the finding that HR mutants don't necessarily abolish HR, but can actually cause an increase in HR activity. HR and NHEJ are also important during specific stages of the cell cycle: HR during the G_2 and S phases where an abundant amount of homologous material is available to repair double stranded breaks with high fidelity versus NHEJ, during the G_0, G_1 and early S phases (Sonoda et al., 2006).

The finding and perspectives outlines above clearly lend support to a study of DNA repair pathways as a cause of CNV destabilization. The first steps in this approach are to decide on appropriate tissues and pathways to examine. Below I provide an outline of my decision to examine DNA repair pathways during organogenesis.

3.2 Organogenesis and Genome Maintenance

During organogenesis, ongoing rapid cell proliferation coupled with the switch from anaerobic to oxidative metabolism drives the need for DNA repair in response to increased oxidative damage by reactive oxygen species (ROS) (Caldecott, 2008; Vinson and Hales, 2002). ROS are a major source of single stranded breaks (SSBs) which are then processed into DSBs during replication and repaired by HR (Kuzminov, 2001).

During embryogenesis, expression levels of many DNA repair genes fluctuate (Jaroudi and SenGupta, 2007). Low expression of a particular pathway during specific developmental stages may represent "bottlenecks" in the repair process, revealing susceptibility to certain types of genotoxic stress. Conversely, elevated expression in a given pathway may indicate that it has a critical role at that time and location (Vinson and Hales, 2002). However, it must be noted that such simplistic models are likely to be presumptive, considering the complex phenotypes and interconnectedness of DNA repair pathways with each other and with replication (West, 2003), for example, as described with X*rcc2*$^{-/-}$ mutants above (Deans et al., 2003). A detailed study of HR- and NHEJ-specific mutations revealed an HR-dependent stage of development between E8.5 – E9.5 during the fast growth phase at the early stages of organogenesis, whereas NHEJ was crucial only after E11.0 (Orii et al., 2006). Taking these aspects of DNA repair into consideration, I decided that the most appropriate time-point to profile DNA repair pathways is at the early stages of organogenesis, at E8.5.

3.3 Experimental Outline

Provided the evidence outlined above, I hypothesized that a regulated response to repair induction will be observed during embryogenesis in hybrid embryos that is distinct from pure-breds. If this were observed, it would help to explain the mechanistic underpinnings of CNV destabilization. Two resources were particularly useful in compiling a list of genes for this study. First, a recent review article outlines many DNA repair genes known to be expressed during embryogenesis, providing an experimentally validated list of 57 genes involved in various repair pathways (Table 4) (Jaroudi and SenGupta, 2007). Second, the online gene ontology resource, (www.geneontology.org) which classifies genes based on biological process, cellular component and molecular function, was queried for genes broadly implicated in DNA Repair. This provided an additional 71 genes not reported by Jaroudi and SenGupta (2007) (Table 4). Candidates taken from the GO database were all confirmed to have embryonic expression using the dbEST viewer at NCBI (www.ncbi.nlm.nih.gov/dbEST). In both instances all genes involved in any form of DNA repair were taken as candidates. Additionally, the most well studied DNA methyltransferases were included, primarily because of their known role in

some types of cancer (Rhee et al., 2002) and point mutations rates in mammalian cells (Chan et al., 2001). Several DNA polymerases were also considered, in particular because certain low-fidelity classes are essential for DNA repair, adding point mutations at repair sites (Bavoux et al., 2005).

Combining accurate and efficient gene expression profiles of many targets in several individuals is best done by qPCR. For this purpose I once again used TaqMan qPCR assays, which are available in a high density array (HDA) format in which lyophilized assays are pre-loaded into 1uL chambers. A buffered mastermix containing genomic DNA and *taq* are then added to the chambers via centrifugation. I manually selected TaqMan assays for my candidate genes and divided them between two 384-well plates. On each plate, an assay is present in quadruplets, meaning I can run four individuals per plate, obtaining one result per assay per individual. Many genes were represented by a single assay, but in cases where all splice variants could not be detected, more than one assay was used. This means that the 128 candidate genes are represented by 164 assays (Table 6). The selection of an appropriate endogenous control is also important for this experiment and so I choose to place 11 standard endogenous control assays for gene expression (mostly housekeeping genes) on both plates, which would allow me to choose the best performing control after obtaining results.

There is an inherent component of ambiguity in surveying internally developing embryos. First, there is uncertainty of when copulation occurred, adding approximately 8 hours of uncertainty to the embryonic age. Second, interspecies variation in developmental timing has never been studied. This is a large undertaking and not within the scope of this project, but could add another degree of variation. Last, the rapid rate of development at E8.5 means that comparing embryos differing in age by only a few hours may influence my results. Therefore I designed a strategy to determine the developmental age of my samples by gene expression. I chose six genes known to be expressed only after a certain point in development between E8.0 and E9.5. I will refer to these as the "developmental age" gene set (Tables 6 & 8).

Table 6. Genes assayed in qPCR experiment and sources

High Density Array GE1

As Determined by GO Terminology		*Assay ID*
Aicda	the B-cell-specific activation-induced cytidine deaminase protein	Mm00507774_m1
Aptx	aprataxin	Mm00481554_m1
Asf1a	ASF1 anti-silencing function 1 homolog A (S. cerevisiae)	Mm00481538_m1
Atr	ataxia telangiectasia and Rad3 related	Mm01223626_m1
Atr		Mm01223652_m1
Bcl-2	B-cell leukemia/lymphoma 2	Mm00477631_m1
Bdp1	B double prime 1, subunit of RNA polymerase III transcription initiation factor IIIB	Mm01283004_m1
Bdp1		Mm01283013_m1
Cebpg	CCAAT/enhancer binding protein (C/EBP), gamma.	Mm01266786_m1
Dclre1c	DNA cross-link repair 1C, PSO2 homolog (S. cerevisiae)	Mm00455364_m1
Dclre1c		Mm00455364_m1
Ddb2	damage specific DNA binding protein 2	Mm01333907_g1
Ddb2		Mm01333911_g1
Dnmt1	DNA methyltransferase (cytosine-5) 1	Mm01151062_g1
Dnmt1	DNA methyltransferase (cytosine-5) 1	Mm01151065_g1
Dnmt3a	DNA methyltransferase 3A	Mm01323808_g1
E130016E03Rik	Uncharacterized	Mm01217421_g1
Eef1e1	eukaryotic translation elongation factor 1 epsilon 1	Mm01349382_m1
Ercc4	excision repair cross-complementing rodent repair deficiency, complementation group 4	Mm01342092_m1
Ercc5	excision repair cross-complementing rodent repair deficiency, complementation group 5	Mm01256322_m1
Ercc6	excision repair cross-complementing rodent repair deficiency, complementation group 6	Mm00621850_m1
Ercc6		Mm01221908_m1
Ercc8	excision repair cross-complementing rodent repair deficiency, complementation group 8	Mm00518465_m1
Ercc8		Mm01730955_m1
Gen1	Gen homolog 1, endonuclease (Drosophila)	Mm00724023_m1
Hmgn1	high mobility group nucleosomal binding domain 1	Mm01626329_g1
Hspa1a	heat shock protein 1A	Mm01159846_s1
Hspa1b	heat shock protein 1B	Mm03038954_s1
Kbtbd5	kelch repeat and BTB (POZ) domain containing 5	Mm01350719_g1
Lig1	ligase I, DNA, ATP-dependent	Mm00495331_m1
Lig4	ligase IV, DNA, ATP-dependent	Mm01221720_m1
Mgmt	O-6-methylguanine-DNA methyltransferase	Mm00485014_m1
Mlh1	mutL homolog 1 (E. coli)	Mm00503449_m1
Msh2	mutS homolog 2 (E. coli)	Mm00500567_m1
Msh3	mutS homolog 3 (E. coli)	Mm00487756_m1
Msh3		Mm01290054_m1
Msh6	mutS homolog 6 (E. coli)	Mm01227378_m1
Mus81	MUS81 endonuclease homolog (yeast)	Mm00472059_g1
Mutyh	mutY homolog (E. coli)	Mm01188300_g1
Neil1	nei endonuclease VIII-like 1 (E. coli)	Mm00452911_g1
Nhej1	nonhomologous end-joining factor 1	Mm01259071_m1
Nthl1	nth (endonuclease III)-like 1 (E.coli)	Mm00476559_m1
Ogg1	8-oxoguanine DNA-glycosylase 1	Mm00501781_m1
Parp1	poly (ADP-ribose) polymerase family, member 1	Mm00500171_g1
Parp2	poly (ADP-ribose) polymerase family, member 2	Mm01319555_m1
PKA	protein kinase A (PKA (geneID: 18747), phosphorylates AID)	Mm01251636_gH
Pms2	postmeiotic segregation increased 2 (S. cerevisiae)	Mm01200871_m1
Pold1	polymerase (DNA directed), delta 1, catalytic subunit	Mm00448264_g1
Polg2	polymerase (DNA directed), gamma 2, accessory subunit	Mm01242536_g1
Polh	polymerase (DNA directed), eta (RAD 30 related)	Mm00453169_m1
Poli	polymerase (DNA directed), iota	Mm01262545_g1
Polk	polymerase (DNA directed), kappa	Mm01282564_m1
Poll	polymerase (DNA directed), lambda	Mm01198394_m1
Polr2g	polymerase (RNA) II (DNA directed) polypeptide G	Mm01230938_g1
Rad52	RAD52 homolog (S. cerevisiae)	Mm00448543_m1
Rad54l	RAD54 like (S. cerevisiae)	Mm00485521_g1
Rad54l		Mm00485528_m1
Recql5	RecQ protein-like 5	Mm00499909_m1
Recql5		Mm00499917_m1
Rev1	REV1 homolog (S. cerevisiae)	Mm00450983_m1
Rev3l	REV3-like, catalytic subunit of DNA polymerase zeta RAD54 like (S. cerevisiae)	Mm00803291_m1
Rev3l		Mm01181860_g1
Rfc5	replication factor C (activator 1) 5	Mm01208090_g1
Rpain	RPA interacting protein	Mm01245732_m1
Rrm2b	ribonucleotide reductase M2 B (TP53 inducible)	Mm01165702_gH
Sod1	superoxide dismutase 1, soluble	Mm01344232_g1
Sod2	superoxide dismutase 2, mitochondrial	Mm00449725_g1
Sod2		Mm00449726_m1
Sumo1	SMT3 suppressor of mif two 3 homolog 1 (yeast)	Mm01609844_g1
Tdg	thymine DNA glycosylase	Mm00834243_g1
Trdmt1	tRNA aspartic acid methyltransferase 1	Mm00438508_m1
Trp53bp1	transformation related protein 53 binding protein 1	Mm00658689_m1
Trp53bp1		Mm01271860_m1
Uvrag	UV radiation resistance associated gene	Mm00724367_m1
Xpa	xeroderma pigmentosum, complementation group A	Mm01345389_m1
Xpc	xeroderma pigmentosum, complementation group C	Mm01183434_m1
Xrcc4	X-ray repair complementing defective repair in Chinese hamster cells 4	Mm01283067_m1
Xrcc5	X-ray repair complementing defective repair in Chinese hamster cells 5	Mm00550142_m1
Xrcc6	X-ray repair complementing defective repair in Chinese hamster cells 6	Mm01310122_m1
Xrcc6		Mm01310126_m1
Xrn2	5'-3' exoribonuclease 2	Mm01275968_m1
Xrn2		Mm01275979_m1
"Develomental Genes" from Gene Expression Database		
BMP10	First detected by in situ hybridization at E9.0 (see Neuhaus et al. 1999)	Mm03024178_s1
Nkx2.1	First detected by reverse transcription PCR at E8.25 (see Serls et al. 2005)	Mm00447558_m1
Rdh16	First detected by reverse transcription PCR at E9.5 (see Ulven et al. 2000)	Mm01625764_s1
Standard Endogenous Controls		
18S	18S RNA	Hs99999901_s1
actb	actin, beta, cytoplasmic	Mm00607939_s1
Arbp	acidic ribosomal phosphoprotein P0	Mm00725448_s1
Arbp	acidic ribosomal phosphoprotein P0	Mm01974474_gH
GAPDH	glyceraldehyde-3-phosphate dehydrogenase pseudogene	Mm99999915_g1
Gusb	glucuronidase, beta	Mm00446954_g1
HPRT1	hypoxanthine guanine phosphoribosyl transferase 1	Mm03024075_m1
Pgk1	phosphoglycerate kinase 1	Mm00435617_m1
Ppia	peptidylprolyl isomerase A	Mm02342429_g1
TBP	TATA box binding protein	Mm01277045_m1
Tfrc	transferrin receptor	Mm00441941_m1

Table 6, Continued.

High Density Array GE2

As Determined by GO Terminology		*Assay ID*
Dnmt3a	DNA methyltransferase 3A	Mm00463987_m1
Dnmt3L	DNA (cytosine-5-)-methyltransferase 3-like	Mm00457635_m1
Rag1	Recombination-Activating Gene 1	Mm01270936_m1
Rag2	Recombination-Activating Gene 2	Mm01270938_m1
As Described in Jaroudi and SenGupta, 2007		
Alkbh8 (Alkb)	alkB, alkylation repair homolog 8 (E. coli)	Mm01251182_m1
Alkbh8 (Alkb)		Mm01251184_m1
Atm	ataxia telangiectasia mutated homolog (human)	Mm01177457_m1
Atm		Mm00431867_m1
Bach1 (Fancj)	BTB and CNC homology 1	Mm01344527_m1
Blm	Bloom syndrome homolog (human)	Mm00476150_m1
Blm		Mm01317898_m1
Brca1	breast cancer 1	Mm01249844_m1
Brca1		Mm01249836_g1
Brca2	breast cancer 2	Mm01218740_g1
Brca2		Mm00464784_m1
Chaf1b	chromatin assembly factor 1, subunit B (p60)	Mm01215604_g1
Chek1	checkpoint kinase 1 homolog (S. pombe)	Mm01176761_g1
Chek1		Mm01176757_m1
Chek2	CHK2 checkpoint homolog (S. pombe)	Mm00443839_m1
Dclre1b (Pso2)	DNA cross-link repair 1B, PSO2 homolog (S. cerevisiae)	Mm00505657_m1
Dclre1b (Pso2)		Mm00505656_m1
Ddb1	damage specific DNA binding protein 1	Mm00497163_g1
Ercc1	excision repair cross-complementing rodent repair deficiency, complementation group 1	Mm00468337_m1
Ercc2 (Xpd)	excision repair cross-complementing rodent repair deficiency, complementation group 2	Mm01307194_g1
Fanca	Fanconi anemia, complementation group A	Mm00516855_m1
Fanca		Mm01243361_g1
Fancc	Fanconi anemia, complementation group C	Mm00514846_m1
Fance	Fanconi anemia, complementation group E	Mm00511654_m1
Fancl	Fanconi anemia, complementation group L	Mm00840321_m1
Fen1	flap structure specific endonuclease 1	Mm01700195_m1
Gtf2h1	general transcription factor II H, polypeptide 1	Mm01202628_m1
Gtf2h2	general transcription factor II H, polypeptide 2	Mm00502499_g1
Gtf2h3	general transcription factor IIH, polypeptide 3	Mm01199634_g1
Gtf2h4	general transcription factor II H, polypeptide 4	Mm00501678_m1
H2afx	H2A histone family, member X	Mm00515990_s1
Hdh (Hap1)	huntingtin-associated protein 1	Mm00468825_m1
Hus1	Hus1 homolog (S. pombe)	Mm01187812_g1
Lig3	ligase III, DNA, ATP-dependent	Mm01309678_m1
Lig3		Mm01303107_m1
Mbd4	methyl-CpG binding domain protein 4	Mm01184338_m1
Mbd4		Mm01184342_m1
Mlh3	mutL homolog 3 (E coli)	Mm01302907_m1
Mms19L	MMS19 (MET18 S. cerevisiae)	Mm00472208_m1
Mms19L		Mm01194228_g1
Mnat1	menage a trois 1	Mm01290617_m1
Mpg	N-methylpurine-DNA glycosylase	Mm01193430_m1
Msh4	mutS homolog 4 (E. coli)	Mm01320231_m1
Msh5	mutS homolog 5 (E. coli)	Mm01132458_g1
Msh5		Mm00488974_m1
Nbn (Nbs1)	nibrin	Mm00449854_m1
Neil3	nei like 3 (E. coli)	Mm00467593_g1
Pcna		Mm00448100_g1
Pms1	postmeiotic segregation increased 1 (S. cerevisiae)	Mm01254621_m1
Polb	polymerase (DNA directed), beta	Mm00448234_m1
Polq	polymerase (DNA directed), theta	Mm01170059_m1
Polq		Mm01170070_g1
Prkdc	protein kinase, DNA activated, catalytic polypeptide	Mm00465092_m1
Prkdc		Mm00465065_m1
Prkdc		Mm01342967_m1
Rad17	RAD17 homolog (S. pombe)	Mm01288365_g1
Rad18	RAD18 homolog (S. cerevisiae)	Mm00451706_m1
Rad23b	RAD23b homolog (S. cerevisiae)	Mm00772280_m1
Rad50	RAD50 homolog (S. cerevisiae)	Mm00485504_m1
Rad50		Mm00485491_g1
Rad51	RAD51 homolog (S. cerevisiae)	Mm01337943_m1
Rad51c	Rad51 homolog c (S. cerevisiae)	Mm01307097_m1
Rad51L3 (Rad51d)	RAD51-like 3 (S. cerevisiae)	Mm01303086_m1
Rdm1 (Rad52b)	RAD52 motif 1	Mm00487918_g1
Rdm1 (Rad52b)		Mm00481760_g1
Rpa1	replication protein A1	Mm00499562_g1
Rpa1		Mm01253368_m1
Shfm1 (Dss1)	split hand/foot malformation (ectrodactyly) type 1	Mm01162165_m1
Smug1	single-strand selective monofunctional uracil DNA glycosylase	Mm00452896_g1
Spo11	sporulation protein, meiosis-specific, SPO11 homolog (S. cerevisiae)	Mm00488871_m1
Tp53	transformation related protein 53	Mm01731287_m1
Tp53		Mm00441964_g1
Ube2n	ubiquitin-conjugating enzyme E2N	Mm00779119_s1
Ung	uracil DNA glycosylase	Mm01201513_m1
Wrn	Werner syndrome homolog (human)	Mm00449247_g1
Xrcc1	X-ray repair complementing defective repair in Chinese hamster cells 1	Mm00494222_m1
Xrcc1		Mm00494232_g1
Xrcc2	X-ray repair complementing defective repair in Chinese hamster cells 2	Mm00445118_m1
"Develomental Genes" from Gene Expression Database		
Nab1	First detected by whole mount in-situ at E8.5 (See Mechta-Grigoriou et al. 2000)	Mm00476263_m1
Nab2	First detected by whole mount in-situ at E8.0 (See Mechta-Grigoriou et al. 2000)	Mm00476267_m1
GATA2	First detected by whole mount in-situ at E9.0, restricted (see Nardellia et al. 1999)	Mm00492299_g1
Standard Endogenous Controls		
18S	18S RNA	Hs99999901_s1
actb	actin, beta, cytoplasmic	Mm00607939_s1
Arbp	acidic ribosomal phosphoprotein P0	Mm00725448_s1
Arbp	acidic ribosomal phosphoprotein P0	Mm01974474_gH
GAPDH	glyceraldehyde-3-phosphate dehydrogenase pseudogene	Mm99999915_g1
Gusb	glucuronidase, beta	Mm00446954_g1
HPRT1	hypoxanthine guanine phosphoribosyl transferase 1	Mm03024075_m1
Pgk1	phosphoglycerate kinase 1	Mm00435617_m1
Ppia	peptidylprolyl isomerase A	Mm02342429_g1
TBP	TATA box binding protein	Mm01277045_m1
Tfrc	transferrin receptor	Mm00441941_m1

For this study, hybrid embryos of diverse genotypic backgrounds are most useful for detecting gene expression signatures associated with genome instability. This allows for the identification of signatures indicative of hybrids in general, or for specific anomalies, observed in only a few individuals. I set up 64 crosses consisting of *M. m. domesticus* and *M. m. musculus* controls, F1 hybrid crosses, intercrosses and backcrosses (Table 7). Mice are most active during the night, and mating events can be detected in the morning by the presence of vaginal plugs. Noon on the day of plug detection is marked as E0.5. Dissections were performed at noon eight days later, i.e. E8.5. In total, 18 crosses produced at least 1 embryo, providing 113 embryos in total (Table 5).

Table 7. Crosses used for obtaining E8.5 hybrid embryos

Summary of All Crosses for E8.5 Embryos

Numer/crosses	Father	Mother	Successful Matings (embryos/female)
Controls			
16	D	D	3 (6, 6, 8)
16	M	M	2 (8, 11)
F Crosses			
9	M	D	3 (2, 8, 9)
1	D	M	1 (1)
Intercrosses			
2	A	A	
2	B	B	2 (6, 8)
2	B	A	
Backcrosses			
1	A	M	-
2	A	D	-
2	B	D	1 (8)
4	D	A	4 (5, 5, 8, 3)
1	D	B	1 (6)
3	M	A	1 (5)
1	M	B	-
Other			
1	F2	B	-
1	M	F2	-

D = Mm domesticus, M = Mm musculus, A = F1 Hybrid with Mm musculus paternity, B = F1 Hybrid with Mm domesticus paternity

Details of Successful Crosses

Family #	Mating Code*	Father	Mother	Offspring used in qPCR analysis
Mm domesticus Controls				
1	DD	CB05F1b	CB09F1b	Dom-1, Dom-4, Dom-8
2	DD	CB107F1b	CB109F1b (#796)	Dom-2, Dom-5
3	DD	CB101F1a (#599)	CB110F2a	Dom-3, Dom-6, Dom-7
Mm musculus Controls				
1	MM	VUB21.27	MHC25.6	Mus-1, Mus-3, Mus-5, Mus-7
2	MM	MHC9.7	MHC1.1	Mus-2, Mus-4, Mus-6, Mus-8
F1 Hybrids				
1	MD	MHC9.7	CB101F2a	F1-1, F1-2, F1-3
2	MD	MHC9.7	CB101F2a	F1-4, F1-5
3	MD	MHC1.2	CB101F2a	
4	DM	CB101F1a (#599)	Vi1F1a1 (#127)	
Intercrosses				
1	BB	RH6F1a2 (#110)	RH5F1b1 (#107)	IC-1, IC-3
2	BB	RH6F1a2 (#110)	RH5F1b1 (#108)	IC-2, IC-4
Backcrosses				
1	BD	RH5F1b2 (#33)	CB07Fa (#323)	BC-1, BC-3, BC-11
2	DA	CB101F1a (#599)	RH1F1b1 (#16)	BC-2, BC-4
3	DB	CB104F1b	RH5F1b3	BC-6, BC-9
4	MA	VUB21.27	RH4F1c (#30)	BC-7, BC-10
5	DA	CB05F1b (#294)	RH3F1a1	BC-8
6	DA	CB101F1a (#599)	RH1F1b2 (#17)	BC-5
7	DA	CB05F1b (#294)	RH2F1b2 (#21)	

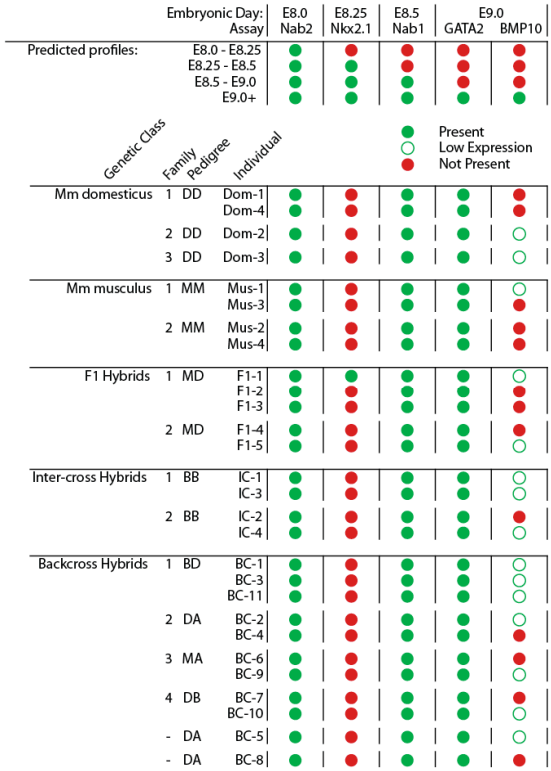
*The genotypic class of the father followed by the mother.

3.4 A Description of Embryonic DNA Repair Pathways

I began with an initial round of qPCR using 28 individuals (Table 7, Fig. 14B). The first priority was to determine which of the 11 ECs is most reliable. The most important requirement of an EC is that the target gene should be expressed at a consistent and adequately high level, independent of the biological state of the sample. The easiest way to determine this is a simple calculation of standard deviation of the Ct for each EC assay among all 28 individuals. Because the EC probes are present on both plates, there are up to 56 data points for each assay for analysis. The assay with the lowest standard variation, a surprisingly low 0.30 Ct, was Arbp-Mm00725448_s1. This assay was used for all subsequenct ΔCt calculations.

The next concern to address before comparing expression profiles was to examine group embryos based on their "developmental age" gene set. The purpose of this was so that later comparisons would only be between the most closely age-matched embryos. These six genes were chosen because their expression profiles would allow me to determine the precise age of the embryo. Although some genes profiles are as expected (e.g. the absence of *Rdh16*, which is expressed late in development), most are inconsistent with the predicted patterns (Table 8). Most of the problems in this assay arise because the original expression patterns of these genes were mostly elucidated by *in situ* hybridization, a technique with much less sensitivity than qPCR. Therefore, genes may be detectable earlier than expected in my assay. The alternative method of staging embryos is by morphological features (Kaufmann, 1992). However without in-house expertise in this field, plus the time pressure during dissection (to preserve mRNA) it is, unfortunately, an impractical option.

Table 8. Profiles of "developmental genes" in test populations.



To avoid complications involving variation in developmental age, I chose to focus on sibling-sibling comparisons. Comparing siblings is ideal, because variation in expression profiles cannot be attributed to differences in age. Divergent expression

profiles between hybrid siblings could be an indicator of an infrequent event and would be in agreement with my previous results that CNV destabilization does not occur in every hybrid individual. The underlying logic relies on unique sets of epistatic interactions, which occur in only a few individuals, leading to genomic destabilization through a reduced capacity for DNA repair. In other words, hybrid siblings, as a group, should encompass more variation in gene expression than pure-bred controls, just as hybrid populations encompass more copy number variation than pure-bred controls.

Principle Component Analysis (PCA) is an ideal method to survey for divergent gene expression profiles. PCA is a data reduction algorithm that is well suited for datasets having many more measurement points than individuals (Ringnér, 2008). A principal component (i.e. axis) is a direction along which the variation in data is maximal. Using a small number of components, each sample can be represented by a few numbers which retain most of the variation in the original data set. Furthermore, the percentage of the original variance retained by each component can be measured. PCA is *not* a statistical measurement, and therefore does not provide information as two how significantly different two groups are. PCA is not concerned with grouping, as it is designed only to identify directions with the largest variation, not directions relevant for separating groups. However, if groups can be distinguished along the first axis, then there is clear evidence for substructure within the data set. At this point it is useful to examine the weight of each measurement (i.e. gene expression value in ΔCt) for each component. Weights are centered around zero and those farthest away carry the greatest weight. A higher weight corresponds to a greater contribution of that measurement point to the overall variation on the axis in question.

I applied PCA to the expression profiles of the 24 individual for which at lest one sibling was present in my initial qPCR dataset. This represented 11 families and 5 genotypic classes (Figure 14A & B). 157 assays were included in the analysis, excluding those which did not work for all individuals, as these cannot be used in a PCA analysis. I used ΔCt (= threshold cycle of experimental – threshold cycle of endogenous control) values for the PCA analysis. Recall, a higher ΔCt is lower expression and vice versa. This test revealed that on the first axis, two individuals are clearly separated from the central cluster of all other samples: IC-2 and BC-1, an intercross and backcross hybrid,

respectively. The positive score for IC-2 is a reflection of a strong downregulation of many genes and likewise, BC-1 has many upregulated genes. Not only are these two individuals separated from the central cluster, but they are also a considerable distance away from their siblings, more than other sibships in the dataset. Additionally, *M. m. domesticus* and *M. m. musculus* individuals are separated only on the second axis, indicating that the divergence of these two hybrid individuals, which is captured on the first axis, encompasses more variation than even the difference between the two purebred subspecies. The F1 hybrids tend to cluster with *M. m. domesticus* and the intercross and backcross hybrids are scattered between the purebreds on the second axis. The first to components retain a respectable 61.8% of the original variance.

One concern with PCA is that size changes can result in outliers. For example, if all measurements are inflated by 30%, that individual will be clearly distinguishable on a PCA plot. This may be of particular concern for qPCR data because a shift in the endogenous control Ct can cause something similar to a size change, all ΔCt values would increase or decrease, resulting in a global change in expression. IC-2 does appear to have a general downregulation of genes, however this is not likely an artifact of a change in EC Ct for several reasons. Firstly, the relative fold change between siblings (i.e. IC-2 versus IC-4) for each assay ranges from 10% to 110%. If the endogenous control Ct had shifted, the relative fold change for all assays would have to be the same. Secondly, the Ct of the endogenous control is consistent with what is observed in all other samples, including IC-4, which I controlled by choosing the EC with the lowest standard deviation. Thus the downregulation in IC-2 is not a systematic artifact.

The identification of outlier hybrids on the first PCA axis is an encouraging initial result. Therefore, I expanded my data set by introducing technical replicates and additional control samples in a second round of qPCR experiments. This brought the total number of samples to 34 individuals, 16 of which were controls and four hybrids present as duplicates (summarized in figure 14B). I used this expanded data set to build another PCA plot, keeping the technical replicates separate to compare how closely they plotted to each other (Fig 14C). This PCA used 153 assays of the 164 available experimental assays; the 11 excluded assays are marked in Table 7.

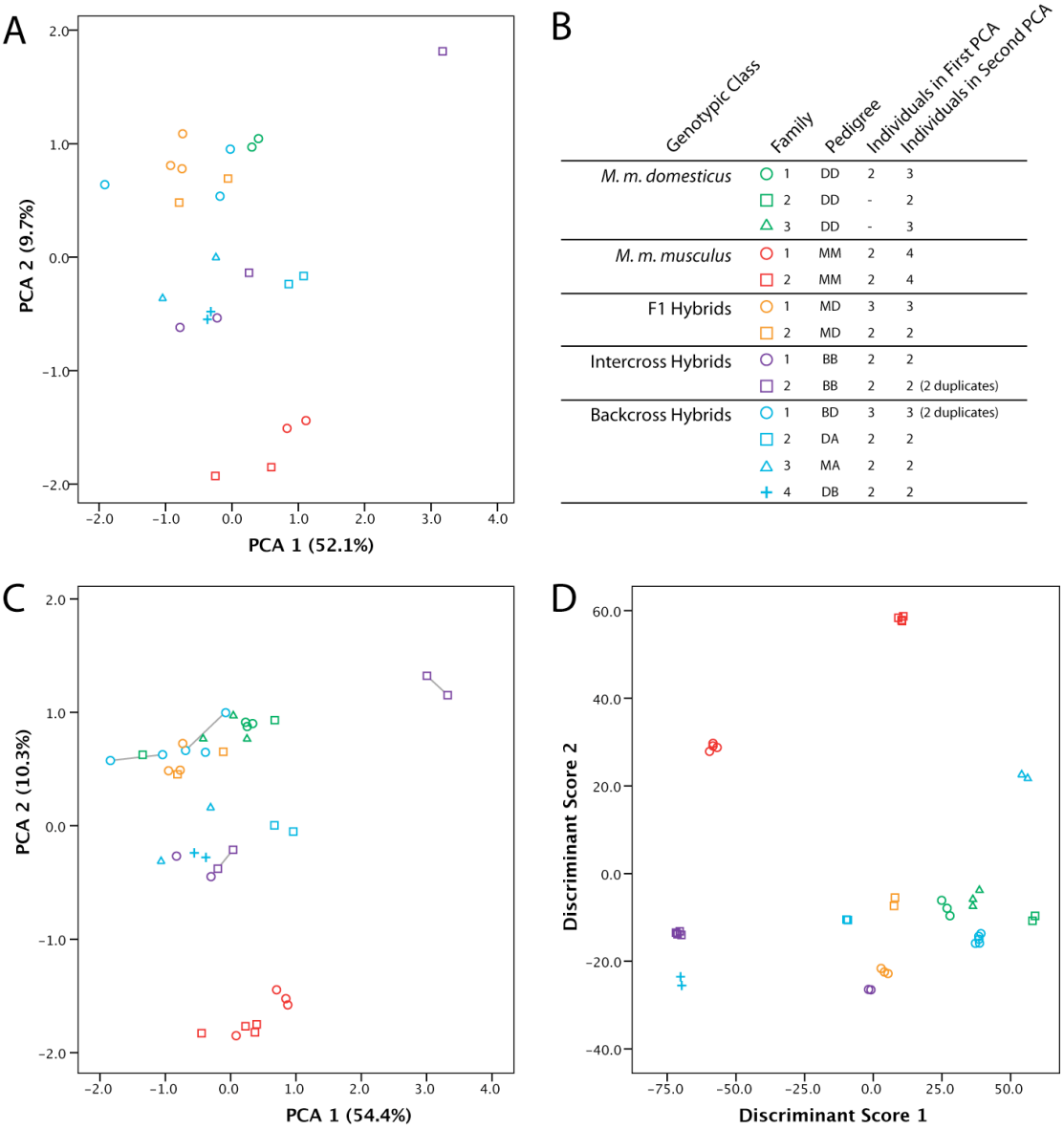


Figure 14. PCA and Discriminant Analysis Plots.
A,C) PCA plots for the first and second rounds of qPCR assays. B)Individuals used in each PCA plot: For pedigree, parental genotypes are listed (father left, mother right), D=*M. m. domesticus*, M=*M. m. musculus*, B=Type B F1 hybrid (having *M. m. domesticus* paternity, i.e. Offspring from a DM cross), A= Type A F1 hybrid (having *M. m. musculus* paternity, i.e. Offspring from a MD cross). D) Discriminant Analysis plot using individuals from the second round of qPCR.

PCA is very much dependent on input, and placement of individuals can change depending on the measurements used. With the expanded data set, the BC-1 lies within the central cluster on the first axis, and there appears to be quite some variation in the second *M. m. domesticus* (number 2) in particular. However, both IC-2 replicates remain distinguishable as outliers on the first axis. Therefore, understanding the gene weights for this component are useful in understanding what changes in gene expression contribute

to IC-2 as an outlier (Fig 14C). For this, I considered those genes which have weights ±0.90 on the first component (Table 9, Fig. 15). The use of PCA to identify the top contributors in this study is also significant because it is an unbiased approach. As mentioned, many genes are downregulated in IC-2 compared to its sibling. One would traditionally consider only those genes which are downregulated beyond a certain threshold to be biologically significant, however the true underlying biological significance of such changes would be known for only a handful of those genes and as such the method is very arbitrary. PCA considers all the measurements of all individual and provides two pieces of information: how are individuals related to each other based on all measurements, and which measurements contribute most to the overall pattern of variation. This is quite a different perspective on gene expression analysis than regarding only differential expression as having some inherent significance. Recently PCA has been applied to microarray gene expression data as a way to compliment traditional statistical practices which survey for the greatest change in gene expression between two well defined experimental groups (Raychaudhuri et al., 2000).

Table 9. High-Weight Genes of the first principle component axis.

Gene	Assay ID	Plate	Weight
Xrcc1	Mm00494222_m1	GE2	0.96
Xrn2	Mm01275968_m1	GE1	0.95
Atr	Mm01223626_m1	GE1	0.93
Rpa1	Mm01253368_m1	GE2	0.93
Gtf2h4	Mm00501678_m1	GE2	0.93
Gtf2h2	Mm00502499_g1	GE2	0.92
Rpa1	Mm00499562_g1	GE2	0.92
Chaf1b	Mm01215604_g1	GE2	0.92
Rfc5	Mm01208090_g1	GE1	0.91
Dnmt1	Mm01151065_g1	GE1	0.91
Tdg	Mm00834243_g1	GE1	0.91
Hus1	Mm01187812_g1	GE2	0.90
Alkbh8	Mm01251184_m1	GE2	0.90
Sumo1	Mm01609844_g1	GE1	0.90

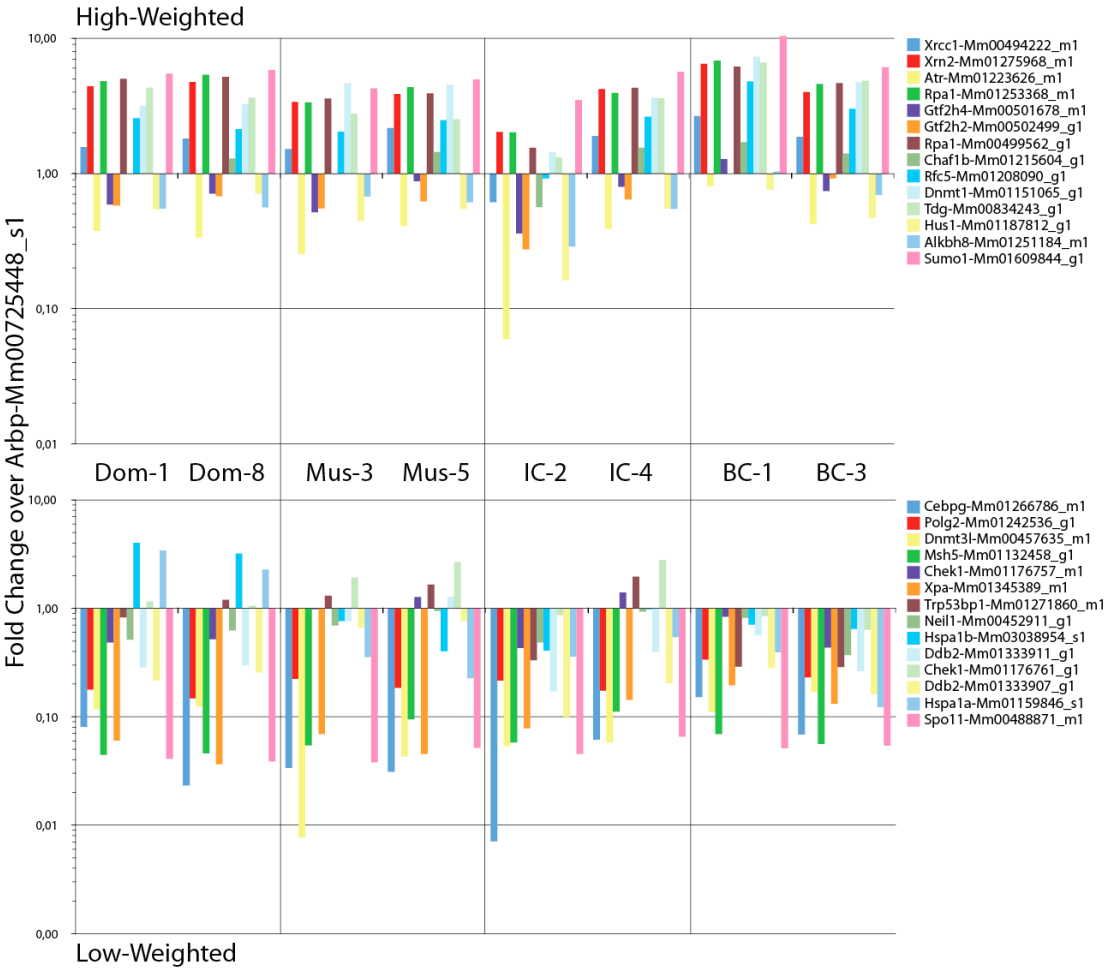


Figure 15. Highest- and Lowest-Weighted Genes on the First PCA Axis. Each vertical section displays the expression value (calculated as fold change to the EC gene *Arbp*) of the highest- (top) and lowest-weight (bottom) genes for the first PCA axis (see Fig 14C) for two siblings each of *M. m. domesticus*, *M. m. musculus* intercross and backcross origins. Of the high weighted genes, IC-2 shows a decrease in expression for all of them compared to its sibling, IC-4. BC-1 has mostly higher expression than its sibling BC-3. The expression between pure-bred siblings is more consistent. Most of the low-weight genes have very low expression levels and therefore did not contribute much to the overall variation in the data set.

The expanded, second PCA reveals that *M. m. domesticus* is more variable than expected. In order to reveal what contributes to the variation in expression profiles of the pure-breds, I recalculated the PCA using only the pure-bred control individuals. Variable gene expression among the control population can be considered neutral. Therefore if a gene carries high weight on the first axis in this control-only PCA and in the PCA with all individuals, it may be of less significance because they likely represent natural variation in gene expression between individuals during development. Four genes have

weights in the first component beyond ±0.90 in both PCAs (with or without hybrid samples): *Xrn2*, *Rcf5*, *Dnmt1* and *Alkbh8*.

Another test that can be done on this data set is a discriminant analysis. Discriminant analysis builds a predictive model for group membership based on linear combinations of predictor variables that provide the best discrimination between groups. Essentially this reveals how confidently individuals in a data set group given information as to the total number of groups but not membership within each group. Using this test all siblings grouped tightly together (Fig 14D), including those that are most divergent by PCA, and group identity was 100% for each of the 13 groups. This indicates that although there is substantial difference between certain siblings, they remain recognizably related.

3.5 Discussion

This study is more focused on discovery rather than precise mechanism elucidation. A mechanistic framework can only be speculated about at this time, but the collection of high-weight genes for the first PCA axis are quite interesting. What is already known about the way these genes interact, their mutational phenotypes and role in genome integrity, hints at a plausible picture of events leading to CNV destabilization.

The highest-weight gene on the first PCA axis is *Xrcc1* (Table 9, Fig. 15). XRCC1 is a non-emzymatic scaffold protein involved in the resolution of SSB repair and has been studied intensively for its involvement in two DNA repair pathways, break-excision repair (BER) and sister chromatid exchange (SCE) (Thompson and West, 2000). A brief review of the relationship between these two repair pathways will help in understanding how they related to CNV destabilization during embryogenesis.

During organogenesis, oxidative damage by ROS is the most prevalent threat to the genome (Vinson and Hales, 2002). ROS are one of the most common causes of SSBs (Caldecott, 2008). SSBs are recognized by the catalytic zinc-finger domain of the *poly(ADP-ribose) polymerase-1* gene product (PARP1), which relays this information to the cell by poly(ADP-ribosyl)ation of histone H1 and H2B (Dantzer et al., 2006). This epigenetic modification allows the chromatin structure to become relaxed, facilitating

accessibility of repair proteins. PARP1 also undergoes auto-poly(ADP-ribosyl)ation which quickly attracts XRCC1 to the damaged site (Okano et al., 2003). *Xrcc1* is one of the most important members in the repair of damaged bases by BER, which produces SSBs as an intermediate step (Caldecott, 2008). These SSBs can remain unresolved if the proper repair components, gathered around the XRCC1 scaffold protein, are not properly assembled. If SSBs are not immediately repaired, they develop into DSBs during replication (Kuzminov, 2001) in which case they must be repaired by SCE, which is essentially homologous recombination between two newly synthesized sister chromatids (Fig. 16). Current estimates of SCE indicate that around 10 DSBs occur per cell division at the replication fork (Haber, 1999).

Because of it's central role in SSB repair, fully functioning *Xrcc1* is a critical link between ROS and DSB repair via SCE during organogenesis. Significantly, *Xrcc1*$^{-/-}$ mutants display a startling high increase in the frequency of spontaneous and damage-induced SCE (Wilson and Thompson, 2007), which appear to arise from homologous recombination at sites of replication-derived DSBs. This process is sometimes referred to as replication-coupled single stranded break repair (RC-SSBR, Fig. 16) (Caldecott, 2003). However, a description of unequal SCE in this process, where strand invasion of the sister chromatid is not completely homologous (i.e. like NAHR), has never been considered as far as I am aware.

The effect of *Xrcc1*$^{-/-}$ null mutations have also been surveyed in mouse embryogenesis where developmental arrest occurs at E6.5 (Tebbs et al., 1999). Thus it is unlikely the *Xrcc1* activity is completely abolished in the samples I am examining, as E8.5 stage embryos would not be recoverable. Additionally, *Xrcc1* is represented by two TaqMan assays in this experiment. The second assay, Mm00494232, recognizes a splice variant with an alternate 3' end and has a gene-weight of 0.70 (rank 101 of 153 assays). This difference in weighting may indicate differential roles of these transcripts.

The connection to single-stranded DNA is also observed in other high-weight genes from the PCA. *Rpa1*, which is represented by two assays, is an essential contributor to the early stages of DSB repair. One of the first steps in DSB repair is the resection of the 5' end, exposing a 3' tail which is then coated with the heterotrimetic Rpa protein

complex, of which *Rpa1* is the largest subunit (Wold, 1997). Rpa coated ssDNA attracts ATR to sites of DNA damage (Zou and Elledge, 2003).

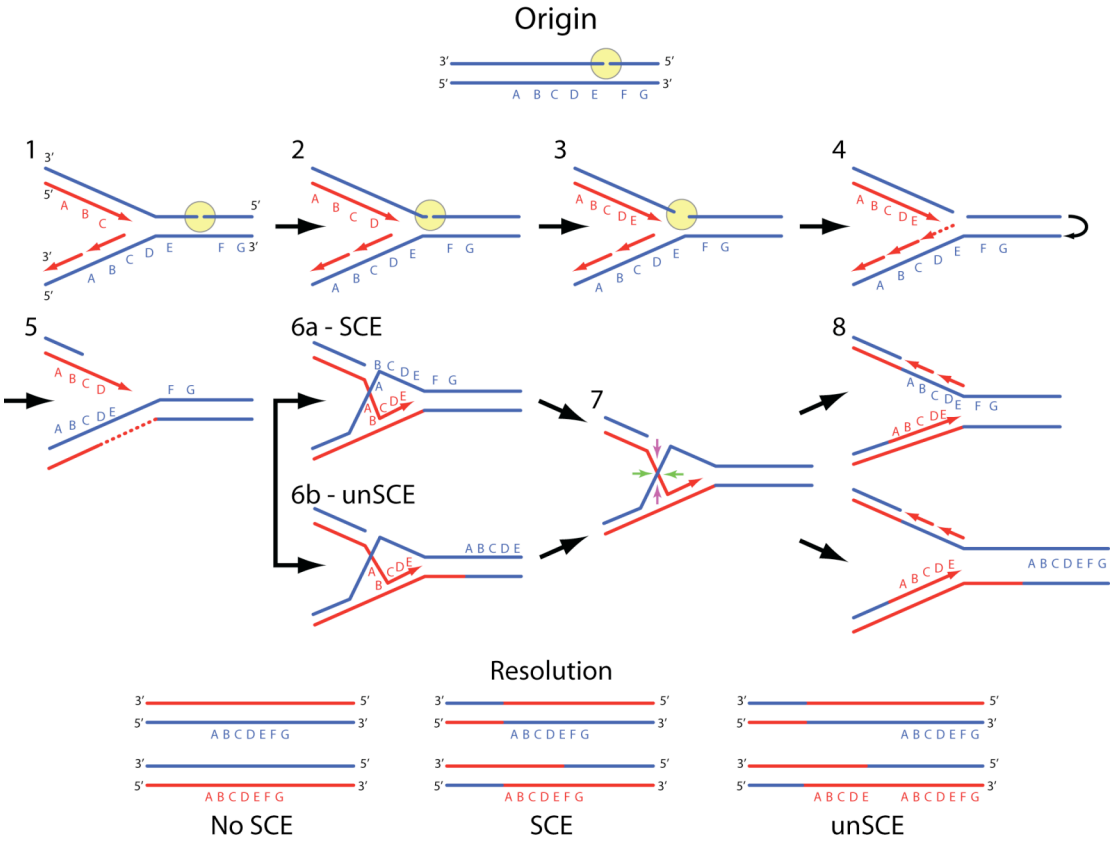


Figure 16. Replication-Coupled Single-Stranded Break Repair Schema
A) RC-SSBR occurs when replication fork encounters an unresolved SSB or gap. Replication progresses (1,2) until the replication fork encounters an SSB or gap (yellow circle) and breaks (3). DNA synthesis continues on the unbroken chromatid (4), The curved black arrow represents a conformational change to facilitate visualization of subsequent events) and the 5' end of the broken strand is resected (5), revealing a 3' single-strand tail. The 3' tail, invades (6a) the sister chromatid to initiate repair. Resolution of the Holliday junction at the green arrows results in SCE, but not at the purples arrows. The replication fork is restored and systhesis continues (8). The critical point for unequal SCE is during strand invasion, as with NAHR, it is possible that stretches of homology cause strand invasion at the wrong loci (6b). Replication after unequal SCE would result in a net gain of DNA (see resolution), in contrast to the overall copy number neutral NAHR in which amplifications are balanced against deletions on another chromosome (Image adapted from Wilson and Thompson, 2007)

Atr, another high-weight gene, represents one of the major DNA damage response pathways and, like XRCC1 also interacts with PARP1 (Kedar et al., 2008). Although *Atr* expression is quite low in IC-2, it still appears to still be functional, as complete lack of ATR results in early embryonic lethality (Brown and Baltimore, 2000; de Klein et al., 2000). It appears that the endogenous role of ATR is in the recovery of

stalled replication forks at fragile sites. Fragile sites are chromosomal regions that are particularly difficult to replicate or recover from replication fork collapse (Glover et al., 2005). In the absence of ATR, stalled replication forks collapse and DSBs accumulate (Paulsen and Cimprich, 2007). Accordingly, ATR-deficient cells have high levels of fragile site breakage (Casper et al., 2002). Significantly, ATR is involved in DSB repair but is recruited by ssDNA, which occurs at stalled replication forks when the 5' end of the DSB has been resected (O'Driscoll and Jeggo, 2006). This is in direct contrast to the other major HR sub-pathway (ATM) which is attracted directly to DSB ends.

Another high-weight gene, *Hus1*, shares a functional relationship with *Atr* as a member of the S-phase DNA damage checkpoint. Like, *Atr*, *Hus1* deficiency results in an increased frequency of fragile site instability (Zhu and Weiss, 2007). *Hus1* null mutants result in genomic instability and embryonic lethality (Weiss et al., 2000). Interestingly both *Rpa1* and *Hus1* are downstream effectors of the ATR kinase signaling cascade (O'Driscoll and Jeggo, 2006).

Chaf1b is yet another high-weight gene operating at the S-phase checkpoint. *Chaf1b* is specifically involved in DNA replication-dependent nucleosome assembly (Kaufman et al., 1995) and is necessary for S-phase progression in mammalian cells (Hoek and Stillman, 2003).

With the presence of *Gtf2h4* and *Gtf2h2* (general transcription factor II H, polypeptide 4 and 2) as high-weight genes, our attention turns to RNA transcription. These two genes encode subunits of the transcription factor TFIIH complex which has a role not only in transcription but also transcription-coupled repair (TCR) (Hanawalt and Spivak, 2008). TCR occurs when a DNA nick is encountered by RNA polymerase II during transcription instead of DNA polmerase during replication. TFIIH interacts with the RNA polymerase II-DNA-RNA complex (Tantin, 1998), remodeling the stalled RNA polymerase II in an ATP-dependent manner (Sarker et al., 2005). During TCR, RNA polymerase II is dislodged from the damaged DNA, the nascent RNA strand is dissociated, and, once again, replication-based DNA repair is invoked. A link between TCR and oxidative damage has been made using plasmids, although a direct link with genomic DNA has yet to be published (Hanawalt and Spivak, 2008; Spivak and Hanawalt, 2006).

As discussed previously, four genes also had high-weight on the pure-bred only PCA: *Xrn2*, *Rfc5*, *Dnmt1* and *Alkbh8*. Of these, *Rfc5* is the only one to bear any direct functional similarity to these genes described above. It is also an S-phase checkpoint gene and yeast mutants have a large increase in genome rearrangements (Myung et al., 2001; Myung and Kolodner, 2002; Naiki et al., 2000). *Xrn2* is a 5' -> 3' exonuclease, with a role in polymerase II termination (West et al., 2004). DNMT1 is a DNA methyl transferase and, like XRCC1, directly interacts with PARP1 (Reale et al., 2005).

As mentioned, PARP1, a central component of SSB repair, interacts with some of the high-weight genes; but *parp1* is also interesting because it is itself *not* a high-weight gene. *parp1* activity is transiently activated by DSBs in the chick DT40 cell line (Sonoda et al., 2006). This results in the inhibition of Ku protein (essential for NHEJ) from binding DNA, and facilitation of HR pathways. In the absence of PARP1, Ku affinity for DNA increases and HR efficiency is reduced. Thus *parp1* is crucial for directing the repair mechanism at DSBs. In my assay, expression of *parp1*, and the closely related *parp2*, are surprisingly variable. Among *M. m. domesticus* embryos, fold-change over the endogenous control ranged from 2.6-5.1 and 0.9-2.0 for *parp1* and *parp2*, respectively. In *M. m. musculus*, a fold change range of 3.0-4.5 and 1.0-1.8 is observed. Interestingly, the expression range of the hybrid individuals overlaps with that of the purebred controls (2.3-5.5 and 06-2.5 for *parp1* and *parp2*, respectively). The lowest expression was, as expected, in IC-2 (2.3 and 0.6) but this is not as dramatic a many of the other downregulated genes, and variation among control pure-breds, even siblings, is also quite high. Therefore, it appears that this crucial link to the HR pathway persists.

Taken alltogether, the relationship between these genes hints at a shift in balance among the various sub-pathways of HR. When BER is not active SCE is required for the accurate repair of the abundant oxidative damage occurring during organogenesis. If critical BER components are insufficient, such as *Xrcc1*, SCE is invoked. However, if we imagine that several members of HR involved in SCE and RC-SSBR are also deficient, we may be presented with a scenario where abundant unequal SCE occurs, leading to gene amplification in the early stages of development (Fig. 16). Several rounds of unequal SCE at common fragile sites (i.e. recurring) could then add to amplification of specific loci.

4.0 Concluding Remarks

4.1 A General Summary

The study of genetics is, in essence, the study of variation: its origins and consequences. Here, I have presented evidence of a hybrid destabilization effect at CNV loci. This represents an increase in the mutation rate of hybrid individuals which results in new alleles (i.e. copy numbers). Further, I present the primary analysis that the causes of this destabilization could be based on misregulation of DNA repair mechanisms.

Although the results presented herein are surprising, they agree with the most recent shifts in thought in biology. For instance, it is clear that somatic mutations are abundant (Gottlieb et al., 2001) not only with SNPs, but also CNVs (Bruder et al., 2008; Liang et al., 2008; Piotrowski et al., 2008). CNVs are also known contributors to many complex diseases (de Vries et al. 2005; Jacquemont et al. 2006; Sebat et al. 2007; Marshall et al. 2008; Walsh et al. 2008). A further elucidation of the causes of destabilization will be interesting in these respects.

4.2 Novel Alleles and Hybrid Speciation

This work offers a new perspective on reproductive isolation. Traditional studies have relied on the identification of incompatible loci leading to sterility, for example for the recently published speciation gene *Prdm9* in mice (Mihola et al., 2009). However, that the genome can undergo extremely fast and dramatic changes in hybrid individuals likely leads to complications in reproductive fitness, an important aspect of the speciation process.

Hybrids benefit from new allele combinations via recombination, but the possibility of novel alleles via mutation also exists. For instance, the "rare allele phenomenon" describes unique point mutations or allozymes specific to hybrid individuals (Bradley et al., 1993; Hoffman and Brown, 1995; Schilthuizen et al., 2001; Smith and Glenn, 1995). Therefore, an interesting perspective is that new hybrid-specific CNV alleles may also be heritable, and even possibly beneficial. Another interesting observation is that at least 10% of animal species readily hybridize in the wild (Mallet, 2005). However, the long term significance of novel hybrid alleles and the extent of these phenomena are, at present, poorly understood.

Another consideration for the long-term impact of CNV destabilization concerns their genic composition. Loci which are most susceptible to copy number mutations by NAHR, and ergo unequal SCE, are those which contain genes that operate at the molecular-environmental boundary and are able to withstand variation in copy number. Several studies have proposed adaptive benefits of CNVs, based on these genic biases (Tuzun et al. 2005; Conrad et al. 2006; McCarroll et al. 2006; Kidd et al. 2008). Hybrid speciation theory postulates that hybrids, in a new environment, may actually be more fit than the parental species (Burke and Arnold, 2001). It will be exciting to see if CNV destabilization plays a role in this process. These considerations provide a good foundation for further investigation into the mode of inheritance for new CNV alleles (i.e. those generated in the somatic lineage before gametogenesis or during meiosis) and possible positive selection on new variants.

4.3 Outlook

The two most distinguishing features of CNVs are their wide range in size and mutation rate size. I predict that these features will lead to studies exploring avenues of research not considered in traditional studies of genetic variation. Two areas will likely predominate: Somatic variation and micro-evolution. Somatic variation has already been documented within humans and given the growing number of human diseases in which somatic mutation plays a role, elucidating the contribution of somatic CNVs to disease will be a difficult but necessary task. Micro-evolutionary studies will benefit by theoretical models of CNV behaviour which will develop as the behaviour and composition of CNVs loci becomes better understood. Given their distinguishing features, it is surprising that CNVs are not more thoroughly studied in a micro-evolutionary context, in particular with respect to adaptive evolution. It appears that the full impact of this form of genetic variation is being largely ignored by researcher who are likely to be heavily, and largely beneficially, affected by it. The results presented in this thesis add to the growing knowledge of CNV mutation dynamics relevant to both areas.

5.0 Materials & Methods

5.1 Materials

The parental animals for the laboratory generated F1 hybrids were first generation individuals born from unrelated animals caught in the wild. The source of *M. m. domesticus* individuals was a Western German population (Cologne/Bonn area), the *M. m. musculus* individuals came from a population caught close to Vienna (obtained from K. Musolf, Konrad Lorenz Institute for Ethology). The *M. m. domesticus* (Germany and France) and *M. m. musculus* (Czech Republic and Kazakhstan) mouse population samples used for the hybrid zone comparisons represent unrelated individuals caught in the wild and have previously been described (Ihle et al., 2006). The wild hybrid mice were collected from a Bavarian transect and DNA was provided by R. Rottscheidt and B. Harr.

For the expression assay of DNA repair genes, embryos were dissected on noon eight days after a female was detected as plug-positive. Plugs were examined every morning between 8:00h and 10:00h. Two people were always involved in the dissections. I would dissect the embryos from the uterus and out of the extraembryonic sac in a 1x PBS solution (pH 7.4, Sigma). The second person would maintain a bowl of liquid nitrogen where the embryo would be places in as soon a possible. The flash-frozen embryos were then transferred to cryotubes pre-chilled on dry-ice. Embryos were stored at -80^{O}C until RNA extraction, see below.

5.2 cDNA Analysis

A 3' RACE kit (Invitrogen, Carlsbad, California) was used to obtain cDNA sequences of *Slx* and *Slx-like*. RNA samples were taken from two mice (Mus1 and Dom1) which showed divergent expression using a microarray platform (Voolstra et al., 2007). In this protocol, RNA is reverse transcribed using Superscript Reverse Transcriptase (Invitrogen) and a tagged poly(T) primer. The single stranded cDNAs are then amplified using a gene specific primer for Slx (ggtgcagttgtgaargtgttc) and the tag added during first-strand synthesis. Amplified sequences of the expected size were cut out of a gel, purified and cloned into a TOPO vector. Bacterial colonies were picked with

pipette tips after growing overnight and diluted into a 96-well plate filled with 20uL of ddH20. High-fidelity Phusion Hot Start taq (Finnzymes, Espoo, Finland) was used to amplify cloned inserts as per the manufacturer protocol. Sequencing was performed as the Cologne centre for genomics.

5.3 Quantitative PCR assays

Quantitative PCR (qPCR) assays were custom designed to determine copy number of *Slx*, *Slx-like* (*4930527E24Rik*), *Xlr* (collectively referred to as the *Slx* qPCR), *L19* and *Sly* gene regions. Given the large number of polymorphisms within these genes, I downloaded paralogous genomic DNA sequences of each gene, aligned them and manually searched for regions of reduced polymorphism. These regions corresponded to *Slx* exon V, the L19 3'UTR and *Sly* exon VII (see Fig. 3). A consensus sequence was compiled, masked and submitted to Applied Biosystems (ABI, Foster City, CA) for design of custom TaqMan assays. The primer and probe combinations provided by ABI are as listed:

Assay	Forward Primer	Reverse Primer	Probe
Slx	CAGGCCAGGCTGTGTTTATTTATG	AGGCATAGTGCCAACATTAGGTT	ATGGCAGCGTTTTGC
2° primers	ANTCAGAAAACGTAAGTTTCTCAGAGG	TTGCTGTTCACCACTTAACAAATTC	
L19	GATCCAAAGCATTGCTGCATATT	CATCTGCCATTGAGGGATGTGAT	ATCCCAGGAAATTTC
Sly	AGAGAAAATGGATGGAAACTTATGTCAAAGA	CTCTCGTTCGTTCGTTTTGCA	CAGCAACCAGAAATT

For *Slx*, the original primer combination amplified a 1.2Kb fragment, normally outside the range for qPCR to function properly and so the 2^{O} primers were added to the assay for a final concentration of 900nM on the recommendation of ABI. qPCR required the use of an endogenous control (EC). For this purpose I chose the ready-made *etd* assay (Mm00558327_s1), a single-copy X-linked gene. Genomic DNA samples were treated with RNAse A to prevent contamination from RNA (although the tissues that we used for DNA extraction should not express these genes anyway). All samples were run in triplicates with high consistency within runs. If the standard deviation from three technical replicates was higher than 0.2 C_t an outlier was defined and removed. Outliers were removed for less than 18% of the technical triplicates. All assays were run on an ABI Prism 7900HT Sequence Detection System using 384 well plates and running SDS v2.1.1.

To validate the efficiency of the qRT-PCR assays, a dilution series (50, 25, 10, 5, 2.5, 1 and 0.5 ng/uL starting concentrations of *M. m. domesticus* DNA from a single individual) was conducted for each custom assay and the slope of a linear regression of the ΔC_t values, measured against the *etd* endogenous control, was calculated. Slopes for *Slx* (0.0312), *L19* (0.0461), and *Sly* (0.0992) are all within the accepted range ($m < 0.1$) for 100% efficient custom TaqMan assays, as suggested by the manufacturer.

Copy number was taken as the fold change over the endogenous control, and calculated for each individual by using $2^{-\Delta Ct \pm standard\ deviation}$. ΔC_t and the standard deviation were calculated as described by the manufacturer. Briefly, C_t is the threshold cycle, at which the fluorescent signal of a PCR reaction is first statistically above background. A higher copy number corresponds to a lower C_t. ΔC_t is calculated as the experimental C_t minue the EC C_t. Due to the nature of the assay, heterozygous CNVs were co-dominant and so we can only estimate that the fold change of a locus over *etd* represents an average of the two chromosomes in females. For the hemizygous gonosomes in males we can state the copy number per chromosome.

For gene expression analysis of DNA repair genes, two custom high density TaqMan arrays in format 96a were ordered with the assays listed in table 7. RNA samples were extracted with Trizol (Invitrogen) by B. Kleinhenz using the standard protocols with recommended modification for small tissue samples. SuperScript III Reverse Transcriptase (Invitrogen) was used to obtain cDNA from 5000ng of RNA in a 20uL reaction. To this, 30uL ddH_20 was added. Gene Expression MasterMix for qPCR was diluted 200uL:196uL with ddH_20 and to this 4uL of the diluted cDNA was added. The 400uL master mix was added to the high density arrays, centrifuged and run on the ABI Prism 7900HT Sequence Detection System as per standard protocol.

5.4 Southern Blotting

To obtain a *Slx* probe in the same region as the *Slx* qPCR assay, a 1.2 kb region was amplified from genomic DNA using the respective *Slx* primers (suppl. files), cloned into a TOPO cloning vector (Invitrogen) and sequenced. A DIG-labelled single-stranded RNA probe was generated using the T7 transcription start site as per standard protocol provided by Roche Applied Sciences (Indianapolis, Indiana). Detection was conducted

by chemiluminescence using CDP-STAR (Roche). Quantification of Southern blot signals was done using the ImageJ application provided by NCBI.

6.0 References

Adams, D. J., Dermitzakis, E. T., Cox, T., Smith, J., Davies, R., Banerjee, R., Bonfield, J., Mullikin, J. C., Chung, Y. J., Rogers, J. et al. (2005). Complex haplotypes, copy number polymorphisms and coding variation in two recently divergent mouse strains. *Nat Genet* **37**, 532-6.

Albertson, D. G. (2006). Gene amplification in cancer. *Trends Genet* **22**, 447-55.

Bavoux, C., Hoffmann, J. S. and Cazaux, C. (2005). Adaptation to DNA damage and stimulation of genetic instability: the double-edged sword mammalian DNA polymerase kappa. *Biochimie* **87**, 637-46.

Boissinot, S. and Boursot, P. (1997). Discordant phylogeographic patterns between the Y chromosome and mitochondrial DNA in the house mouse: selection on the Y chromosome? *Genetics* **146**, 1019-34.

Bradley, R. D., Bull, J. J., Johnson, A. D. and Hillis, D. M. (1993). Origin of a novel allele in a mammalian hybrid zone. *Proc Natl Acad Sci USA* **90**, 8939-41.

Brown, E. J. and Baltimore, D. (2000). ATR disruption leads to chromosomal fragmentation and early embryonic lethality. *Genes Dev* **14**, 397-402.

Bruder, C. E., Piotrowski, A., Gijsbers, A. A., Andersson, R., Erickson, S., de Ståhl, T. D., Menzel, U., Sandgren, J., von Tell, D., Poplawski, A. et al. (2008). Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am J Hum Genet* **82**, 763-71.

Burke, J. M. and Arnold, M. L. (2001). Genetics and the fitness of hybrids. *Annu Rev Genet* **35**, 31-52.

Caldecott, K. W. (2003). XRCC1 and DNA strand break repair. *DNA Repair (Amst)* **2**, 955-69.

Caldecott, K. W. (2008). Single-strand break repair and genetic disease. *Nature Reviews Genetics* **9**, 619-31.

Calenda, A., Allenet, B., Escalier, D., Bach, J. F. and Garchon, H. J. (1994). The meiosis-specific Xmr gene product is homologous to the lymphocyte Xlr protein and is a component of the XY body. *EMBO J* **13**, 100-9.

Cañestro, C., Yokoi, H. and Postlethwait, J. H. (2007). Evolutionary developmental biology and genomics. *Nat Rev Genet* **8**, 932-42.

Casper, A. M., Nghiem, P., Arlt, M. F. and Glover, T. W. (2002). ATR regulates fragile site stability. *Cell* **111**, 779-89.

Castro, J. P. and Carareto, C. M. (2004). Drosophila melanogaster P transposable elements: mechanisms of transposition and regulation. *Genetica* **121**, 107-18.

Chan, M. F., van Amerongen, R., Nijjar, T., Cuppen, E., Jones, P. A. and Laird, P. W. (2001). Reduced rates of gene loss, gene silencing, and gene mutation in Dnmt1-deficient embryonic stem cells. *Mol Cell Biol* **21**, 7587-600.

Chia, R., Achilli, F., Festing, M. F. W. and Fisher, E. M. (2005). The origins and uses of mouse outbred stocks. *Nat Genet* **37**, 1181-6.

Cohen, D. I., Hedrick, S. M., Nielsen, E. A., D'Eustachio, P., Ruddle, F., Steinberg, A. D., Paul, W. E. and Davis, M. M. (1985a). Isolation of a cDNA clone corresponding to an X-linked gene family (XLR) closely linked to the murine immunodeficiency disorder xid. *Nature* **314**, 369-72.

Cohen, D. I., Steinberg, A. D., Paul, W. E. and Davis, M. M. (1985b). Expression of an X-linked gene family (XLR) in late-stage B cells and its alteration by the xid mutation. *Nature* **314**, 372-4.

Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E. and Pritchard, J. K. (2006). A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* **38**, 75-81.

Cutler, G., Marshall, L. A., Chin, N., Baribault, H. and Kassner, P. D. (2007). Significant gene content variation characterizes the genomes of inbred mouse strains. *Genome Research* **17**, 1743-54.

Dantzer, F., Amé, J. C., Schreiber, V., Nakamura, J., Ménissier-de Murcia, J. and de Murcia, G. (2006). Poly(ADP-ribose) polymerase-1 activation during DNA damage and repair. *Methods in Enzymology* **409**, 493-510.

de Klein, A., Muijtjens, M., van Os, R., Verhoeven, Y., Smit, B., Carr, A. M., Lehmann, A. R. and Hoeijmakers, J. H. (2000). Targeted disruption of the cell-cycle checkpoint gene ATR leads to early embryonic lethality in mice. *Curr Biol* **10**, 479-82.

de Smith, A. J., Walters, R. G., Coin, L. J., Steinfeld, I., Yakhini, Z., Sladek, R., Froguel, P. and Blakemore, A. I. (2008). Small deletion variants have stable breakpoints commonly associated with alu elements. *PLoS ONE* **3**, e3104.

de Vries, B. B., Pfundt, R., Leisink, M., Koolen, D. A., Vissers, L. E., Janssen, I. M., Reijmersdal, S., Nillesen, W. M., Huys, E. H., Leeuw, N. et al. (2005). Diagnostic genome profiling in mental retardation. *Am J Hum Genet* **77**, 606-16.

Deans, B., Griffin, C. S., O'regan, P., Jasin, M. and Thacker, J. (2003). Homologous Recombination Deficiency Leads to Profound Genetic Instability in Cells Derived from Xrcc2-Knockout Mice. *Cancer Res* **63**, 8181-7.

Degenhardt, J. D., de Candia, P., Chabot, A., Schwartz, S., Henderson, L., Ling, B., Hunter, M., Jiang, Z., Palermo, R. E., Katze, M. et al. (2009). Copy number variation of CCL3-like genes affects rate of progression to simian-AIDS in Rhesus Macaques (Macaca mulatta). *PLoS Genet* **5**, e1000346.

Derothe, J. M., Porcherie, A., Perriat-Sanguinet, M., Loubès, C. and Moulia, C. (2004). Recombination does not generate pinworm susceptibility during experimental crosses between two mouse subspecies. *Parasitol Res* **93**, 356-63.

Dod, B., Jermiin, L., Boursot, P., Chapman, V., Nielsen, J. and Bonhomme, F. (1993). Counterselection on sex-chromosomes in the Mus musculus European hybrid zone. *J Evol Biol* **6**, 529-46.

Dumas, L., Kim, Y. H., Karimpour-Fard, A., Cox, M., Hopkins, J., Pollack, J. R. and Sikela, J. M. (2007). Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Research* **17**, 1266-77.

Egan, C. M., Sridhar, S., Wigler, M. and Hall, I. M. (2007). Recurrent DNA copy number variation in the laboratory mouse. *Nat Genet* **39**, 1384-9.

Ellis, P. J., Clemente, E. J., Ball, P., Touré, A., Ferguson, L., Turner, J. M., Loveland, K. L., Affara, N. A. and Burgoyne, P. S. (2005). Deletions on mouse Yq lead to upregulation of multiple X- and Y-linked transcripts in spermatids. *Hum Mol Genet* **14**, 2705-15.

Escalier, D., Allenet, B., Badrichani, A. and Garchon, H. J. (1999). High level expression of the Xlr nuclear protein in immature thymocytes and colocalization with the matrix-associated region-binding SATB1 protein. *J Immunol* **162**, 292-8.

Escalier, D., Eloy, L. and Garchon, H. J. (2002). Sex-specific gene expression during meiotic prophase I: Xlr (X linked, lymphocyte regulated), not its male homologue Xmr (Xlr related, meiosis regulated), is expressed in mouse oocytes. *Biol Reprod* **67**, 1646-52.

Escalier, D. and Garchon, H. (2005). XMR, a dual location protein in the XY pair and in its associated nucleolus in mouse spermatocytes. *Mol. Reprod. Dev.* **72**, 105-12.

Escalier, D. and Garchon, H. J. (2000). XMR is associated with the asynapsed segments of sex chromosomes in the XY body of mouse primary spermatocytes. *Chromosoma* **109**, 259-65.

Forejt, J. and Iványi, P. (1974). Genetic studies on male sterility of hybrids between laboratory and wild mice (Mus musculus L.). *Genet Res* **24**, 189-206.

Frazer, K. A., Eskin, E., Kang, H. M., Bogue, M. A., Hinds, D. A., Beilharz, E. J., Gupta, R. V., Montgomery, J., Morenzoni, M. M., Nilsen, G. B. et al. (2007). A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* **449**, 851-62.

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., Aburatani, H., Jones, K. W., Tyler-Smith, C., Hurles, M. E. et al. (2006). Copy number variation: new insights in genome diversity. *Genome Res* **16**, 949-61.

Garchon, H. J., Loh, E., Ho, W. Y., Amar, L., Avner, P. and Davis, M. M. (1989). The XLR sequence family: dispersion on the X and Y chromosomes of a large set of closely related sequences, most of which are pseudogenes. *Nucleic Acids Res* **17**, 9871-88.

Geigl, J. B., Obenauf, A. C., Schwarzbraun, T. and Speicher, M. R. (2008). Defining 'chromosomal instability'. *Trends Genet* **24**, 64-9.

Glover, T. W., Arlt, M. F., Casper, A. M. and Durkin, S. G. (2005). Mechanisms of common fragile site instability. *Hum Mol Genet* **14 Spec No. 2**, R197-205.

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R. J., Freedman, B. I., Quinones, M. P., Bamshad, M. J. et al. (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434-40.

Good, J., Dean, M. D. and Nachman, M. W. (2008). A complex genetic basis to X-linked hybrid male sterility between two species of house mice. *Genetics* **179**, 2213-28.

Gottlieb, B., Beitel, L. K. and Trifiro, M. A. (2001). Somatic mosaicism and variable expressivity. *Trends Genet* **17**, 79-82.

Graubert, T. A., Cahan, P., Edwin, D., Selzer, R. R., Richmond, T. A., Eis, P. S., Shannon, W. D., Li, X., McLeod, H. L., Cheverud, J. M. et al. (2007). A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genet* **3**, e3.

Guénet, J. L. and Bonhomme, F. (2003). Wild mice: an ever-increasing contribution to a popular mammalian model. *Trends Genet* **19**, 24-31.

Haber, J. E. (1999). DNA recombination: the replication connection. *Trends in Biochemical Sciences* **24**, 271-5.

Hanawalt, P. C. and Spivak, G. (2008). Transcription-coupled DNA repair: two decades of progress and surprises. *Nat Rev Mol Cell Biol* **9**, 958-70.

Harr, B. (2006). Genomic islands of differentiation between house mouse subspecies. *Genome Research* **16**, 730-7.

Harrison, R. G. (1993). Hybrid Zones and the Evolutionary Process. Oxford: Oxford University Press.

Hayashi, K., Yoshida, K. and Matsui, Y. (2005). A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. *Nature* **438**, 374-8.

Hoek, M. and Stillman, B. (2003). Chromatin assembly factor 1 is essential and couples chromatin assembly to DNA replication in vivo. *Proc Natl Acad Sci USA* **100**, 12183-8.

Hoffman, S. M. and Brown, W. M. (1995). The molecular mechanism underlying the "rare allele phenomenon" in a subspecific hybrid zone of the California field mouse, Peromyscus californicus. *J Mol Evol* **41**, 1165-9.

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W. and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nat Genet* **36**, 949-51.

Ihle, S., Ravaoarimanana, I., Thomas, M. and Tautz, D. (2006). An analysis of signatures of selective sweeps in natural populations of the house mouse. *Mol Biol Evol* **23**, 790-7.

Inoue, K. and Lupski, J. R. (2002). Molecular mechanisms for genomic disorders. *Annual review of genomics and human genetics* **3**, 199-242.

Jacquemont, M., Sanlaville, D., Redon, R., Raoul, O., Cormier-Daire, V., Lyonnet, S., Amiel, J., Le Merrer, M., Heron, D., De Blois, M. et al. (2006). Array-based comparative genomic hybridisation identifies high frequency of cryptic chromosomal rearrangements in patients with syndromic autism spectrum disorders. *Journal of Medical Genetics* **43**, 843-9.

Jaroudi, S. and SenGupta, S. (2007). DNA repair in mammalian embryos. *Mutat Res* **635**, 53-77.

Karanjawala, Z. E., Grawunder, U., Hsieh, C. L. and Lieber, M. R. (1999). The nonhomologous DNA end joining pathway is important for chromosome stability in primary fibroblasts. *Curr Biol* **9**, 1501-4.

Kaufman, P. D., Kobayashi, R., Kessler, N. and Stillman, B. (1995). The p150 and p60 subunits of chromatin assembly factor I: a molecular link between newly synthesized histones and DNA replication. *Cell* **81**, 1105-14.

Kaufmann, M. H. (1992). The Atlas of Mouse Development, Revised Edition. London: Elsevier Academic Press.

Kedar, P. S., Stefanick, D. F., Horton, J. K. and Wilson, S. H. (2008). Interaction between PARP-1 and ATR in mouse fibroblasts is blocked by PARP inhibition. *DNA Repair (Amst)* **7**, 1787-98.

Kidd, J., Cooper, G., Donahue, W., Hayden, H., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F. et al. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64.

Kidwell, M. G., Kidwell, J. F. and Sved, J. A. (1977). Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: A Syndrome of Aberrant Traits Including Mutation, Sterility and Male Recombination. *Genetics* **86**, 813-833.

Kim, P. M., Lam, H. Y., Urban, A. E., Korbel, J. O., Affourtit, J., Grubert, F., Chen, X., Weissman, S., Snyder, M. and Gerstein, M. B. (2008). Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome Research* **18**, 1865-74.

Kuzminov, A. (2001). Single-strand interruptions in replicating chromosomes cause double-strand breaks. *Proc Natl Acad Sci USA* **98**, 8241-6.

Li, J., Jiang, T., Mao, J. H., Balmain, A., Peterson, L., Harris, C., Rao, P. H., Havlak, P., Gibbs, R. and Cai, W. W. (2004). Genomic segmental polymorphisms in inbred mouse strains. *Nat Genet* **36**, 952-4.

Liang, Q., Conte, N., Skarnes, W. C. and Bradley, A. (2008). Extensive genomic copy number variation in embryonic stem cells. *Proc Natl Acad Sci USA*.

Macholán, M., Munclinger, P., Sugerková, M., Dufková, P., Bímová, B., Bozíková, E., Zima, J. and Piálek, J. (2007). Genetic analysis of autosomal and X-linked markers across a mouse hybrid zone. *Evolution* **61**, 746-71.

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends Ecol Evol (Amst)* **20**, 229-37.

Marshall, C. R., Noor, A., Vincent, J. B., Lionel, A. C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y. et al. (2008). Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* **82**, 477-88.

McCarroll, S. A., Hadnott, T. N., Perry, G. H., Sabeti, P. C., Zody, M. C., Barrett, J. C., Dallaire, S., Gabriel, S. B., Lee, C., Daly, M. J. et al. (2006). Common deletion polymorphisms in the human genome. *Nat Genet* **38**, 86-92.

McKinney, C., Merriman, M. E., Chapman, P. T., Gow, P. J., Harrison, A. A., Highton, J., Jones, P. B., McLean, L., O'donnell, J. L., Pokorny, V. et al. (2007). Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis. *Ann Rheum Dis*.

McVey, M. and Lee, S. E. (2008). MMEJ repair of double-stranded breaks (director's cut): deleted sequences and alternate endings. *Trends in Genetics* **24**, 525-582.

Mihola, O., Trachtulec, Z., Vlcek, C., Schimenti, J. C. and Forejt, J. (2009). A mouse speciation gene encodes a meiotic histone h3 methyltransferase. *Science* **323**, 373-5.

Moulia, C., Le Brun, N., Dallas, J., Orth, A. and Renaud, F. (1993). Experimental evidence of genetic determinism in high susceptibility to intestinal pinworm infection in mice: a hybrid zone model. *Parasitology* **106 (Pt 4)**, 387-93.

Moulia, C., Le Brun, N., Loubes, C., Marin, R. and Renaud, F. (1995). Hybrid vigour against parasites in interspecific crosses between two mice species. *Heredity* **74 (Pt 1)**, 48-52.

Mueller, J., Mahadevaiah, S. K., Park, P., Warburton, P., Page, D. C. and Turner, J. M. (2008). The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression. *Nat Genet* **40**, 794-9.

Myung, K., Datta, A. and Kolodner, R. D. (2001). Suppression of spontaneous chromosomal rearrangements by S phase checkpoint functions in Saccharomyces cerevisiae. *Cell* **104**, 397-408.

Myung, K. and Kolodner, R. D. (2002). Suppression of genome instability by redundant S-phase checkpoint pathways in Saccharomyces cerevisiae. *Proc Natl Acad Sci USA* **99**, 4500-7.

Naiki, T., Shimomura, T., Kondo, T., Matsumoto, K. and Sugimoto, K. (2000). Rfc5, in cooperation with rad24, controls DNA damage checkpoints throughout the cell cycle in Saccharomyces cerevisiae. *Mol Cell Biol* **20**, 5888-96.

Nair, S., Miller, B., Barends, M., Jaidee, A., Patel, J., Mayxay, M., Newton, P., Nosten, F., Ferdig, M. T. and Anderson, T. J. (2008). Adaptive copy number evolution in malaria parasites. *PLoS Genet* **4**, e1000243.

O'Driscoll, M. and Jeggo, P. A. (2006). The role of double-strand break repair - insights from human genetics. *Nature Reviews Genetics* **7**, 45-54.

Okano, S., Lan, L., Caldecott, K. W., Mori, T. and Yasui, A. (2003). Spatial and temporal cellular responses to single-strand breaks in human cells. *Mol Cell Biol* **23**, 3974-81.

Orii, K. E., Lee, Y., Kondo, N. and McKinnon, P. J. (2006). Selective utilization of nonhomologous end-joining and homologous recombination DNA repair pathways during nervous system development. *Proc Natl Acad Sci USA* **103**, 10017-22.

Paulsen, R. D. and Cimprich, K. A. (2007). The ATR pathway: fine-tuning the fork. *DNA Repair (Amst)* **6**, 953-66.

Payseur, B. A., Krenz, J. G. and Nachman, M. W. (2004). Differential patterns of introgression across the X chromosome in a hybrid zone between two species of house mice. *Evolution* **58**, 2064-78.

Perry, G. H., Ben-Dor, A., Tsalenko, A., Sampas, N., Rodriguez-Revenga, L., Tran, C. W., Scheffer, A., Steinfeld, I., Tsang, P., Yamada, N. A. et al. (2008a). The fine-scale and complex architecture of human copy-number variation. *Am J Hum Genet* **82**, 685-95.

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A., Mountain, J. L., Misra, R. et al. (2007). Diet and the evolution of human amylase gene copy number variation. *Nat Genet* **39**, 1256-60.

Perry, G. H., Tchinda, J., McGrath, S. D., Zhang, J., Picker, S. R., Cáceres, A. M., Iafrate, A. J., Tyler-Smith, C., Scherer, S. W., Eichler, E. E. et al. (2006). Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci USA* **103**, 8006-11.

Perry, G. H., Yang, F., Marques-Bonet, T., Murphy, C., Fitzgerald, T., Lee, A. S., Hyland, C., Stone, A. C., Hurles, M. E., Tyler-Smith, C. et al. (2008b). Copy number variation and evolution in humans and chimpanzees. *Genome Research* **18**, 1698-710.

Pinkel, D. and Albertson, D. G. (2005a). Array comparative genomic hybridization and its applications in cancer. *Nat Genet* **37 Suppl**, S11-7.

Pinkel, D. and Albertson, D. G. (2005b). Comparative genomic hybridization. *Annual review of genomics and human genetics* **6**, 331-54.

Piotrowski, A., Bruder, C. E., Andersson, R., de Ståhl, T. D., Menzel, U., Sandgren, J., Poplawski, A., von Tell, D., Crasto, C., Bogdan, A. et al. (2008). Somatic mosaicism for copy number variation in differentiated human tissues. *Hum Mutat*.

Raychaudhuri, S., Stuart, J. M. and Altman, R. B. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, 455-66.

Read, L. R., Raynard, S. J., Rukść, A. and Baker, M. D. (2004). Gene repeat expansion and contraction by spontaneous intrachromosomal homologous recombination in mammalian cells. *Nucl Acids Res* **32**, 1184-96.

Reale, A., Matteis, G. D., Galleazzi, G., Zampieri, M. and Caiafa, P. (2005). Modulation of DNMT1 activity by ADP-ribose polymers. *Oncogene* **24**, 13-9.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W. et al. (2006). Global variation in copy number in the human genome. *Nature* **444**, 444-54.

Reynard, L. N., Turner, J. M., Cocquet, J., Mahadevaiah, S. K., Touré, A., Höög, C. and Burgoyne, P. S. (2007). Expression analysis of the mouse multi-copy X-linked gene Xlr-related, meiosis-regulated (Xmr), reveals that Xmr encodes a spermatid-expressed cytoplasmic protein, SLX/XMR. *Biol Reprod* **77**, 329-35.

Rhee, I., Bachman, K. E., Park, B. H., Jair, K. W., Yen, R. W., Schuebel, K. E., Cui, H., Feinberg, A. P., Lengauer, C., Kinzler, K. W. et al. (2002). DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature* **416**, 552-6.

Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology, Published online: 23 November 2008; | doi:10.1038/nbt.1509* **26**, 303-4.

Sage, R. D., Heyneman, D., Lim, K. C. and Wilson, A. C. (1986). Wormy mice in a hybrid zone. *Nature* **324**, 60-3.

Salcedo, T., Geraldes, A. and Nachman, M. W. (2007). 10.1534/genetics.107.079988. In *Genetics*, vol. 177 (ed., pp. 2277.

Sarker, A. H., Tsutakawa, S. E., Kostek, S., Ng, C., Shin, D. S., Peris, M., Campeau, E., Tainer, J. A., Nogales, E. and Cooper, P. K. (2005). Recognition of RNA polymerase II and transcription bubbles by XPG, CSB, and TFIIH: insights for transcription-coupled repair and Cockayne Syndrome. *Mol Cell* **20**, 187-98.

Schilthuizen, M., Hoekstra, R. F. and Gittenberger, E. (2001). The 'rare allele phenomenon' in a ribosomal spacer. *Mol Ecol* **10**, 1341-5.

Schlötterer, C. and Tautz, D. (1994). Chromosomal homogeneity of Drosophila ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution. *Curr Biol* **4**, 777-83.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J. et al. (2007). Strong association of de novo copy number mutations with autism. *Science* **316**, 445-9.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M. et al. (2004). Large-scale copy number polymorphism in the human genome. *Science* **305**, 525-8.

Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., Pertz, L. M., Clark, R. A., Schwartz, S., Segraves, R. et al. (2005). Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* **77**, 78-88.

She, X., Cheng, Z., Zöllner, S., Church, D. M. and Eichler, E. E. (2008). Mouse segmental duplication and copy number variation. *Nat Genet* **40**, 909-14.

Sieber, O. M., Heinimann, K. and Tomlinson, I. P. (2003). Genomic instability--the engine of tumorigenesis? *Nature Reviews Cancer, Published online: 01 December 2007; | doi:10.1038/nrc2251* **3**, 701-8.

Smith, D. R. and Glenn, T. C. (1995). Allozyme polymorphisms in Spanish honeybees (Apis mellifera iberica). *J Hered* **86**, 12-6.

Snijders, A. M., Nowak, N. J., Huey, B., Fridlyand, J., Law, S., Conroy, J., Tokuyasu, T., Demir, K., Chiu, R., Mao, J. H. et al. (2005). Mapping segmental and sequence variations among laboratory mice using BAC array CGH. *Genome Res* **15**, 302-11.

Sonoda, E., Hochegger, H., Saberi, A., Taniguchi, Y. and Takeda, S. (2006). Differential usage of non-homologous end-joining and homologous recombination in double strand break repair. *DNA Repair (Amst)* **5**, 1021-9.

Spivak, G. and Hanawalt, P. C. (2006). Host cell reactivation of plasmids containing oxidative DNA lesions is defective in Cockayne syndrome but normal in UV-sensitive syndrome fibroblasts. *DNA Repair (Amst)* **5**, 13-22.

Stevenson, B. J., Iseli, C., Panji, S., Zahn-Zabal, M., Hide, W., Old, L. J., Simpson, A. J. and Jongeneel, C. V. (2007). Rapid evolution of cancer/testis genes on the X chromosome. *BMC Genomics* **8**, 129.

Storchová, R., Gregorová, S., Buckiová, D., Kyselová, V., Divina, P. and Forejt, J. (2004). Genetic analysis of X-linked hybrid sterility in the house mouse. *Mamm Genome* **15**, 515-24.

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., de Grassi, A., Lee, C. et al.

(2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848-53.

Tantin, D. (1998). RNA polymerase II elongation complexes containing the Cockayne syndrome group B protein interact with a molecular complex containing the transcription factor IIH components xeroderma pigmentosum B and p62. *J Biol Chem* **273**, 27794-9.

Tebbs, R. S., Flannery, M. L., Meneses, J. J., Hartmann, A., Tucker, J. D., Thompson, L. H., Cleaver, J. E. and Pedersen, R. A. (1999). Requirement for the Xrcc1 DNA base excision repair gene during early mouse development. *Dev Biol* **208**, 513-29.

Teeter, K. C., Payseur, B. A., Harris, L. W., Bakewell, M. A., Thibodeau, L. M., O'Brien, J. E., Krenz, J. G., Sans-Fuentes, M. A., Nachman, M. W. and Tucker, P. K. (2008). Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome Res* **18**, 67-76.

Thacker, J. and Zdzienicka, M. Z. (2004). The XRCC genes: expanding roles in DNA double-strand break repair. *DNA Repair* **3**, 1081-90.

The International Consortium, H. (2003). The International HapMap Project. *Nature* **426**, 789-96.

Thompson, L. H. and West, M. G. (2000). XRCC1 keeps DNA from getting stranded. *Mutat Res* **459**, 1-18.

Touré, A., Clemente, E. J., Ellis, P., Mahadevaiah, S. K., Ojarikre, O. A., Ball, P. A., Reynard, L., Loveland, K. L., Burgoyne, P. S. and Affara, N. A. (2005). Identification of novel Y chromosome encoded transcripts by testis transcriptome analysis of mice with deletions of the Y chromosome long arm. *Genome Biol* **6**, R102.

Tucker, P. K., Sage, R. D., Warner, J., Wilson, A. C. and Eicher, E. M. (1992). Abrupt Cline for Sex Chromosomes in a Hybrid Zone between Two Species of Mice. *Evolution* **46**, 1146-53.

Turner, D. J., Miretti, M., Rajan, D., Fiegler, H., Carter, N. P., Blayney, M., Beck, S. and Hurles, M. E. (2008). Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nat Genet* **40**, 90-5.

Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, V. A., Pertz, L. M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D. et al. (2005). Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727-32.

Vanlerberghe, F., Dod, B., Boursot, P., Bellis, M. and Bonhomme, F. (1986). Absence of Y-chromosome introgression across the hybrid zone between Mus musculus domesticus and Mus musculus musculus. *Genet Res* **48**, 191-7.

Vijg, J., Busuttil, R., Bahar, R. and Dollé, M. E. (2005). Aging and genome maintenance. *Annals of the New York Academy of Sciences* **1055**, 35-47.

Vinson, R. K. and Hales, B. F. (2002). DNA repair during organogenesis. *Mutat Res* **509**, 79-91.

Voolstra, C., Tautz, D., Farbrother, P., Eichinger, L. and Harr, B. (2007). Contrasting evolution of expression differences in the testis between species and subspecies of the house mouse. *Genome Research* **17**, 42-9.

Walsh, T., McClellan, J. M., McCarthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M., Nord, A. S., Kusenda, M., Malhotra, D., Bhandari, A. et al. (2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539-43.

Wang, P. J., McCarrey, J. R., Yang, F. and Page, D. C. (2001). An abundance of X-linked genes expressed in spermatogonia. *Nat Genet* **27**, 422-6.

Watkins-Chow, D. E. and Pavan, W. J. (2008). Genomic copy number and expression variation within the C57BL/6J inbred mouse strain. *Genome Research* **18**, 60-6.

Weiss, R. S., Enoch, T. and Leder, P. (2000). Inactivation of mouse Hus1 results in genomic instability and impaired responses to genotoxic stress. *Genes Dev* **14**, 1886-98.

West, S., Gromak, N. and Proudfoot, N. J. (2004). Human 5' --> 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. *Nature* **432**, 522-5.

West, S. C. (2003). Molecular views of recombination proteins and their control. *Nat Rev Mol Cell Biol* **4**, 435-45.

Wilson, D. and Thompson, L. H. (2007). Molecular mechanisms of sister-chromatid exchange. *Mutat Res* **616**, 11-23.

Wold, M. S. (1997). Replication protein A: a heterotrimeric, single-stranded DNA-binding protein required for eukaryotic DNA metabolism. *Annu Rev Biochem* **66**, 61-92.

Yang, H., Bell, T. A., Churchill, G. A. and Pardo-Manuel de Villena, F. (2007). On the subspecific origin of the laboratory mouse. *Nat Genet*.

Yonekawa, H., Moriwaki, K., Gotoh, O., Miyashita, N., Matsushima, Y., Shi, L. M., Cho, W. S., Zhen, X. L. and Tagashira, Y. (1988). Hybrid origin of Japanese mice "Mus musculus molossinus": evidence from restriction analysis of mitochondrial DNA. *Molecular Biology and Evolution* **5**, 63-78.

Zhu, M. and Weiss, R. S. (2007). Increased common fragile site expression, cell proliferation defects, and apoptosis following conditional inactivation of mouse Hus1 in primary cultured cells. *Molecular Biology of the Cell* **18**, 1044-55.

Zou, L. and Elledge, S. J. (2003). Sensing DNA damage through ATRIP recognition of RPA-ssDNA complexes. *Science* **300**, 1542-8.