2012

SPECIATION OF THE HOUSE MOUSE: GENE FLOW AND MOLECULAR EVOLUTION IN A HYBRID SYSTEM

Matthew R. Lindeman
Northern Michigan University

SPECIATION OF THE HOUSE MOUSE: GENE FLOW AND MOLECULAR EVOLUTION IN A HYBRID SYSTEM

By

Matthew R. Lindeman

THESIS

Submitted to
Northern Michigan University
In partial fulfillment of the requirements
For the degree of

MASTER OF SCIENCE

Office of Graduate Education and Research

2012

SIGNATURE APPROVAL FORM

This thesis by Matthew R. Lindeman is recommended for approval by the student's thesis committee in the Department of Biology and by the Assistant Provost of Graduate Education and Research.

Committee Chair: Katherine Teeter, Ph.D. Date

First Reader: John Rebers, Ph.D. Date

Second Reader: Kurt Galbreath, Ph.D. Date

Department Head: John Rebers, Ph.D. Date

Asst Provost of Graduate Education and Research: Brian D. Cherry, Ph.D. Date

ABSTRACT

SPECIATION OF THE HOUSE MOUSE: GENE FLOW AND MOLECULAR EVOLUTION IN A HYBRID SYSTEM

By

Matthew R. Lindeman

Mus domesticus and *Mus musculus* are two species of house mice which have evolutionarily diverged and recently come into contact again. These species are closely related enough to interbreed (hybridize) resulting in gene flow between the two species. This secondary contact has occurred in Europe, resulting in a hybrid zone stretching north to south across the continent. The *t*-haplotype is a gene complex in *Mus* species causing non-Mendelian inheritance, transmitting to the next generation at an unusually high rate. This unusual genetic element has been studied in natural populations but has received little attention in the *M. domesticus* – *M. musculus* hybrid system. A Polymerase Chain Reaction (PCR) based assay was used to investigate the distribution of the *t*-haplotype in the Saxony transect of the *M. domesticus* – *M. musculus* hybrid zone. Sequences from the Sanger Institute's Mouse Genomes project were used to investigate rates of gene evolution in the *t*-complex and in highly introgressing genomic regions. The *t*-haplotype was found to not cross the hybrid zone readily. Complete *t*-haplotypes were only found in the *M. domesticus* side of the hybrid zone, with one partial *t*-haplotype in the *M. musculus* side. The *t*-complex contains multiple rapidly evolving genes which likely contributed to the evolution of transmission ratio distortion and may contribute to

reproductive isolation. Highly introgressing genomic regions were found to be evolving more slowly, introgressing due to neutral forces or weak positive selection. This study illustrates the interplay between gene flow and molecular evolution in the *M. domesticus* – *M. musculus* hybrid zone.

ACKNOWLEDGEMENTS

I want to thank my advisor, Dr. Katherine Teeter, for her guidance and support throughout this project; my committee members Drs. John Rebers and Kurt Galbreath for their valuable advice and suggestions contributing to this work; Dr. Priscilla Tucker for providing SNP introgression data; Václav Janoušek for providing genomic estimates of gene evolution; and all of my friends at the Northern Michigan University Department of Biology.

PREFACE

This project was supported by an NMU Faculty Research Grant and the Excellence in Education Grant.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Biological evolution is a continuous process of change. Over many generations, mutations inevitably accumulate in the gene pool and cause a population's genetic structure to change over time. Physical and molecular changes to the organism occur as a result of accumulating genetic differences. In the allopatric model of speciation, a barrier to gene flow prevents new mutations from being exchanged by sister populations. The genetic structures of such related populations consequently diverge, becoming less similar over time. Populations which have evolved in isolation for sufficient time can develop incompatible new alleles leading to reproductive isolation. Hybrid Incompatibility (HI) can manifest as reduced fitness, sterile offspring, and even lethality of hybrid progeny.

In recent years, the field of population genetics has experienced profound advances in its ability to analyze the way in which genes move in a population. Particularly the appearance of high-throughput or "next generation" sequencing is making an impact on the way that research is done. The ability to determine DNA sequence data more rapidly, on a larger scale, and at reduced cost now grants greater power to identify and study genes of interest in hybrid systems and determine the level of divergence in individual gene sequences. This study used an established genotyping strategy as well as newly available genome sequences from the Sanger Institute to investigate gene flow and molecular evolution in a *Mus domesticus – Mus musculus* hybrid zone.

Chapter I: Literature Review

New alleles evolving in a species may be incompatible in the genetic background of closely related species. In these cases, a new allele can disrupt coadapted gene complexes, being functionally diverged from the homologous allele. The failure of a new allele to interact as its homologue does can cause reproductive incompatibility in hybrids. Well documented examples include cases of lethality and sterility in *Drosophila* hybrids (Barbash et al., 2003; Presgraves et al., 2003) and *Mus* (Vyskocilova et al., 2008).

Having diverged an estimated 2-3 million years ago, *Drosophila* species *D. melanogaster* and *D. simulans* provide an example of the disruption of coadapted gene complexes. Presgraves (2003) estimates there are approximately 191 genomic regions causing lethal X-autosomal epistatic interactions. The known genes involved in HI are limited, but a few cases between these species are well documented. The *D. melanogaster* and *D. simulans* forms of the genes *Hmr* (Barbash et al., 2003) and *Nup96*, described in Presgraves et al., 2003, have functionally diverged under recurrent positive selection. The wild type *D. melanogaster Hmr* gene causes male lethality, female infertility, and high temperature female lethality in *D. melanogaster – D. simulans* hybrids (Barbash et al., 2003). The *Drosophila* gene *Nup96* codes for a nucleoporin protein functioning in RNA export. Mutations in the *D. simulans* version of this gene prevent it from functioning correctly on a *D. melanogaster* background, resulting in the inviability of males with only the *D. melanogaster* X-chromosome. These two cases illustrate the impact that divergence in a single gene can have on the capability for gene flow between related species. Similarly, the Hst-1 locus is known to cause male sterility in *Mus domesticus – Mus musculus* crosses (Vyskocilova et al., 2008). Hybrid males

have underdeveloped epididymi and testes and do not produce spermatozoa. Hayashida and Kohno (2008) report a mitochondrial deletion which causes male sterility in *M. domesticus – M. musculus* hybrids.

The *Mus domesticus – Mus musculus* contact zone in Europe (Figure 1) is a hybrid system which has held the attention of researchers for decades. *M. domesticus* and *M. musculus* are two species of house mouse which are estimated to have diverged, along with their sister taxon *M. castaneus*, between 350,000 and 900,000 years ago (She et al., 1990; Boursot et al., 1996; Suzuki et al., 2004). *M. domesticus* and *M. musculus* are believed to have come into contact again within the past 3,000 years (Cucchi et al., 2005). The contact zone divides Europe, running north to south through Germany and Austria before turning eastward toward the Black Sea. *M. domesticus* populates western Europe, and *M. musculus* is prevalent on the east side of the hybrid zone.

Fitness of hybrids can vary from that of the parent species in different ways. Heterosis, or "hybrid vigor", can occur and fitness of hybrids is greater due to an increase in genetic diversity. The exact mechanism of heterosis is unclear, but the effect may be a matter of dominance, overdominance, or pseudo-overdominance (Birchler et al., 2010). Alternatively, if the parent taxa have diverged enough hybrids can be less fit due to the disruption of coadapted gene complexes (Clarke, 1993). Sage et al. (1986) found increased parasite loads in *M. domesticus – M. musculus* hybrids. Lowered ability of hybrids to resist parasitism may indicate a reduction in fitness.

Lowered fitness of hybrids would indicate that the *M. domesticus – M. musculus* hybrid zone is a tension zone. A tension zone is a hybrid system where the more-fit parent taxa on either side continuously contribute individuals to the hybrid zone as the

less fit hybrids are selected against (Boursot et al., 1993). Selection against hybrids results in a reproductive barrier to gene flow. This reduces the exchange of genes between diverging species and allows genetic mutations to accumulate, creating more differences between them. Continued divergence of parent species' genomes eventually results in complete reproductive isolation.

Varying size differences between bilaterally symmetrical characteristics (Fluctuating Asymmetry) can be used as a measure of developmental stability. High Fluctuating Asymmetry indicates lower developmental stability, while low Fluctuating Asymmetry indicates higher developmental stability and has been used as an analog of fitness in hybrid systems. Measures of Fluctuating Asymmetry in molars of *M. domesticus* – *M. musculus* hybrids, reported by Alibert et al., (1994; 1997) seemed to indicate a heterotic effect, even in sterile hybrids. This result indicates that individuals with a mixture of *M. domesticus* and *M. musculus* genes were more developmentally fit. Furthermore, genetic marker data from the Bavaria and Czech transects (Wang et al., 2011) indicate multi-generation hybrids. Hybrids in the contact zone are evidently fit enough to produce viable and fertile offspring themselves, allowing for gene flow across the hybrid zone. These studies suggest that incompatibility between *M. domesticus* and *M. musculus* occurs due to the disruption of a small number of coadapted genes, while heterozygosity at other loci is beneficial.

Gene flow between *M. domesticus* and *M. musculus* has been investigated previously by a number of studies. Tucker et al. (1992) found that the two X-linked markers and one Y-linked marker they studied introgressed less than the four autosomal markers included in the study. Macholan et al. (2007) estimated that there are approximately 4-7

times more X-linked than autosomal loci under selection, also reporting much lower introgression and fitness for X-linked loci. Payseur et al. (2004) identified a region on the X chromosome with a containing a hybrid male sterility locus. In the Bavaria (BV) transect, Teeter et al. (2008) found a handful of biological processes with genes significantly correlated with wide clines (G-protein-mediated signaling, pheromone response, chromatin packaging and remodeling, mesoderm development, cell-adhesion-mediated signaling, cell adhesion, and olfaction), as well processes with genes significantly associated with narrow clines and, presumably, reproductive isolation (MHCI-mediated immunity, steroid hormone metabolism, and cell structure).

The speciation event occurring between *M. domesticus* and *M. musculus* does not fit a strict allopatric model. Contact between the two species with the capability of hybridization and therefore gene-flow between the two populations precludes this system from being strictly allopatric. The capability of gene flow between the two species begs a line of questioning: Are genes of one species introgressing onto the genetic background of the other, and if so, what genes are introgressing? Are highly introgressing genes experiencing recurrent positive selection? What genes are resistant to introgression, and does this lack of introgression indicate hybrid incompatibility between coadapted gene complexes? The projects described in the following aim to help elucidate the genetic basis of isolation and gene flow between *M. domesticus* and *M. musculus* by searching for the signature of selection in the DNA sequences of the *t*-haplotype and genes associated with highly introgressing genetic markers.

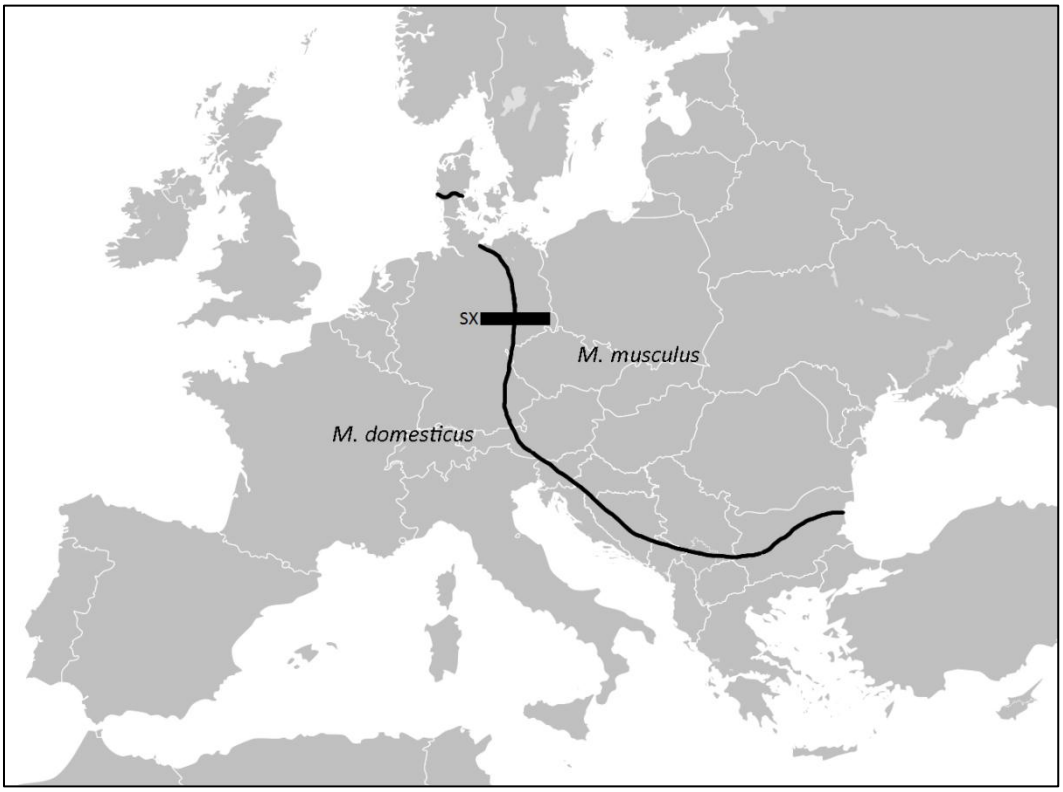

Figure 1: Hybrid Zone Map. The *M. domesticus* – *M. musculus* hybrid zone runs roughly north to south through Europe, represented by the solid black line. Samples to be used in this study were obtained from the Saxony transect, represented by the horizontal bar marked 'SX'.

Chapter II: *t*-Haplotype Genotyping

Background

The *t*-complex is a large gene-containing region making up the proximal third of chromosome 17. *t*-Haplotypes are abnormal forms of this region containing recessive lethal alleles and alleles causing sterility and Transmission Ratio Distortion (TRD) (Silver, 1985). TRD is a process in which an allele is transmitted to the next generation significantly more than the competing allele. The *t*-haplotype is such a genetic element, outcompeting the wild-type *t*-complex by a large margin.

t-Haplotypes were originally thought to be forms of the mutant *T* (brachyury) locus. The *T* locus produces short tails in *T/+* individuals and *T/T* is embryonic-lethal. Recessive *t*-alleles cause a tailless phenotype in *T/t* individuals, and were thought to cause sterility and TRD through pleiotropic effects. It was later determined through mapping studies that the *t*-haplotypes actually consist of a large complex of genes spanning the proximal third of chromosome 17 (Silver 1985).

Typical Mendelian inheritance would result in both copies of chromosome 17 to be equally transmitted to the next generation. The *t*-haplotype has been observed to transmit with as high as 99% frequency (Johnson et al., 1995). This departure from the normal 50% transmission occurs due to distorter loci which disrupt flagella function in sperm which do not contain the rescue locus, *Tcr* (Dod et al., 2003).

Investigations by Bauer et al. (2005; 2007;2012) have identified three of the distorter loci, *Fgd2, Tagap, and Nme3*. *Fgd2* and *Tagap1* regulate the activity of small Rho G proteins, and *Nme3* is a nucleotide diphosphate kinase gene which is also thought to affect the activity of these proteins. In male *t*-heterozygous mice, misregulation of Rho

small G proteins occurs in developing *t* and wild-type sperm, which are connected by syncitium. This misregulation causes the flagella of wild-type sperm to develop abnormally, impairing motility. Sperm carrying the *t*-haplotype also have the rescue locus, *Smok(Tcr)*, which prevents abnormal flagella development and allows *t*-sperm to be successful. A series of four inversions in the complex (Figure 2) reduce recombination with wild-type chromosomes during meiosis. Reduced recombination keeps the deleterious genes, distorter loci, and the rescue locus in the complex together (Hammer et al., 1989; Miller Baker, 2008).

Mus individuals heterozygous for the *t*-haplotype act as carriers for the deleterious alleles maintained by TRD and are estimated to make up 10-20% of the population (Miller Baker 2008). Homozygosity for *t* results in male sterility or death (Hammer et al., 1989). The markers Tcp-1 and Hba are located in inversions 2 and 4, respectively (Dod et al., 2003). Both inversions also contain insertions, making them detectable by a PCR-based screening method. Amplification of these regions using specially designed PCR primers produces a normal-length DNA product for wild-type mice. A longer DNA product is also produced for *t*-heterozygous mice. The size difference in PCR products can be resolved by standard agarose gel electrophoresis, allowing for hundreds of mice to be screened for the *t*-haplotype with relative ease.

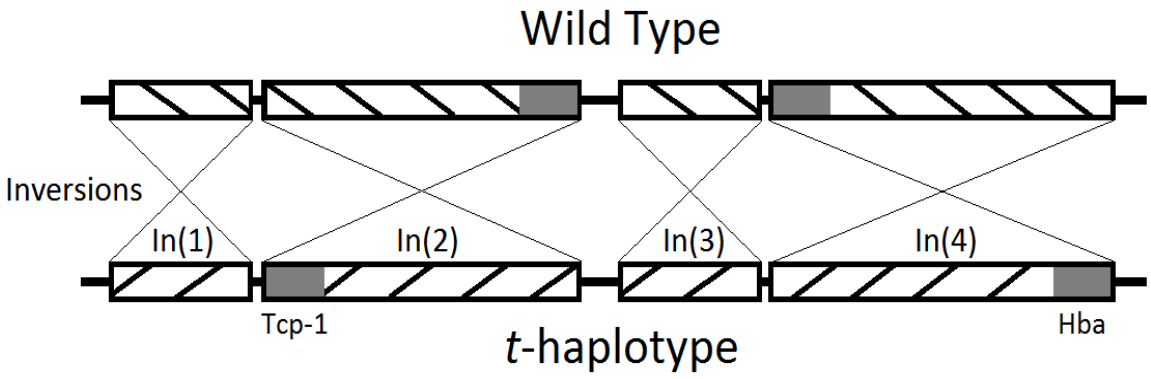

Figure 2: *t*-Complex. The *t*-complex consists of four inversions in chromosome 17. Mouse DNAs were genotyped for the *t*-haplotype using a PCR-based method targeting insertions in the Tcp-1 and Hba loci. Figure adapted from Dod et al., 2003.

Methods

Mus DNAs were obtained previously in 2001–2003 (Teeter et al., 2009). A total of 322 commensal *Mus* individuals were collected from 34 localities in the Saxony transect (SX), which stretches through the German states of Thuringia, Saxony-Anhalt, and Saxony. DNA was extracted from spleen or kidney tissue by either standard proteinase K / phenol-chloroform protocol, or the commercially available MasterPure™ DNA Purification Kit, manufactured by Epicentre Biotechnologies (Madison, WI).

The *Mus* DNA samples were genotyped for the presence of *t*/+ heterozygotes using a PCR-based screening method. Heterozygotes were identified using multiple sets of primers targeting the Tcp-1 and Hba markers. Diagnostic primer sequences for Tcp-1 and Hba markers are listed in Table 1. Reactions were conducted at a volume of 15 µL using GoTaq® 5X PCR buffer. GoTaq® 5X PCR buffer has a proprietary composition with a pH of 8.5 and $MgCl_2$ concentration of 7.5 mM. Hba reactions were run at a working $MgCl_2$ concentration of 1.5 mM. An additional 0.9 µL 25 mM $MgCl_2$ per reaction was added to Tcp-1, Tcp-1-Jad, and Tcp-1-Cl PCR mixes for a final concentration of 3.0 mM. PCR was run with a 3 minute 95°C initial denaturation step, followed by 30 cycles of 95°C denaturation, 53°C annealing, and 72°C elongation steps for 30 seconds each. Final elongation was 72°C for 7 minutes.

The Tcp-1 PCR products were run on 0.7% agarose gel at 100V for 90 minutes to resolve the 1.4 kb wild type fragment and the 1.6 kb *t*-fragment. These primers did not amplify as reliably as desired, possibly due to sub-optimal primer design and the large size of the PCR products. It was necessary to use alternate primer sets targeting the Tcp-1 region in order to complete genotyping. The Tcp-1-Jad and Tcp-1-Cl primer sets target

regions within the region defined by the Tcp-1 primer set. Wild type mice produce one 475 bp product for with Tcp-1-Jad primers, and *t*-heterozygotes produce one 475 bp product and an additional 600 bp product. Wild type mice produce one 517 bp product for Tcp-1-Cl primers, and *t*-heterozygous mice produce one 517 bp product and an additional 692 bp product. Tcp-1-Jad and Tcp-1-Cl PCR products were run on 0.7% agarose gel at 100V for 60 minutes. The Hba PCR products were run on a 3.0% Metaphor® agarose gel at 100V for 120 minutes to resolve the 198 bp wild type and 214 bp *t*-fragments.

Representative wild-type and *t*-specific bands were cut out of agarose gels with a clean razor blade for DNA extraction and sequencing. DNA was extracted from gel fragments with a QIAgen gel extraction kit and sequencing reactions were performed using a BigDye sequencing kit. Sequencing reactions were centrifuged through Sephadex columns for cleanup before being run on the ABI 3100 Avant at Northern Michigan University. The newly generated sequences were compared to reference sequences for Tcp-1 and Hba to confirm the identity of bands used for genotyping. Once confirmed, the PCR genotyping method was used to determine the frequency and distribution of the *t*-haplotype in the SX transect.

Table 1: PCR Primers for Genotyping. Primer sets used to genotype mice for the *t*-haplotype. (A) Primer sets targeting the Tcp-1 marker (Morita et al., 1993; Miller Baker, 2008). (B) Primer sets targeting the Hba marker (Schimenti & Hammer, 1990).

A

Primer:	Tcp-1 Forward	AGG AAA GCT TGC CCA AGA GAA TAG TTA ATG C
	Tcp-1 Reverse	AGG CGA ATT CCA TAT CAT CAA TGC CAC CAG
	Tcp-1-Jad Forward	GAC AAT CAT AGC CTT GTC TCA G
	Tcp-1-Jad Reverse	GCA GTG TTA TCT TTC ACT GG
	Tcp-1-Cl Forward	CTA TGT GGG GCT TGA TTT TCT GTC
	Tcp-1-Cl Reverse	TGC AAC ATG CTT CAG GTC TCG

B

Primer:	Hba Forward	GAG TGA CCT GCA TGC CCA CAA GCT GTG
	Hba Reverse	GAG CTG TGG AGA CAG GAA GGG TCA GTG

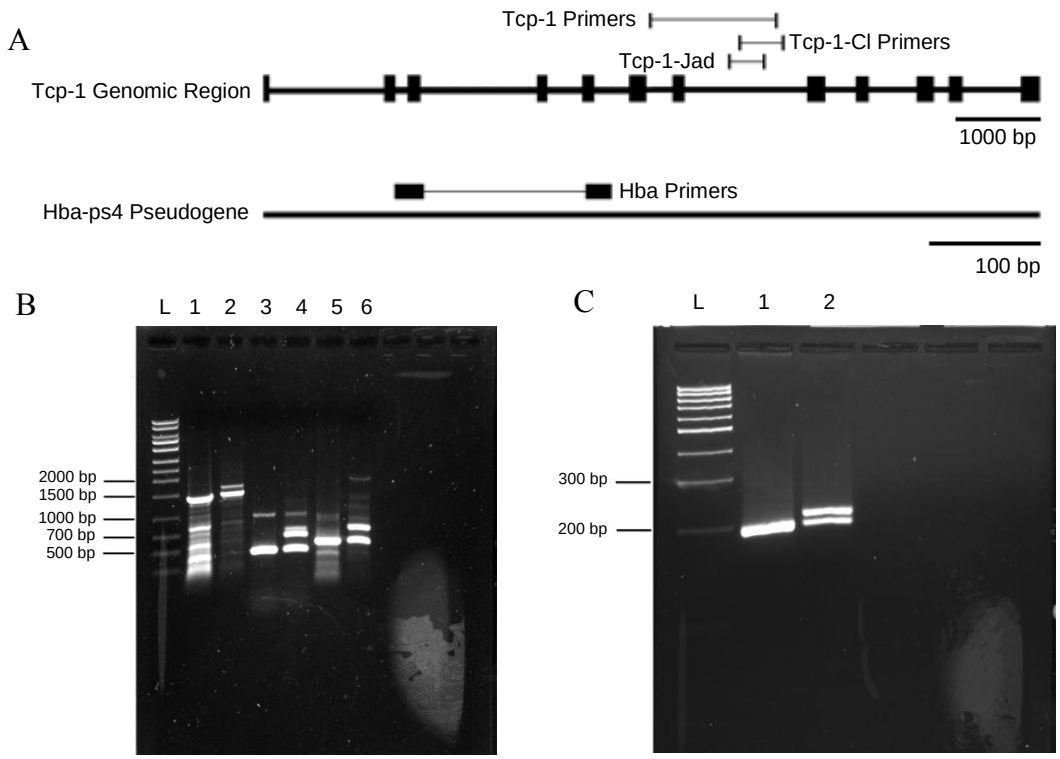

Figure 3: PCR Genotyping. (A) PCR primers are shown with respect to Tcp-1 and Hba genomic regions. Thickened segments represent primer sequence and thin segments represent amplified sequence. Schematic representations of the Tcp-1 genomic region and Hba-ps4 are shown. Thickened line segments represent coding regions, and thin segments represent non-coding regions. (B) Amplification of the Tcp region using Tcp-1, Tcp-Jad, and Tcp-cl sets of primers. Tcp-1 primers result in one 1.4 kb product (lane 1) in wild-type mice, and an additional 1.6 kb product (2) in *t*-heterozygous mice. Tcp-Jad primers result in one 475 bp product (3) for wild type with an additional 600 bp product (4) for *t*-heterozygotes, and Tcp-cl primers result in one 517 bp product (5) for wild type and an additional 692 bp product (6) for *t*-heterozygotes. (C) The Hba region produces one 198 base-pair product (lane 1) for wild-type mice, and an additional 214 base-pair product (2) for *t*-heterozygous mice.

Results

The Tcp-1, Tcp-1-Jad, Tcp-1-Cl, and Hba primer sets were used for PCR genotyping of all *Mus* DNA samples. Samples which were noted as species other than *Mus domesticus* or *Mus musculus*, or were captured outside of the hybrid zone were excluded from the data set. Of the 303 samples which were genotyped, 70 were not successfully genotyped with the Tcp-1 primer set after multiple attempts. Agreement between Tcp-1, Tcp-1-Jad, and Tcp-1-Cl made it possible to use the latter two in place of the Tcp-1 primer set for these samples.

Genotyping data was cross-referenced with trapping records to determine the frequency of the *t*-haplotype in each of the 34 localities. Values are listed in Table 2. The overall frequency of *t*-heterozygous mice in the SX transect was calculated to be 0.175. Frequencies of individual localities range from 0 to 1, but it should be noted that only one mouse was trapped in both localities with frequencies of 1. Numbers of wild-type and *t*-heterozygous mice were plotted by locality in Figure 4. The frequency of the *t*-haplotype was calculated for each locality and plotted against distance across the SX transect in Figure 5. Genotyping of Tcp-1 and Hba markers agreed in most cases. Four mice were genotyped as having the Tcp-1 portion of the *t*-haplotype but not the Hba region. This indicates a low frequency of partial *t*-haplotypes in the hybrid zone. Overall, the *t*-haplotype is more prevalent in the west (*M. domesticus*) side of the hybrid zone, with only one *t*-heterozygous mouse detected in the east (*M. musculus*) side of the hybrid zone.

Table 2: Locality Data. West-to-east distances, number of mice captured, and *t* frequency is shown for each locality in the SX transect of the hybrid zone. Localities 1-24 are in the *M. domesticus* side of the hybrid zone. Localities 25-34 are in the *M. musculus* side of the hybrid zone.

Locality	**Name**	**Distance (km)**	**Mice**	***t* Freq**
1	Remderoda bei Jena	0	35	0
2	Benkendorf bei Salzmuende	21	1	0
3	Doellnitz-Halle, Gut Dollnitz	34	5	0.40
4	Borau bei Weissenfels	34	3	0
5	Burgliebenau, S. of Halle	36	1	0
6	Muschwitz bei Weissenfels	42	1	0
7	Zeitz	43	1	0
8	Grosspoerthen bei Zeitz	45	4	0
9	Nissma bei Kayna, E. of Altenbur	53	2	0
10	Borna	70	6	0.83
11	N. of Floessberg	74	4	0
12	Trebishain bei Floessberg	76	4	0.50
13	Thallwitz, N. of Wurzen	81	10	0.40
14	Nischwitz	82	1	0
15	Dehnitz bei Wurzen	84	37	0.35
16	Dehnitz/NSI	84	5	0.40
17	Lueptitz	86	43	0.09
18	Gniebitz bei Trossin	87	51	0.33
19	Trebelshain, E. of Wurzen	90	14	0
20	Zschirla	92	1	1
21	Mehderitzsch/Losswig	104	1	0
22	Kreischau	104	2	0.50
23	Hohenlauft, by Rosswein, by Doebel	113	8	0.13
24	Troischau, by Rosswein, by Doebeln	114	16	0
25	Wilsdruff	139	24	0
	Mus domesticus – *Mus musculus* Boundary			
26	Lohmen, Kastanienalle 57	173	1	0
27	Pulsnitz	173	1	0
28	Kamenz, Museum der Westlausitz	178	1	1
30	Kamenz OT Wiesa	179	6	0
31	Deutschbaselitz	182	1	0
32	Piskowitz	185	3	0
33	Skerbersdorf bei Goerlitz	227	3	0
34	Friedersdorf bei Goerlitz	231	4	0
35	Goerlitz, Tierpark	239	3	0

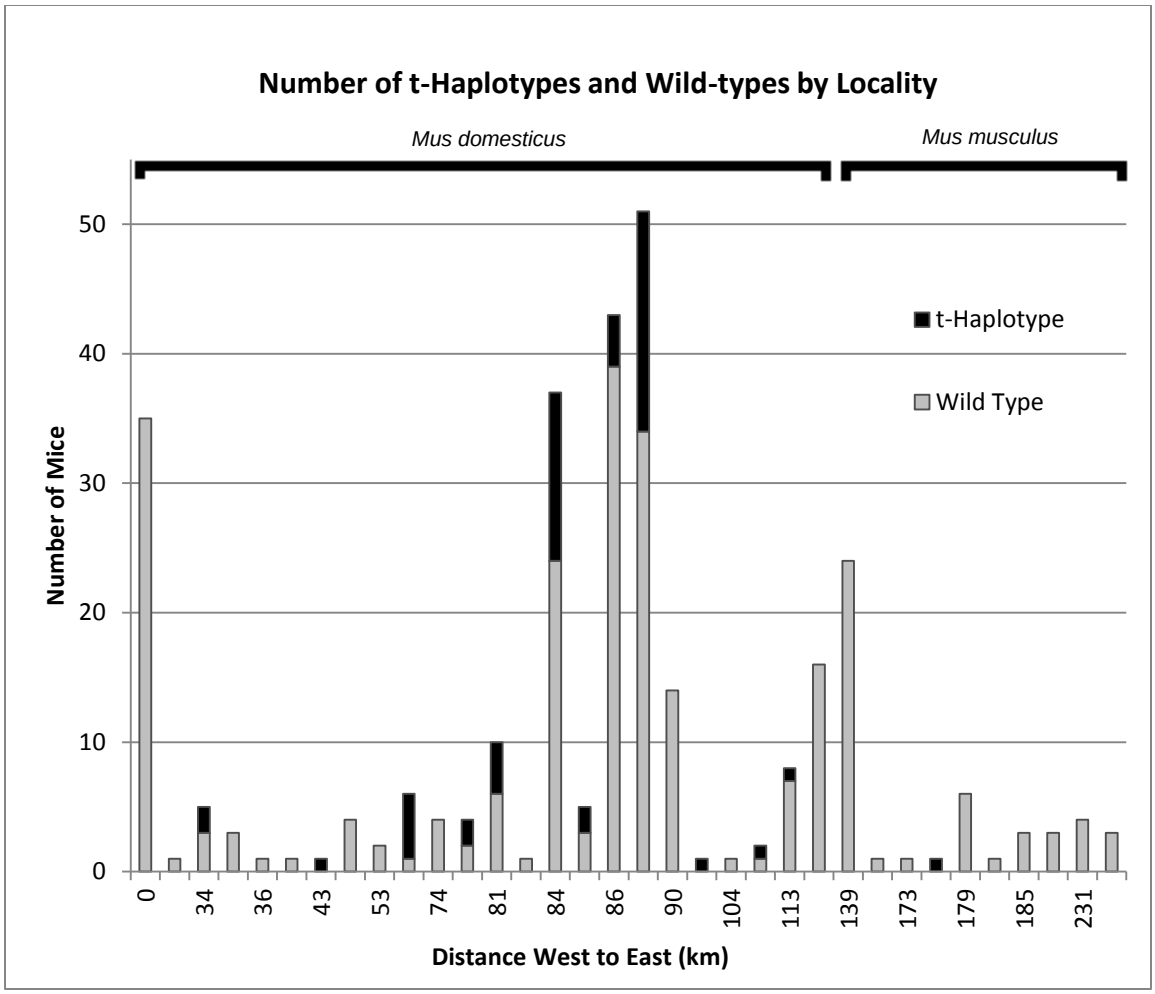

Figure 4: Genotypes by Locality. Mice were trapped in 34 localities across the Saxony transect. Numbers of wild-type mice and *t*-heterozygous mice are shown by the distance of each locality from the west end of the hybrid zone. Mice at localities from 0-114 km have > 80% *M. domesticus* alleles. Mice at localities from 139-239 km have > 80% *M. musculus* alleles.

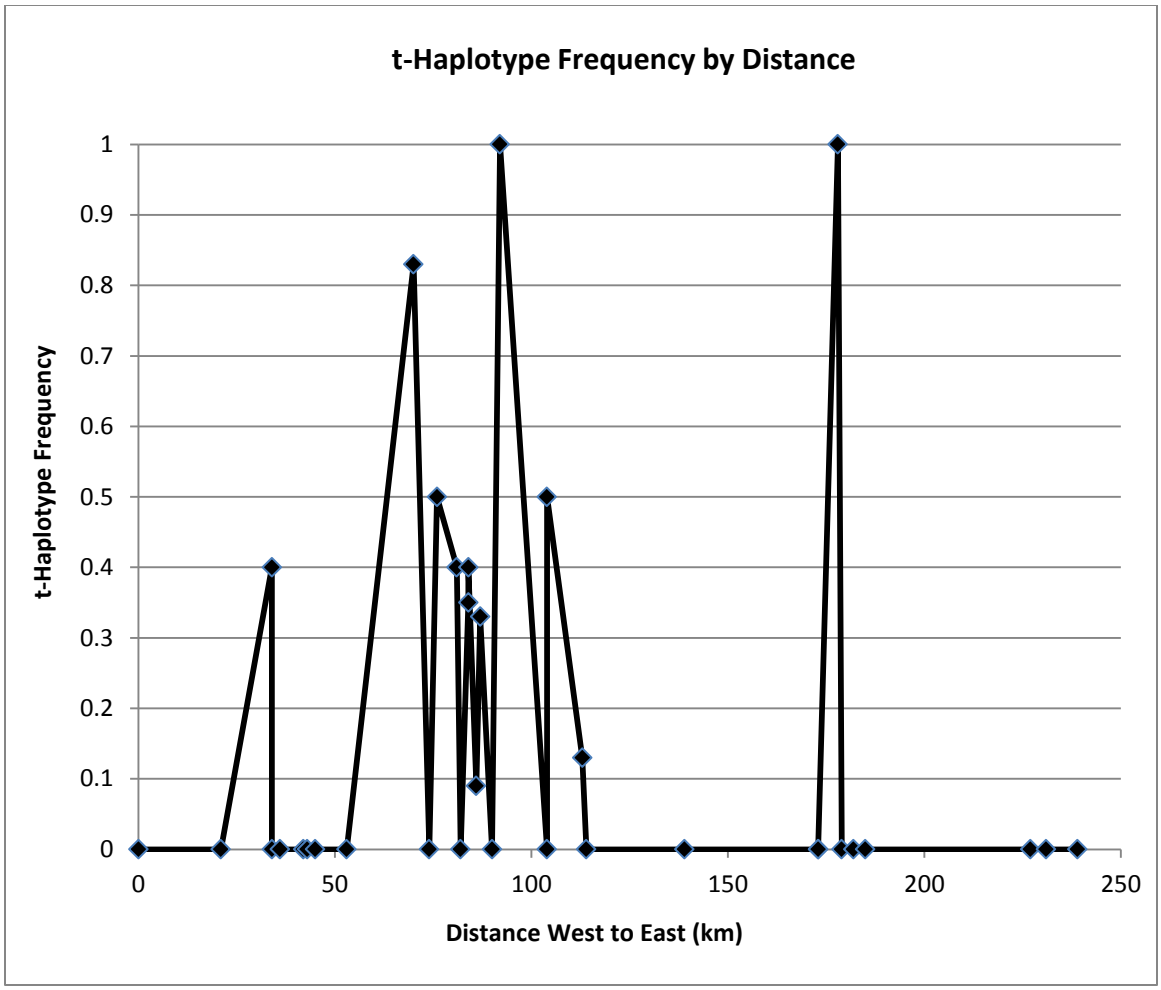

Figure 5: *t*-Haplotype Frequency Distribution. Frequency of *t*-heterozygous mice was calculated for each locality. Frequencies are plotted against the distance of each locality from locality 1 at the western end of the hybrid zone.

Discussion

Genotyping of 303 mice from localities across the SX transect detected the *t*-haplotype in 17.8% of mice. This observation is consistent with previous findings reporting the *t*-haplotype at 10-25% in natural populations (Ardlie & Silver 1998; Huang et al., 2001). Recessive alleles causing sterility and lethality in *t*-homozygous mice are thought to keep the *t*-haplotype at low frequencies despite strong transmission ratio distortion, and the results of this study in the Saxony transect are consistent with this hypothesis.

Genotyping revealed a sharp contrast between the west and east sides of the hybrid zone. This is unusual, as the *t*-haplotype has been described previously in *M. domesticus, M. musculus* and *M. castaneus* and is thought to move freely between these species (Huang et al., 2001). The western half of the transect (0-114 km; localities 1-24) had the *t*-haplotype present in 11 localities. In the eastern half of the transect, only locality 28, Kamenz, Museum der Westlausitz, had a *t*-heterozygous mouse. This was the only mouse from this locality, resulting in a *t*-haplotype frequency of 1. This is most likely not representative of the population. It is interesting, however, that this mouse is the only *t*-heterozygote collected from the eastern side of the transect.

The *t*-heterozygote found in Kamenz is contrary to the rest of the genotyping data, which suggests a trend of the *t*-haplotype being exclusively found in the *M. domesticus* side of the hybrid zone. Genome-wide SNP data provided by K. C. Teeter reveals that the Kamenz mouse is primarily *M.musculus*, being homozygous for 77% of *M.musculus* versions of SNPs, heterozygous for 20%, and homozygous for 3% of *M. domesticus* versions of SNPs. Chromosome 17 has a slightly higher proportion of *M. domesticus* alleles, being heterozygous for 23% of SNPs and homozygous for the *M. domesticus*

version of 6% of SNPs. The Kamenz mouse is heterozygous for SNP 17-12921658, the SNP closest to the Tcp-1 marker (chr17:13109331-13117933), and homozygous for the *M. musculus* version of SNP 17-26695767, which is closest to the Hba marker (chr17:26423308-26424006). The Tcp-1-Cl and Tcp-1-Jad primer sets both detected the *t*-haplotype in this mouse, however genotyping with the Hba primer set did not detect the *t*-haplotype. Although recombination in the *t*-complex is typically prevented by inversions in this region, recombination does occur between the *t*-haplotype and wild type chromosomes about 1 in 1000 times (Howard et al., 1990). It appears that a *M. domesticus* chromosomal region containing a partial *t*-haplotype has introgressed into the *M. musculus* genetic background, which was detected in the Kamenz mouse.

The overall frequency of the *t*-haplotype was within the expected range based on previous investigations by Ardlie & Silver (1998), Huang et al. (2001) and Miller Baker (2008). The *t*-haplotype has also been described previously in *M. domesticus*, *M. musculus*, and *M. castaneus* populations and was believed to cross between populations (Huang et al., 2001). This was not observed in this study, as the *t*-haplotype was predominantly found on the *M. domesticus* side of the Saxony transect. The *t*-heterozygous mouse detected in the eastern side of the SX transect appears to have a partial *t*-haplotype, meaning that complete *t*-haplotypes were only found on the *M. domesticus* side of the zone. This suggests that one or more alleles in *M. domesticus t*-haplotypes studied may be incompatible with some part of the *M. musculus* genetic landscape. Determining which components of the *t*-haplotype interact negatively with what parts of the *M. musculus* genome will be a question for future investigations.

Chapter III: Tests of Selection in *t*-Complex Genes

Background

An allele which is continuously favored by natural selection as new mutations occur experiences recurrent positive selection**.** A way of detecting the action of recurrent positive selection on an allele is to compare the rate of non-synonymous mutations to the rate of synonymous mutations. A synonymous or "silent" mutation occurs when the mutation does not change the amino acid sequence. For example, the mRNA codons ACU, ACC, ACA, and ACG all translate to the amino acid, threonine. A mutation of the third base to any other will still result in a threonine codon and is therefore a synonymous mutation. A non-synonymous mutation, such as ACU to AUU, will change the coded amino acid from threonine to isoleucine and subsequently change the resulting protein. The impact of a non-synonymous mutation can vary, ranging from the severe case of a stop codon (effectively deleting the rest of the protein), to the unlikely yet possible event of a beneficial change. The number of non-synonymous mutations (Ka) increases relative to the number of synonymous mutations (Ks) when an allele experiences recurrent positive selection. Therefore, by calculating the Ka/Ks ratio in fixed sequence differences between gene orthologs, the presence and extent of recurrent positive selection can be inferred. A Ka/K_S ratio greater than 1.0 provides definitive evidence of positive selection, though a Ka/Ks ratio which is elevated above the genomic norm but less than 1.0 may still be suggestive (Presgraves et al., 2003). Furthermore, a sliding-window approach to calculating Ka/Ks ratios can narrow down sequence evolution to specific regions of a gene and reveal high Ka/Ks peaks in a gene for which the overall

Ka/Ks ratio is not significant. The sliding window method increases the power of the test to detect selection as well as determine what parts of a gene selection may be acting on.

Six genes from the *t*-complex were analyzed for evidence of selection (Table 3). *T* codes for the brachyury protein. Investigations of deformities caused by mutant forms of this gene led to the discovery of the *t*-haplotype (Howard et al., 1990). *Tagap1*, *Fgd2*, and *Nme3* are genes which contribute additively to the transmission ratio distortion caused by the *t*-haplotype (Bauer et al., 2005; Bauer et al., 2007; Bauer et al., 2012). *Smok4a* is the wild type homologue to the *t*-complex responder, *Smok(Tcr)* (Herrmann et al., 1999). *Tcp-1* encodes the *t*-complex protein 1 subunit alpha, which was targeted to detect the *t*-haplotype in the SX transect.

Table 3: *t*-Complex Genes. Genes associated with the *t*-complex were investigated for molecular signatures of positive selection. *Fgd2*, *Tagap1*, and *Nme3* have been implicated in transmission ratio distortion. *Smok(Tcr)* is the responder locus contributing to TRD. The *Tcp-1* locus is used for genotyping in detecting the *t*-complex. *T* produces the brachyury protein, mutant forms of which led to the discovery of the *t*-complex.

Gene Symbol	**Gene Name**	**Location**	**Conserved Domain Database Matches**
Fgd2	FYVE, RhoGEF and PH domain-containing protein 2	chr17:29,497,859-29,516,480	Rho Guanine Exchange Factor (GEF), FGD Pleckstrin Homology domain, FYVE Zinc-binding domain, Pleckstrin homology domain
Smok(Tcr)	Sperm motility kinase 4A (*Smok4a*)	chr17:13,714,322-13,721,300	Catalytic domain of the Protein Serine/Threonine Kinase
Tagap1	T-cell activation GTPase-activating protein 1	chr17:7,159,314-7,165,505	GTPase-activator protein [GAP] for Rho-like small GTPases
Tcp-1	T-complex protein 1 subunit alpha	chr17:13,109,331-13,117,933	*TCP-1* (CTT or eukaryotic type II) chaperonin family, alpha subunit
T	Brachyury protein	chr17:8,627,288-8,635,361	T-box DNA binding domain
Nme3	Nucleoside diphosphate kinase 3	chr17:25,033,445-25,034,474	Nucleoside diphosphate kinase Group I (NDPk_I)-like

Methods

Whole-genome sequencing data of 17 strains of mice are available online from the Sanger Institute (http://www.sanger.ac.uk/resources/mouse/genomes/). This source was used to make sequence comparisons of Dom (*M. domesticus*) and PWK (*M. musculus*) strains, as well as CAST (*M. castaneus*) and SPRET (*M. spretus*). Mouse strains are abbreviated here as "Dom", "Musc", "Cast", and "Spret" for clarity.

De novo assemblies of chromosomal sequences were downloaded from the Sanger website and reduced to the regions of interest using the EmEditor Pro program for editing large text files (Emurasoft, Inc.). Borders were checked against the July 2007 NCBI37/mm9 mouse genome assembly using the UCSC BLAT tool. Sequences were aligned with ClustalW using MEGA 5.05 and proofread by eye. Coding sequences were built in MEGA 5.05 from *de novo* assemblies based on intron-exon boundaries and coding start and stop sites described in the Refseq profile of each gene. Coding sequences which were missing from *de novo* assemblies were reconstructed manually from the original read alignments of the Mouse Genomes project. Read alignments were navigated using Tablet 1.12.

Aligned coding sequences were loaded into KaKs_Calculator 2.0 which calculated Ka and K_S values for each pair of gene sequences using the Nei & Gojobori (1986) method. Significance of pairwise comparisons was tested by KaKs_Calculator using Fisher's exact test. Fisher's exact test consisted of a 2x2 contingency table comparing non-synonymous substitutions and non-synonymous sites to synonymous substitutions and synonymous sites. Low p-values indicate a significant difference between Ka and Ks. This can occur either due to a very low or very high Ka/Ks ratio. A very low Ka/Ks

ratio would indicate purifying selection. Elevated Ka/Ks ratios are typically less than 1.0. These cases are not considered significant by Fisher's exact test due to Ka and Ks being similar values, but are indicative of positive selection.

A sliding window calculation of Ka and K_S values for each gene was performed in DnaSP using a 200bp window and 25bp step size. The Ks value of one synonymous substitution was added to the Ks value of each window to prevent divide-by-zero errors in Ka/Ks ratios. Ka/K_S ratios were graphed using Microsoft Excel 2010.

Results

Whole-Gene Ka/Ks Ratios

Genome-wide calculations of synonymous and non-synonymous substitutions for Musc and Dom genomes were provided by V. Janoušek. Median Ka/Ks ratios were determined to be 0.092784 across the genome, and 0.120773 for chromosome 17. Whole-gene Ka/Ks ratios were calculated for putative distorter genes *Fgd2*, *Tagap1*, and *Nme3*, rescue locus *Smok(Tcr)*, and other *t*-related genes *T* and *Tcp-1* (Table 4).

Comparisons of distorter candidate *Fgd2* sequences resulted in significantly low Ka/Ks ratios, as determined by Fisher's exact test. In addition, the Dom-Musc comparison yielded a Ka/Ks of 0.07071, which is lower than the median Ka/Ks across the genome and for chromosome 17.

Ka/Ks calculations for pairings of rescue locus *Smok(Tcr)* sequences resulted in elevated ratios. While Ka/Ks ratios were elevated, these were not found to be significant

by Fisher's exact test. The Dom-Musc comparison had a Ka/Ks value of 0.316976, which is elevated with respect to the genome-wide and chromosome 17 median Ka/Ks.

Pairwise comparisons of the *T* locus, encoding the brachyury protein, produced low Ka/Ks ratios. All Ka/Ks ratios were found to be significantly lower than 1 by Fisher's exact test, and only comparisons including Spret sequence showed non-synonymous substitutions.

Comparisons of putative distorter *Tagap1* produced Ka/Ks ratios which were all above genome-wide and chromosome 17 median Ka/Ks. Interestingly, the Dom-Musc comparison yielded the highest Ka/Ks ratio (0.831637). Musc and Dom are two of the most closely related strains studied, with the exception of Musc and Cast. While comparisons of the more distantly related Spret strain with others were typically highest, this was not the case with *Tagap1* comparisons. Ka/Ks ratios in these comparisons are elevated relative to genomic and chromosome 17 median Ka/Ks.

Comparisons of *Tcp-1* gene sequences produced low Ka/Ks ratios. Comparisons of Spret *Tcp-1* with other sequences produced low but non-significant P-values, while the remaining comparisons lacked non-synonymous substitutions. Fisher's exact test found these comparisons to be significant, indicating purifying selection.

Comparisons of putative distorter *Nme3* sequences resulted in low Ka/Ks ratios. Only comparisons of Musc to other sources resulted in non-zero Ka, which is caused by a single non-synonymous substitution in the *M. musculus* sequence. The Cast-Musc comparison showed non-synonymous substitutions and no synonymous substitutions. Dom-Musc comparison yielded a Ka/Ks value of 0.328399**,** which is elevated compared to genome-wide and chromosome 17 median Ka/Ks values.

Table 4: Ka/Ks Ratios for *t*-Complex Genes. Whole-gene Ka/Ks ratios were calculated for genes of interest associated with the *t*-complex. Calculations were made using the Nei & Gojobori (1986) method in KaKs_Calculator 2.0. P-Values result from Fisher's exact test comparing synonymous substitutions and sites with non-synonymous substitutions and sites. Bonferroni correction for multiple tests reduces the critical value to 0.0083. Starred P-Values indicate significance.

Fgd2	**Ka**	**Ks**	**Ka/Ks**	**P-Value**
Cast-Spret	0.001345	0.021146	0.063602	2.77E-05*
Cast-Musc	0.001345	0.016883	0.079643	0.000328*
Cast-Dom	0.001345	0.019013	0.070727	9.63E-05*
Spret-Musc	0.001345	0.02115	0.063586	2.77E-05*
Spret-Dom	0.002692	0.02329	0.115606	8.39E-05*
Dom-Musc	0.001345	0.019017	0.07071	9.62E-05*
***T* (brachyury protein)**	**Ka**	**Ks**	**Ka/Ks**	**P-Value**
Cast-Spret	0.002162	0.0315	0.068647	8.09E-05*
Cast-Musc	0	0.020852	0	0*
Cast-Dom	0	0.024385	0	0*
Spret-Musc	0.002162	0.024381	0.088694	0.000972*
Spret-Dom	0.002162	0.02793	0.077425	1.81E-05*
Dom-Musc	0	0.010352	0	0*
Smok4a	**Ka**	**Ks**	**Ka/Ks**	**P-Value**
Cast-Spret	0.032293	0.039142	0.825034	0.476657
Cast-Musc	0.016248	0.028469	0.570714	0.05823
Cast-Dom	0.010176	0.014075	0.723007	0.465092
Spret-Musc	0.025391	0.043337	0.585893	0.048669
Spret-Dom	0.025382	0.039145	0.648411	0.122686
Dom-Musc	0.008355	0.026357	0.316976	0.00337*
Tcp-1	**Ka**	**Ks**	**Ka/Ks**	**P-Value**
Cast-Spret	0.005542	0.017701	0.313077	0.030355
Cast-Musc	0	0.002503	0	0*
Cast-Dom	0	0.002503	0	0*
Spret-Musc	0.005542	0.015146	0.365876	0.066618
Spret-Dom	0.005542	0.015146	0.365876	0.066618
Dom-Musc	0	0	NA	0*
Tagap1	**Ka**	**Ks**	**Ka/Ks**	**P-Value**
Cast-Spret	0.01146	0.018182	0.63028	0.276915
Cast-Musc	0.001825	0.008299	0.219845	0.100953
Cast-Dom	0.002752	0.008317	0.330845	0.167671
Spret-Musc	0.011466	0.015328	0.748063	0.580164
Spret-Dom	0.014313	0.015361	0.931799	0.796731
Dom-Musc	0.004594	0.005524	0.831637	0.86677
Nme3	**Ka**	**Ks**	**Ka/Ks**	**P-Value**
Cast-Spret	0	0.024167	0	0*
Cast-Musc	0.002627	0	NA	0*
Cast-Dom	0	0.007969	0	0*
Spret-Musc	0.002627	0.024259	0.10829	0.045269
Spret-Dom	0	0.032399	0	0*
Dom-Musc	0.002627	0.008	0.328399	0.373609

Sliding Window Ka/Ks

Sequence data from four inbred strains of mice were used to make interspecific pairwise comparisons of genes associated with the *t*-complex. Plots of sliding window Ka/Ks ratios are shown in Figures 6A-6F. Windows which have non-synonymous substitutions but do not have synonymous substitutions result in a divide-by-zero error. This error is indistinguishable from zero when graphed, resulting in a graph which does not show all of the non-synonymous sites. This error was corrected in these plots by adding the Ks value of one synonymous substitution to each window.

The sliding window analysis of putative distorter gene FYVE, RhoGEF and PH domain-containing protein 2 (*Fgd2*) in Figure 6A shows a series of low Ka/Ks peaks across the sequence. Low Ka/Ks peaks were observed upstream of the Rho guanine exchange factor (Rho GEF) domain and between FYVE and Pleckstrin homology domains in multiple comparisons, including Dom-Musc. All comparisons except Dom-Musc had low peaks in the Rho Guanine exchange factor domain.

Figure 6B shows sliding window analysis of *T* sequence comparisons. Low Ka/Ks peaks were observed downstream of the T-box DNA binding domain in comparisons with Spret sequence. Ka/Ks sliding window comparisons of *Smok4a* sequences in Figure 6C yielded numerous moderate and high peaks. Peaks occurred within and downstream of the Serine/Threonine kinase domain.

Tcp-1 sequences (Figure 6D) produced a low plateau and a high peak in Spret comparisons with each of the other sequences. Figure 6E shows sliding window Ka/Ks comparisons of *Tagap1* sequences. Each of these comparisons yielded moderate and high peaks, all of which are downstream of a short CDD match for a GTPase activating

protein domain. Comparisons with Spret produced a low peak around this domain. Figure 6F shows Ka/Ks sliding window analyses of Nucleoside diphosphate kinase 3 (*Nme3*) sequences. *Nme3* contains a Nucleoside diphosphate kinase Group I domain which spans most of the sequence. Only comparisons of Musc with other sequences elevated Ka/Ks ratios, while *Nme3* was conserved in other strains.

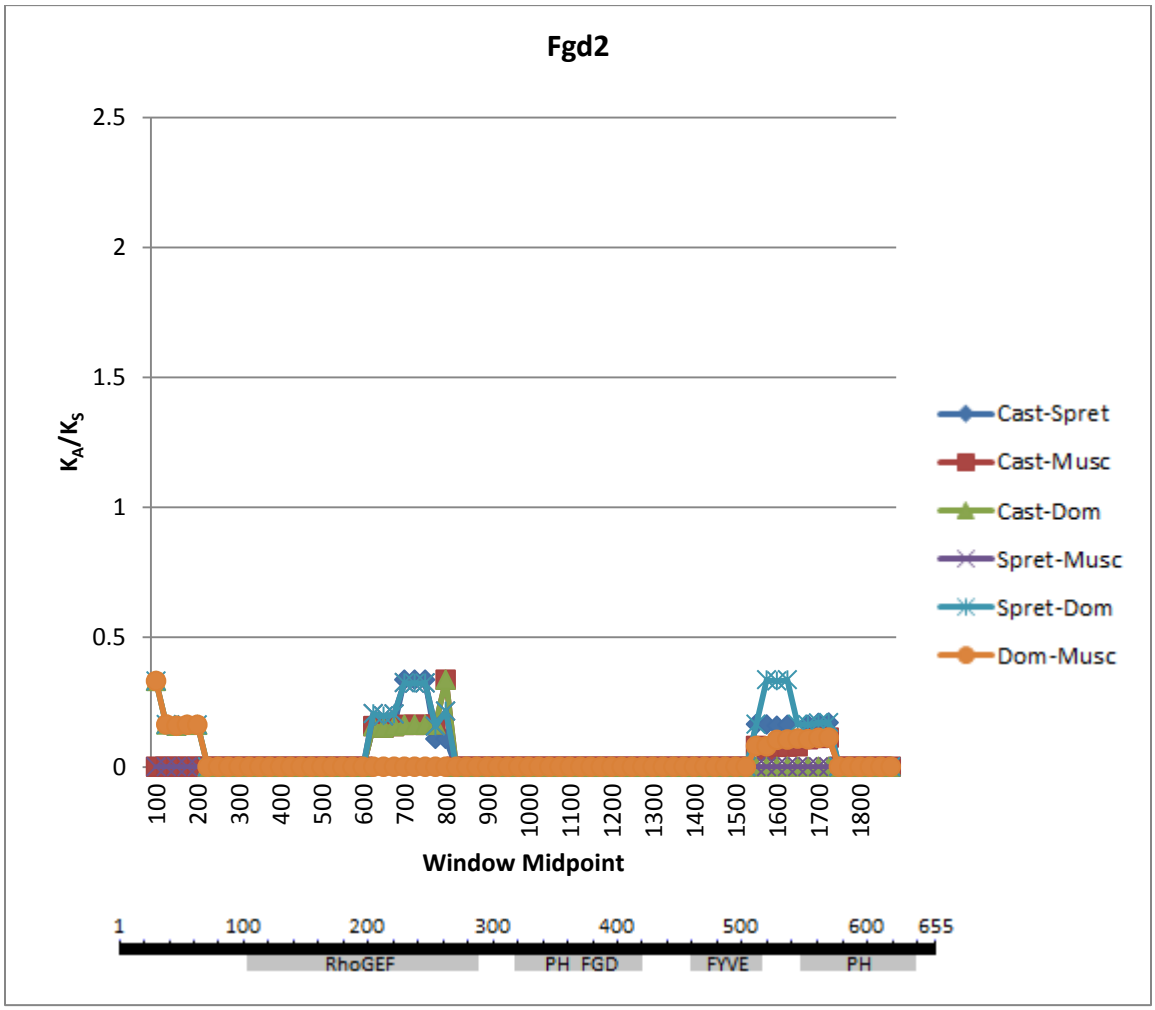

Figure 6A: Ka/Ks Sliding Window for *Fgd2*. Ka and Ks were calculated in a 200 bp sliding window using DnaSP 5.10. Pairwise comparisons were made between coding sequences extracted from *de novo* chromosomal assemblies for each strain. The Ks value for one synonymous substitution was added to each window to prevent divide-by-zero errors. Conserved Domain Database information is displayed below the horizontal axis.

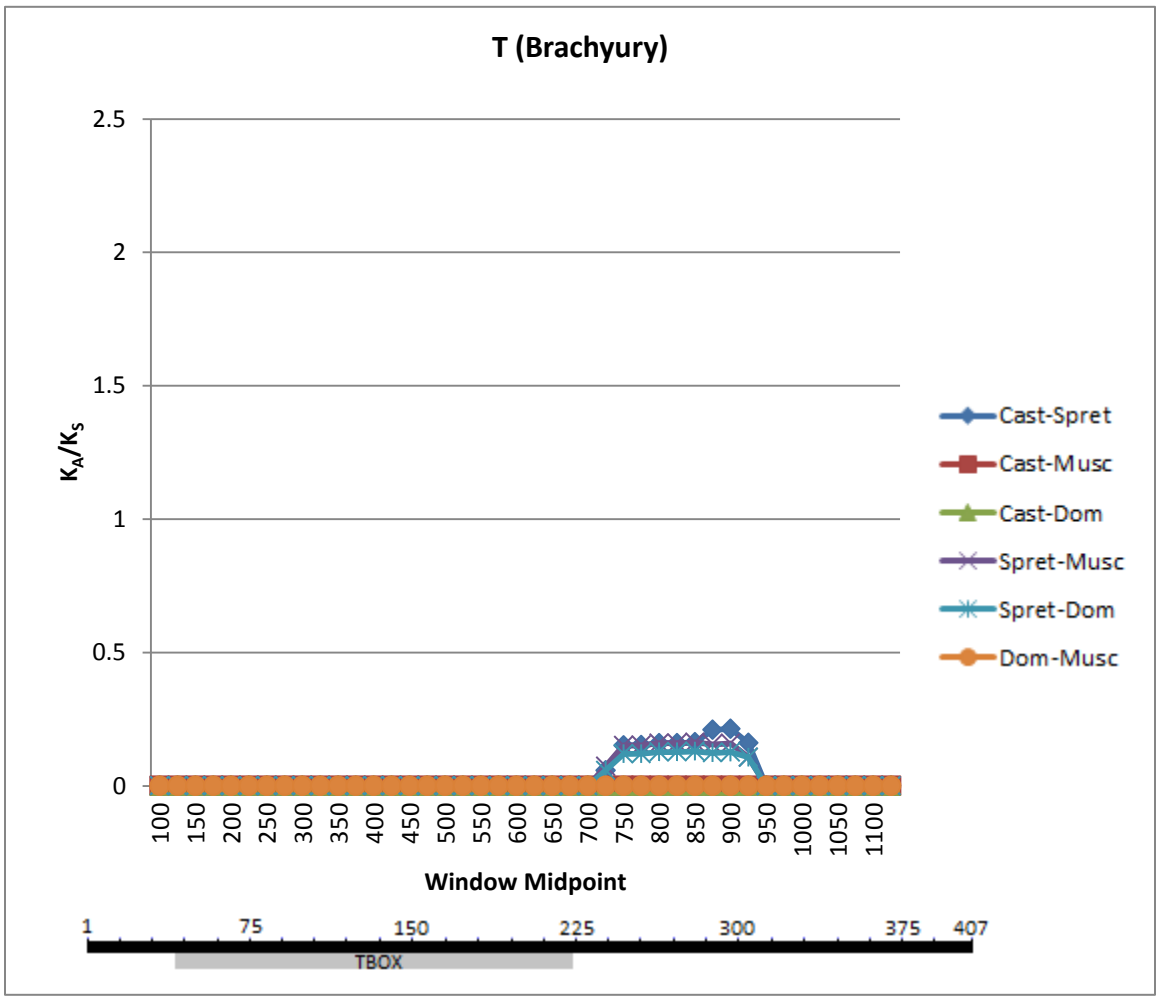

Figure 6B. Ka/Ks Sliding Window for *T* (Brachyury Protein). Ka and Ks were calculated in a 200 bp sliding window using DnaSP 5.10. Pairwise comparisons were made between coding sequences extracted from *de novo* chromosomal assemblies for each strain. The Ks value for one synonymous substitution was added to each window to prevent divide-by-zero errors. Conserved Domain Database information is displayed below the horizontal axis.

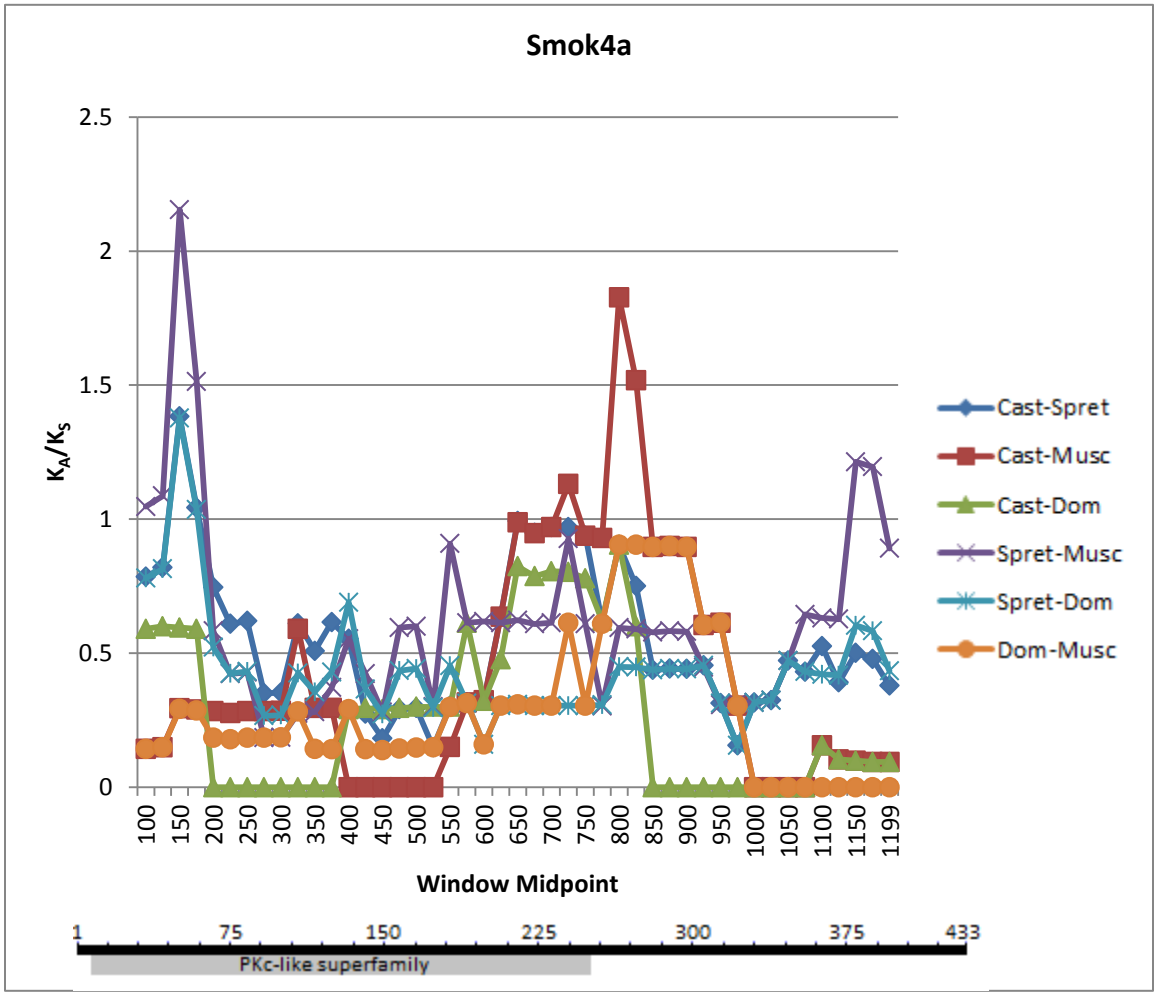

Figure 6C. Ka/Ks Sliding Window for *Smok4a*. Ka and Ks were calculated in a 200 bp sliding window using DnaSP 5.10. Pairwise comparisons were made between coding sequences extracted from *de novo* chromosomal assemblies for each strain. The Ks value for one synonymous substitution was added to each window to prevent divide-by-zero errors. Conserved Domain Database information is displayed below the horizontal axis.

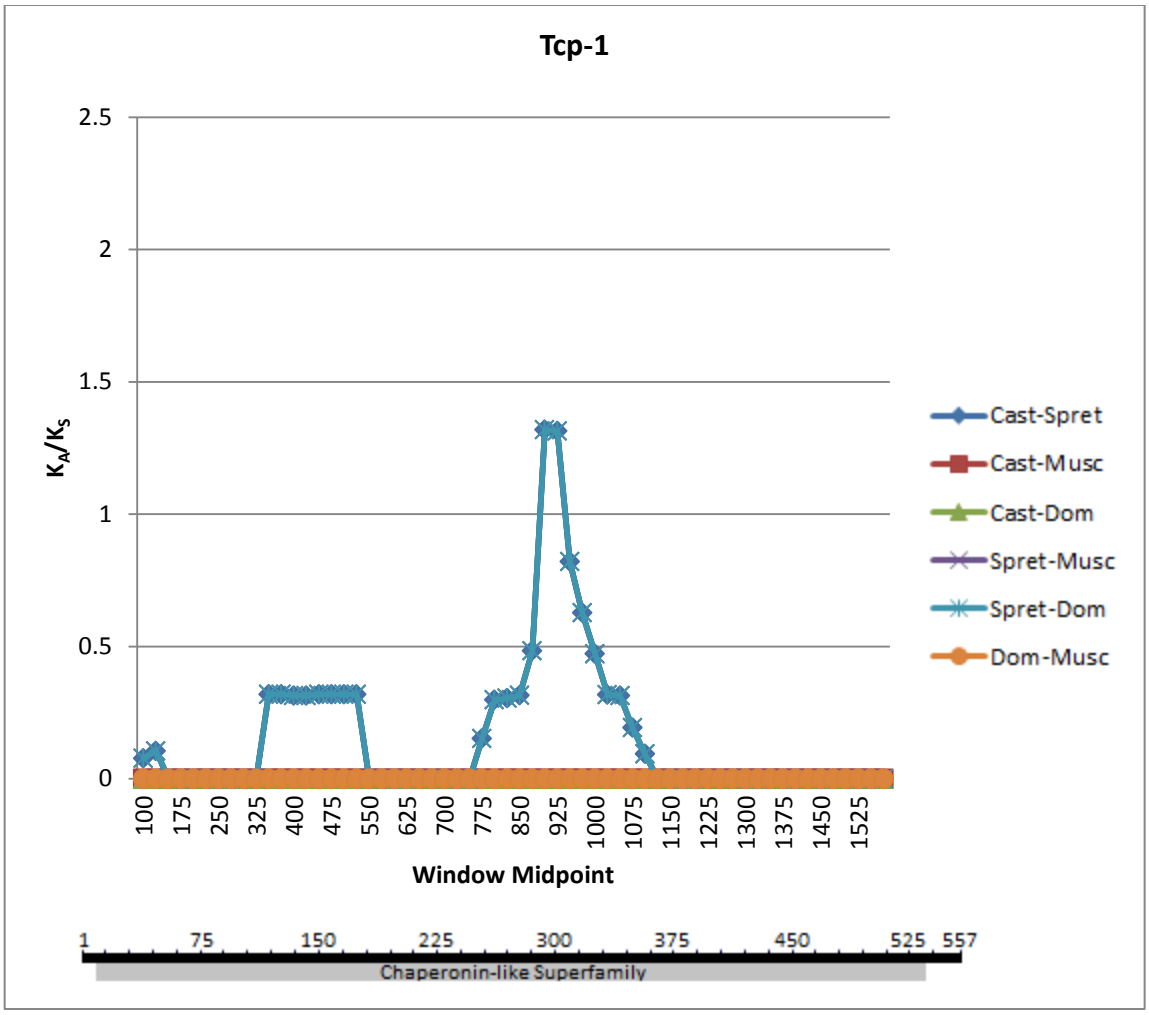

Figure 6D. Ka/Ks Sliding Window for *Tcp-1*. Ka and Ks were calculated in a 200 bp sliding window using DnaSP 5.10. Pairwise comparisons were made between coding sequences extracted from *de novo* chromosomal assemblies for each strain. The Ks value for one synonymous substitution was added to each window to prevent divide-by-zero errors. Conserved Domain Database information is displayed below the horizontal axis.

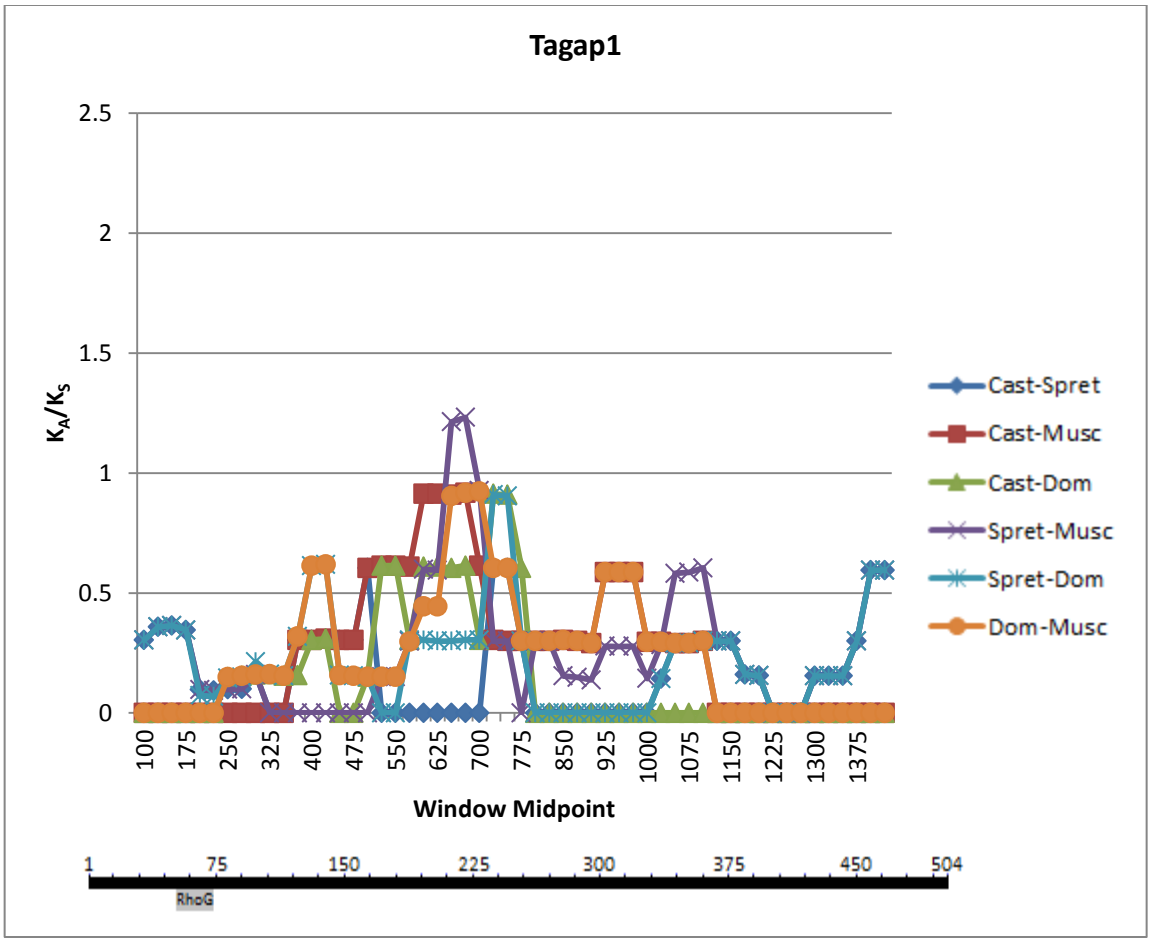

Figure 6E. Ka/Ks Sliding Window for *Tagap1*. Ka and Ks were calculated in a 200 bp sliding window using DnaSP 5.10. Pairwise comparisons were made between coding sequences extracted from *de novo* chromosomal assemblies for each strain. The Ks value for one synonymous substitution was added to each window to prevent divide-by-zero errors. Conserved Domain Database information is displayed below the horizontal axis.

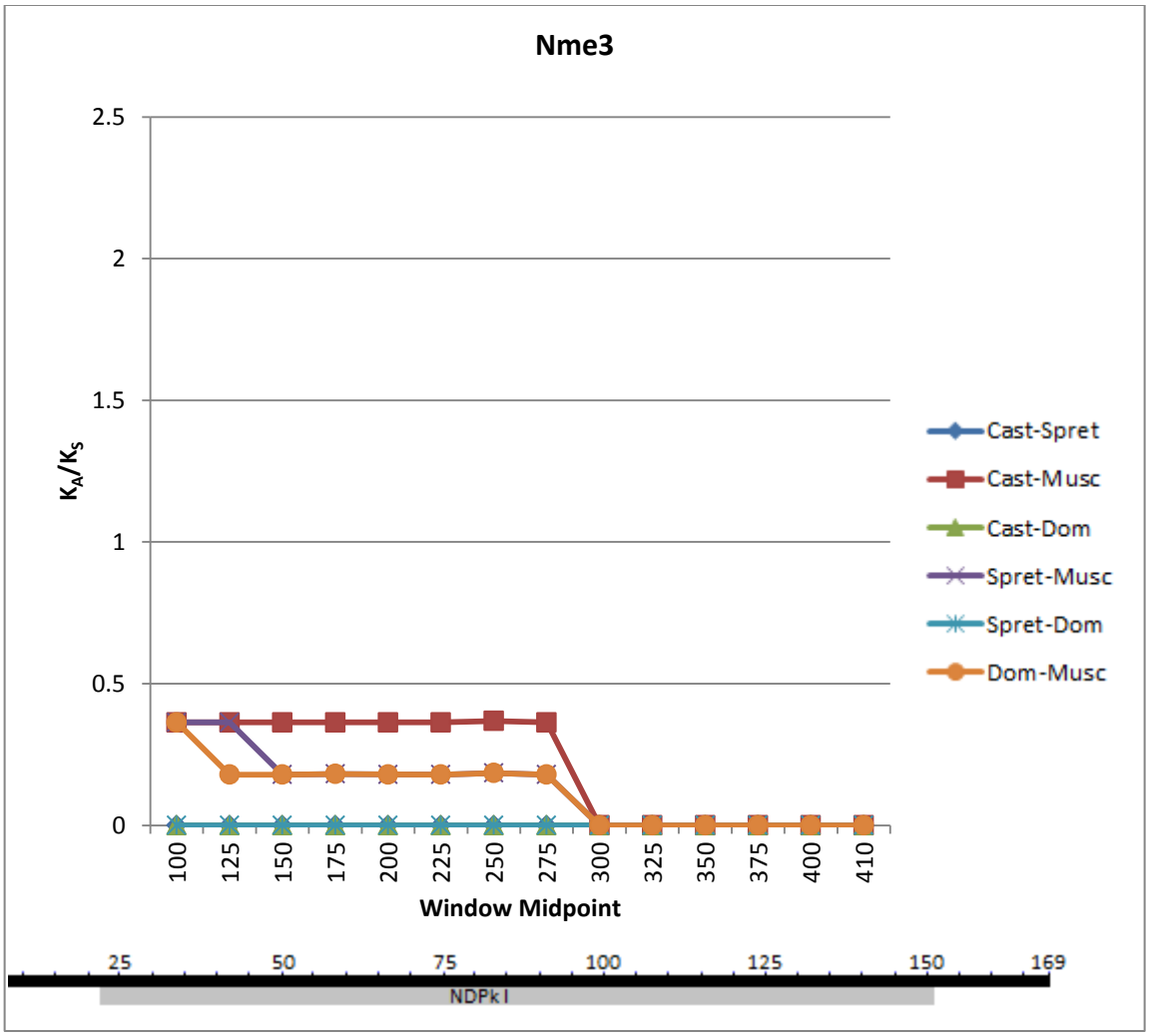

Figure 6F. Ka/Ks Sliding Window for *Nme3*. Ka and Ks were calculated in a 200 bp sliding window using DnaSP 5.10. Pairwise comparisons were made between coding sequences extracted from *de novo* chromosomal assemblies for each strain. The Ks value for one synonymous substitution was added to each window to prevent divide-by-zero errors. Conserved Domain Database information is displayed below the horizontal axis.

Discussion

Comparisons of rates of synonymous and non-synonymous substitutions revealed different kinds of selection acting on genes associated with the *t*-complex. Calculation of whole-gene Ka/Ks ratios indicated whether genes had experienced purifying selection, positive selection, or were not significantly different from neutrality. Sliding window calculation of Ka/Ks ratios provided a closer look at synonymous and non-synonymous substitutions across the gene sequences, and how they correspond to domains within those genes.

Whole-gene comparisons of *Fgd2* yielded indicated the action of purifying selection. Sliding window analyses yielded a low peak in each comparison with Dom upstream of the RhoGEF domain. These peaks are caused by a Serine-to-Isoleucine replacement in the Dom sequence, which changes the site from polar to non-polar and may affect protein shape. The Dom-Musc Ka/Ks ratio was lower than genome-wide and chromosome 17 median Ka/Ks, indicating purifying selection. All comparisons except Dom-Musc yielded low peaks in the GTPase interaction region of the RhoGEF domain. The peak between the FYVE and Pleckstrin homology (PH) domain is caused by Aspartic acid ↔ Glutamic acid substitutions. These substitutions do not cause a change in polarity and their position between domains suggest that these peaks represent neutral changes to a gene under purifying selection.

The *t*-haplotype form of *Fgd2* contains a single non-synonymous substitution in the RhoGEF domain (Bauer et al., 2007). This single amino-acid substitution causes a functional change which results in up-regulation of Rho-G proteins, contributing to abnormal flagella development. The highly conserved sequences of wild-type strains

suggest that *Fgd2* may be a more recent distorter in the *t*-haplotype. Disruption of the function of this protein is likely deleterious without the responder, *Tcr*. It is probable that the *t*-haplotype evolved transmission ratio distortion before the *t*-allele of *Fgd2* became a distorter.

The *T* gene, which codes for the brachyury protein, was found to be highly conserved. The calculated whole-gene Ka/Ks ratios were significantly low, indicating purifying selection in all comparisons. Ka/Ks sliding windows produced a low peak in some comparisons. Comparisons of Spret with other sequences produced peaks downstream of the T-box DNA binding domain. Amino acid substitutions in this gene may affect protein shape in *M. spretus*, but the position and low frequency of these substitutions suggest *T* is under purifying selection.

Whole-gene comparisons of *Smok4a* sequences produced multiple elevated Ka/Ks ratios. Comparisons with Spret were highest, but less than 1.0. Fisher's exact test for the Dom-Musc comparison yielded a significant P-value, indicating purifying selection. However, the Ka/Ks ratio for this comparison is more than double and triple the median Ka/Ks for chromosome 17 and genome-wide, respectively. Elevated Ka/Ks indicates positive selection in all comparisons. Sliding window analysis shows Ka/Ks is elevated throughout the gene sequence. Dom-Musc in particular shows low-to-moderate Ka/Ks throughout the Serine/Threonine kinase catalytic domain and high Ka/Ks downstream of the domain. All comparisons of *Smok4a* sequence appear to indicate positive selection.

The *t*-haplotype form of *Smok4a* (*Tcr*) contributes to transmission ratio distortion by rescuing *t*-sperm from abnormal flagella development. Non-synonymous mutations in the PKc-like domain are thought to enhance signaling to rescue *t*-sperm from impaired

motility (Hermann et al., 1999). Rapid evolution of this gene likely contributed to the development of the rescue effect of *Tcr*, which allows the *t*-haplotype to cause transmission ratio distortion.

Comparisons of Spret *Tcp-1* sequences to other strains yielded slightly elevated Ka/Ks ratios. P-values for these comparisons were low but not significant to indicate purifying selection. Sliding window plots show low Ka/Ks peaks in the proximal half of the sequence, and a tall peak around the 925 bp window. This tall peak occurs in the middle of the chaperonin-like domain, and indicates positive selection on this region which is less apparent in the whole-gene comparisons. Musc and Dom *Tcp-1* sequences are identical, and Cast only has non-synonymous differences.

Comparisons of *Tagap1* sequences show elevated Ka/Ks ratios indicating positive selection. Comparisons of Spret with the other sequences had the highest Ka/Ks. The Dom-Musc comparison produced a Ka/Ks ratio of 0.831637, which is greatly elevated above chromosome 17 and genome-wide median Ka/Ks. Dom-Musc Ka/Ks for *Tagap1* is greater than 93% of genes on chromosome 17. Sliding window analysis of *Tagap1* pairwise comparisons showed Ka/Ks is elevated throughout the sequence downstream of a Rho GTPase interaction site. Comparisons with Spret had a low peak in this region as well. Whole-gene Ka/Ks as well as sliding window analysis indicate positive selection in *Tagap1*.

The remarkable Ka/Ks ratios found in wild-type *Tagap1* sequences suggest that diversification of this gene has been favored by positive selection. Rapid evolution of *Tagap1* likely occurred in the *t*-haplotype as well, creating a new allele which contributes to TRD. Bauer et al. (2005) showed that the *t*-haplotype version of *Tagap1* has a

premature stop codon downstream of the functional domain. While it is not clear exactly how changes to this region contribute to TRD, it is possible that mutation of this region leads to mislocalization of the protein, as in the case of the segregation distorter *Sd* in *Drosophila melanogaster* (Merrill et al., 1999).

Whole-gene Ka/Ks showed comparisons of Musc *Nme3* with other sequences to have elevated non-synonymous substitutions. Dom-Musc Ka/Ks for *Nme3* is higher than 74% of chromosome 17 genes. Sliding window comparisons of Musc with the other sequences showed low Ka/Ks across the first half of the NDPkI domain. The Musc *Nme3* sequence has a single Arginine-to-Glutamine substitution, while the other sequences have no non-synonymous substitutions. *Nme3* is conserved in most strains but appears to show positive selection in Musc.

Ka/Ks tests of selection have found that genes which contribute to transmission ratio distortion in the *t*-haplotype are rapidly evolving, while other genes in the complex were being maintained by purifying selection. Comparisons of wild-type alleles from *M. domesticus, M. musculus, M. castaneus,* and *M. spretus* have shown that the TRD genes tend to accumulate protein-changing mutations without the inversions which prevent recombination in the *t*-haplotype. Hammer et al. (1999) proposed that the proximal inversion of the *t*-haplotype suppressed recombination allowing TRD to evolve, a selfish DNA strategy which accelerated as additional inversions and distorter loci became part of the *t*-haplotype. The results of this study add to this model of *t*-haplotype evolution by showing that TRD genes in the *t*-haplotype are rapidly evolving. It is likely that the propensity of these genes to accumulate non-synonymous mutations played an important role in the evolution of the *t*-haplotype. If these genes had instead been maintained by

purifying selection it is likely that the *t*-haplotype would have not have developed TRD and would be a simple structural variant of the proximal third of chromosome 17.

Chapter IV: Tests of Selection in Highly-Introgressed SNP Regions

Background

Multiple studies of the *Mus domesticus* – *Mus musculus* hybrid zone using molecular markers have found both low-introgressing genomic regions, potentially containing genes contributing to reproductive isolation, and highly introgressing genomic regions, potentially containing genes which raise fitness (Macholan et al., 2007; Payseur et al., 2004; Teeter et al., 2008; 2009; Tucker et al., 2011). Prior work has also shown that divergence in individual genes can cause hybrid incompatibility and such divergence can occur under recurrent positive selection (Barbash et al., 2003; Presgraves et al., 2003). Unpublished data provided by P. K. Tucker identified several genomic regions associated with highly introgressing SNPs in the Saxony transect (Table 5). Highly introgressing markers may indicate adaptive introgression by one or more linked genes (Barton & Hewitt, 1985; Riesberg et al., 1999). Hybridization is believed to provide genetic variation for adaptation. Alleles responding to strong selection may overcome barriers to gene flow (Payseur et al., 2004), such as the reduced fitness of hybrids which acts as a barrier to gene flow in the *M. domesticus* – *M. musculus* contact zone. Coding sequences of alleles involved in adaptive introgression may contain a signature of selection. Sufficiently strong signatures of selection may be detected by comparing the number on non-synonymous nucleotide substitutions to synonymous changes.

The number of non-synonymous mutations (Ka) increases relative to the number of synonymous mutations (K_S) when an allele experiences recurrent positive selection. Therefore, by calculating the Ka/K_S ratio in sequence differences between gene

orthologs, the action of positive selection can be inferred. A Ka/K_S ratio greater than 1.0 provides definitive evidence of positive selection, though an elevated ratio which is less than 1.0 may still be suggestive (Presgraves et al., 2003). Furthermore, a sliding-window approach to calculating Ka/K_S ratios can narrow down sequence evolution to specific regions of a gene and reveal significant Ka/K_S peaks in a gene for which the overall Ka/K_S ratio is not significant. The sliding window method increases the power of the test to detect selection as well as determine what parts of a gene selection may be acting on. Calculation of Ka/K_S ratios for alignments of *M. domesticus* and *M. musculus* forms of highly-introgressing genomic regions may reveal evidence of positive selection.

The SNPs 11.0154, 16.0587, and 19.0381 have displayed high introgression between *M. domesticus* and *M. musculus* genetic backgrounds. SNPs 11.0154 and 16.0587 have introgressed from the *M. musculus* side of the hybrid zone into the *M. domesticus* side, and 19.0381 has introgressed in the opposite direction. Introgression of these SNPs may be due to positive selection on genes contained within these genomic regions. While tests such as Tajima's D and the McDonald–Kreitman test require sequences from multiple individuals per species, Ka/K_S ratios can infer selection using only one protein-coding sequence from each of the two species.

Table 5: Introgressed SNP Regions. (A) The SNPs 11.0154, 16.0587, and 19.0381 were identified by P. K. Tucker (unpublished data) as demonstrating significant gene flow across the hybrid zone. Direction of gene flow is indicated above as *M. musculus* to *M. domesticus* (M2D) and *M. domesticus* to *M. musculus* (D2M). Genomic regions were expanded to include genes which had boundaries extending outside the genomic regions. (B) Names and genomic locations are listed for genes found within introgressed SNP regions. Amino acid sequences were used to query the Conserved Domain Database.

A

	SNP 11.0154	**SNP 16.0587**	**SNP 19.0381**
Direction of Introgression	M2D	M2D	D2M
Genomic Region	chr11:15,396,612-15,414,638 (18 Kb)	chr16:58,511,239-58,715,719 (204 Kb)	chr19:38,055,137-38,117,191 (62 Kb)
Expanded Region	N/A	chr16:58,470,655-58,715,719 (245 Kb)	chr19:37,973,526-38,118,067 (144 Kb)
Genes	N/A	*Cpox*, *E330017A01RiK*, *Gm813*, *St3gal6*	*Myof*

B

Gene Symbol	**Gene Name**	**Location**	**Conserved Domain Database Matches**
Cpox	Coproporphyrinogen oxidase	chr16:58,670,321-58,680,502	Coprogen Oxidase
E330017A01RiK	RIKEN cDNA E330017A01 gene	chr16:58,635,375-58,638,516	Euk-Ferritin
Gm813	Mus musculus predicted gene 813	chr16:58,613,799-58,617,091	Ferritin-like Superfamily
St3gal6	ST3 beta-galactoside alpha-2,3-sialyltransferase 6	chr16:58,470,655-58,523,425	Glycosyltransferase family 29
Myof	Myoferlin	chr19:37,973,526-38,118,067	C2A_Ferlin, C2B_Ferlin, FerI, C2C_Ferlin, FerA, FerB, Dysferlin, C2D_Ferlin, C2 Superfamily, C2E_Ferlin, C2F_Ferlin

Methods

The regions of interest surrounding SNPs 11.0154, 16.0587, and 19.0381 were checked for genes using the UCSC Genome Browser (genome.ucsc.edu) with the July 2007 release of the mouse genome. The region around SNP 11.0154 (chr11: 15,396,612-15,414,638) is approximately 18 kb and contains no genes. The region around SNP 16.0587 (chr16: 58,511,239-58,715,719) is approximately 204 kb and contains the genes *St3gal6*, *Gm813*, *E330017A01RiK*, and *Cpox*. The proximal border of this region occurs inside the *St3gal6* gene so the region of interest was expanded to 58,470,655-58,715,719 (245 kb) to include the entire gene. The region around SNP 19.0381 (chr19: 38,055,137-38,117,191) is approximately 62 kb and is entirely within the borders of the *Myof* gene. This region of interest was expanded to 37,973,526-38,118,067 (144 kb) to include the whole of the *Myof* gene.

De novo assemblies of chromosomal sequences were downloaded from the Sanger website and reduced to the regions of interest using the EmEditor Pro program for editing large text files (Emurasoft, Inc.). Coding sequences were built in MEGA 5.05 from *de novo* assemblies based on intron-exon boundaries and coding start and stop sites described in the Refseq profile of each gene. Missing coding sequence was reconstructed manually from the original read alignments of the Mouse Genomes project. Read alignments (.bam files) were navigated using Tablet 1.12. Aligned coding sequences were loaded into KaKs_Calculator 2.0 which calculated Ka/K_S ratios for each pair of gene sequences. A sliding window calculation of Ka and K_S values for each gene was also performed in DnaSP using a 200bp window and 25bp step size. Ka/K_S ratios were graphed using Microsoft Excel 2010.

Results – Whole Gene Ka/Ks

Sequence data from four inbred strains of mice were used to make interspecific pairwise comparisons of genes in the highly introgressed SNP regions. Whole-gene Ka/Ks ratios were calculated using the Nei & Gojobori (1986) method in KaKs_Calculator 2.0 (Zhang et al., 2006). Results of these comparisons are listed in Table 6. Synonymous substitutions and sites were compared with non-synonymous substitutions and sites by Fisher's exact test. Critical values were corrected for multiple tests by Bonferroni correction.

Genome-wide calculations of synonymous and non-synonymous substitutions for Musc and Dom genomes were provided by V. Janoušek. Median Ka/Ks ratios were determined to be 0.092784 across the genome. The median Ka/Ks ratio for chromosome 11 was 0.065924. No genes were found in the SNP 11.0154 region for comparison with this value. The median Ka/Ks ratio for chromosome 16 was 0.120299. The 245 kb SNP 16.0587 region contains the genes *Cpox*, *E330017A01RiK*, *Gm813*, and *St3gal6*.

Pairwise comparisons of *Cpox* (Coproporphyrinogen oxidase) sequence primarily showed evidence of purifying selection. The exception to this is the Cast-Musc comparison, which had non-synonymous substitutions but no synonymous substitutions. This causes a divide-by-zero error, resulting in a N/A output from KaKs_Calculator. Elevated non-synonymous substitutions over synonymous substitutions were found to be significant by Fisher's exact test, indicating positive selection.

Pairwise comparisons of *E330017A01RiK* were variable. Comparisons of Spret sequences to other strains showed elevated non-synonymous substitutions relative to other comparisons for this gene, resulting in low but non-significant P-values and a

moderately elevated Ka/Ks ratio of 0.566352 when paired with Dom. Pairwise comparisons among Musc, Dom, and Cast sequences were suggestive of purifying selection, with no substitutions occurring between Cast and Musc.

Ka/Ks ratios of predicted gene *Gm813* were elevated but not significant, and may suggest weak or recent positive selection. The exception is the Cast-Dom comparison, which showed no substitutions in the sequence. Ka/Ks for Dom-Musc was 0.268812, which is not high but is elevated from median Ka/Ks across the genome (0.1) and in chromosome 16 (0.136842).

Pairwise comparisons of *St3gal6* (ST3 beta-galactoside alpha-2,3-sialyltransferase 6) were all suggestive of purifying selection, with the exception of Dom-Musc. Ka/Ks ratios for Cast-Dom and Spret-Dom were low but not significant. Conversely, Dom-Musc had only non-synonymous substitutions, indicating positive selection.

The 144 kb region around SNP 19.0381 contains the gene *Myof* (Myoferlin). Data from V. Janoušek shows the median Ka/Ks for chromosome 19 is 0.084848. Ka/Ks values for pairwise comparisons of *Myof* are all significantly low and comparable to median Ka/Ks values for chromosome 19 and across the genome.

Table 6: Ka/Ks Ratios for Genes in Introgressed SNP Regions. Ka/Ks ratios were calculated for genes in regions surrounding SNP 16.0587 (*Cpox*, *E330017A01RiK*, *Gm813*, and *St3gal6*) and SNP 19.0381 (*Myof*). Calculations were made using the Nei & Gojobori (1986) method in KaKs_Calculator 2.0. P-Values result from Fisher's exact test comparing synonymous substitutions and sites with non-synonymous substitutions and sites. Bonferroni correction for multiple tests reduces the critical value to 0.0083. Starred P-Values indicate significance.

Cpox	**Ka**	**Ks**	**Ka/Ks**	**P-Value**
Cast-Spret	0.003995	0.028199	0.141689	0.000785*
Cast-Musc	0.000997	0	NA	0*
Cast-Dom	0.003995	0.021843	0.18291	0.006441*
Spret-Musc	0.002995	0.028199	0.106196	0.000307*
Spret-Dom	0	0.031386	0	0*
Dom-Musc	0.002994	0.021843	0.137091	0.002957*
E330017A01RiK	**Ka**	**Ks**	**Ka/Ks**	**P-Value**
Cast-Spret	0.009069	0	NA	0.033382
Cast-Musc	0	0	NA	0*
Cast-Dom	0	0.015888	0	0*
Spret-Musc	0.009069	0	NA	0.033382
Spret-Dom	0.009069	0.016012	0.566352	0.397426
Dom-Musc	0	0.015888	0	0*
Gm813	**Ka**	**Ks**	**Ka/Ks**	**P-Value**
Cast-Spret	0.014559	0.02731	0.533096	0.292774
Cast-Musc	0.002409	0.008962	0.268812	0.335427
Cast-Dom	0	0	NA	0*
Spret-Musc	0.017006	0.036697	0.463408	0.185556
Spret-Dom	0.014559	0.02731	0.533096	0.292774
Dom-Musc	0.002409	0.008962	0.268812	0.335427
St3gal6	**Ka**	**Ks**	**Ka/Ks**	**P-Value**
Cast-Spret	0.002594	0.037687	0.068831	0.000145*
Cast-Musc	0	0.013934	0	0*
Cast-Dom	0.001295	0.013945	0.092865	0.032691
Spret-Musc	0.002594	0.023331	0.111181	0.006982*
Spret-Dom	0.003893	0.02335	0.166743	0.015299
Dom-Musc	0.001295	0	NA	0*
Myof	**Ka**	**Ks**	**Ka/Ks**	**P-Value**
Cast-Spret	0.002356	0.03072	0.076687	4.66E-18*
Cast-Musc	0.001499	0.015188	0.098682	7.55E-09*
Cast-Dom	0.001498	0.013739	0.109065	8.05E-08*
Spret-Musc	0.001692	0.028153	0.060112	3.53E-18*
Spret-Dom	0.002116	0.022309	0.09483	1.14E-12*
Dom-Musc	0.000846	0.01214	0.069668	2.93E-08*

Results – Sliding Window Ka/Ks

Sequence data from four inbred strains of mice were used to make interspecific pairwise comparisons of genes in the highly introgressed SNP regions. Sliding windows of Ka/Ks ratios were calculated using DnaSP 5.10. Plots of sliding window Ka/Ks ratios are shown in Figures 7A-7E. Windows which have non-synonymous substitutions but do not have synonymous substitutions result in a divide-by-zero error. This error is indistinguishable from zero when graphed, resulting in a graph which does not show all of the non-synonymous sites. This error was corrected in these plots by adding the Ks value of one synonymous substitution to each window.

As seen in Figure 7A, Ka/Ks is highest at the beginning of the *Cpox* gene for all pairwise comparisons. This part of the coding region is outside of the coprogen oxidase domain, and less prominent Ka/Ks peaks were found at the beginning of the domain. Figure 7B shows sliding window Ka/Ks for E330017ARiK. Comparisons of Spret to other sequences produced low-to-moderate peaks within the Ferritin domain, while other comparisons did not produce peaks. *Gm813*, seen in Figure 7C, also has a Ferritin-like domain. Sequences produced low-to-moderate peaks across the domain, with the exception of the Cast-Dom comparison which did not produce peaks. Figure 7D shows Ka/Ks sliding window comparisons of *St3gal6* sequences. *St3gal6* contains a Glycosyltransferase family 29 domain. The highest Ka/Ks peaks occur before the start of the domain, and comparisons of Spret to other sequences has additional peaks further downstream within the domain. The Dom-Musc pairing had no synonymous substitutions. The plateau around 0.3 in this comparison occurs due to a single non-synonymous substitution before the Glycosyltransferase family 29 domain.

The final Ka/Ks sliding window analysis focused on the SNP 19.0381 region, containing *Myof*. Myoferlin is relatively long gene containing regions matching the domains C2A_Ferlin, C2B_Ferlin, FerI, C2C_Ferlin, FerA, FerB, Dysferlin, C2D_Ferlin, C2 Superfamily, C2E_Ferlin, and C2F_Ferlin. Pairwise comparisons produced low-to-moderate peaks in the first domain, C2A_Ferlin. Ka/Ks comparisons of Spret to all others produced a short peak in the first Dysferlin domain. Moderate peaks were also found in-between Dysferlin, C2D_Ferlin, C2E_Ferlin, and C2F_Ferlin domains.

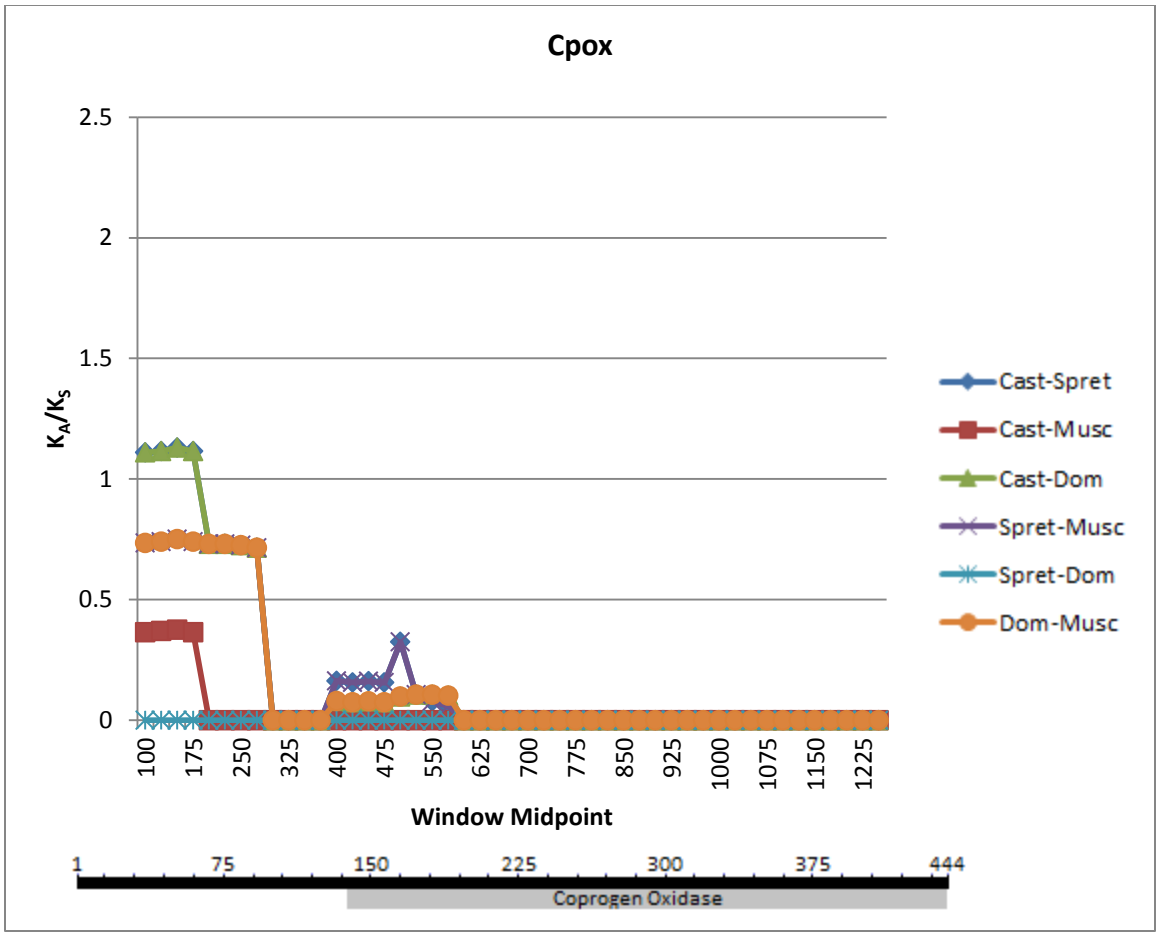

Figure 7A: Ka/Ks Sliding Window for *Cpox*. Ka and Ks were calculated in a 200 bp sliding window using DnaSP 5.10. Pairwise comparisons were made between coding sequences extracted from *de novo* chromosomal assemblies for each strain. The Ks value for one synonymous substitution was added to each window to prevent divide-by-zero errors. Conserved Domain Database information is displayed below the horizontal axis.

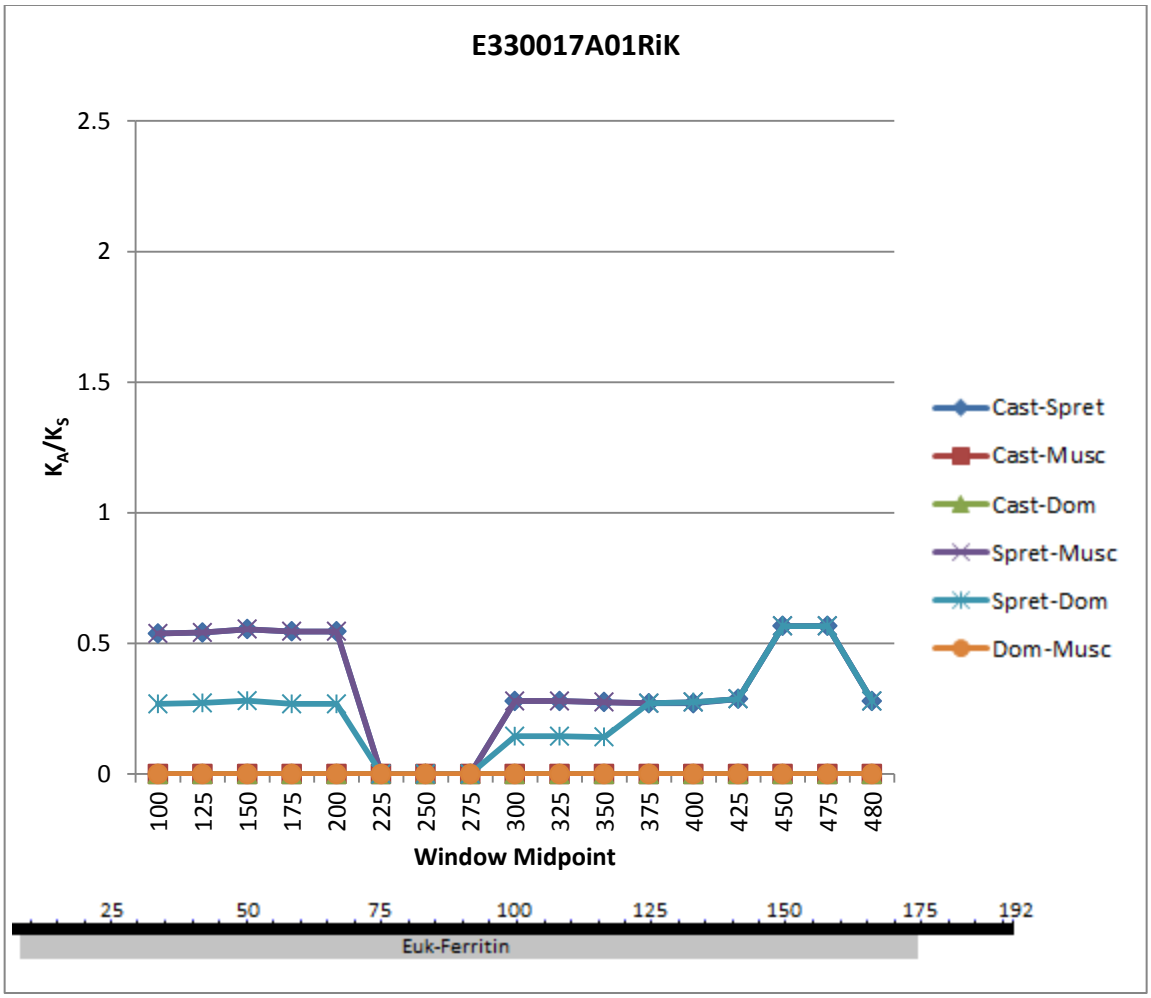

Figure 7B: Ka/Ks Sliding Window for *E330017A01RiK*. Ka and Ks were calculated in a 200 bp sliding window using DnaSP 5.10. Pairwise comparisons were made between coding sequences extracted from *de novo* chromosomal assemblies for each strain. The Ks value for one synonymous substitution was added to each window to prevent divide-by-zero errors. Conserved Domain Database information is displayed below the horizontal axis.

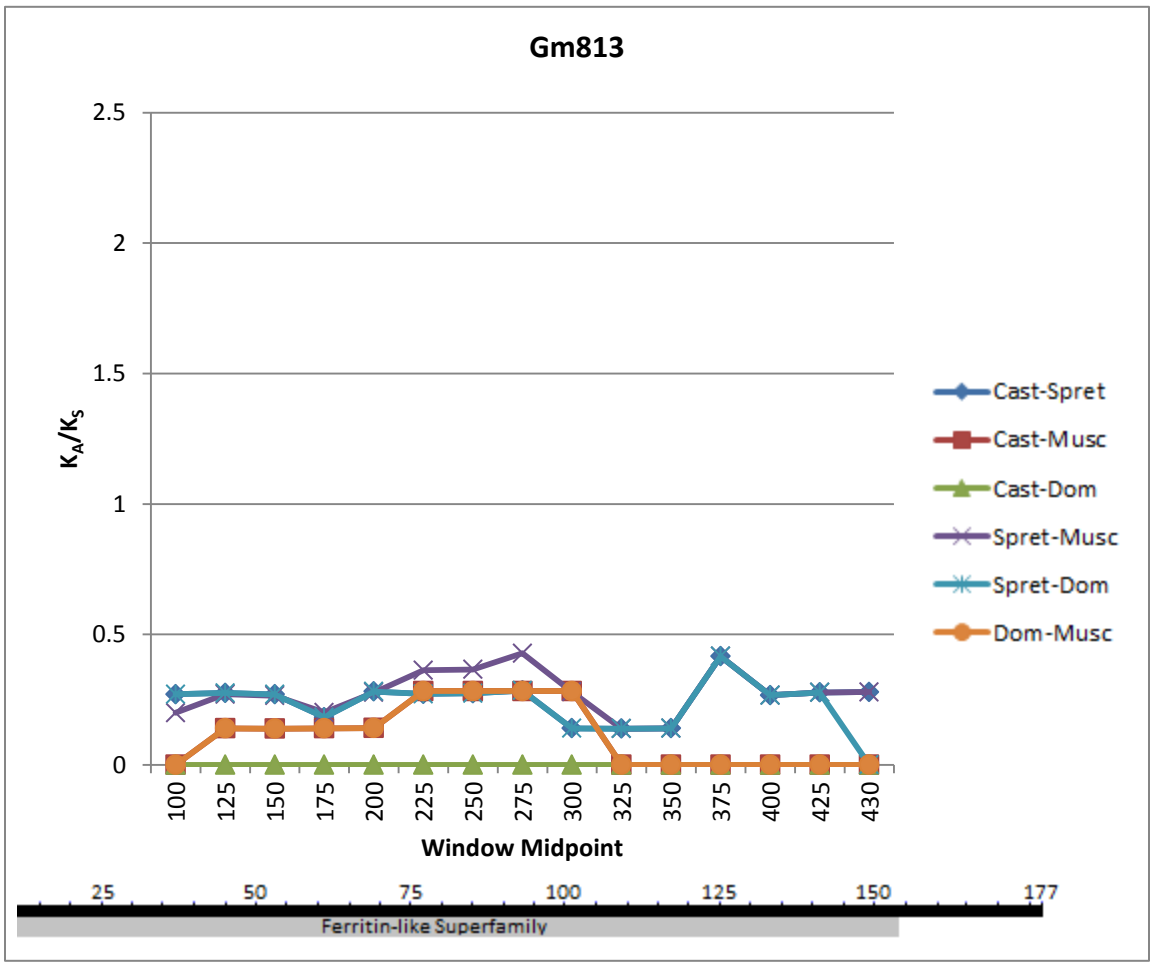

Figure 7C: Ka/Ks Sliding Window for *Gm813*. Ka and Ks were calculated in a 200 bp sliding window using DnaSP 5.10. Pairwise comparisons were made between coding sequences extracted from *de novo* chromosomal assemblies for each strain. The Ks value for one synonymous substitution was added to each window to prevent divide-by-zero errors. Conserved Domain Database information is displayed below the horizontal axis.

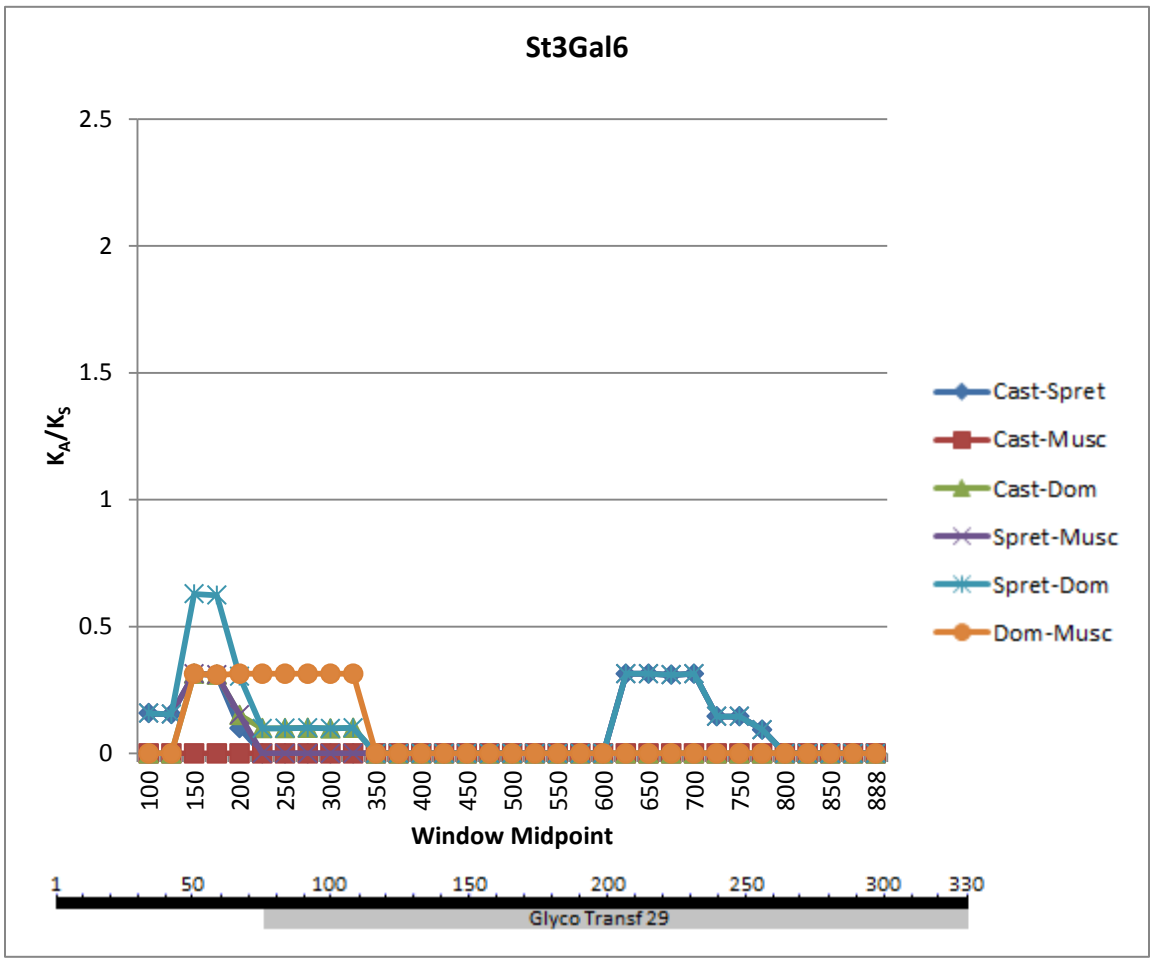

Figure 7D: Ka/Ks Sliding Window for *St3gal6*. Ka and Ks were calculated in a 200 bp sliding window using DnaSP 5.10. Pairwise comparisons were made between coding sequences extracted from *de novo* chromosomal assemblies for each strain. The Ks value for one synonymous substitution was added to each window to prevent divide-by-zero errors. Conserved Domain Database information is displayed below the horizontal axis.

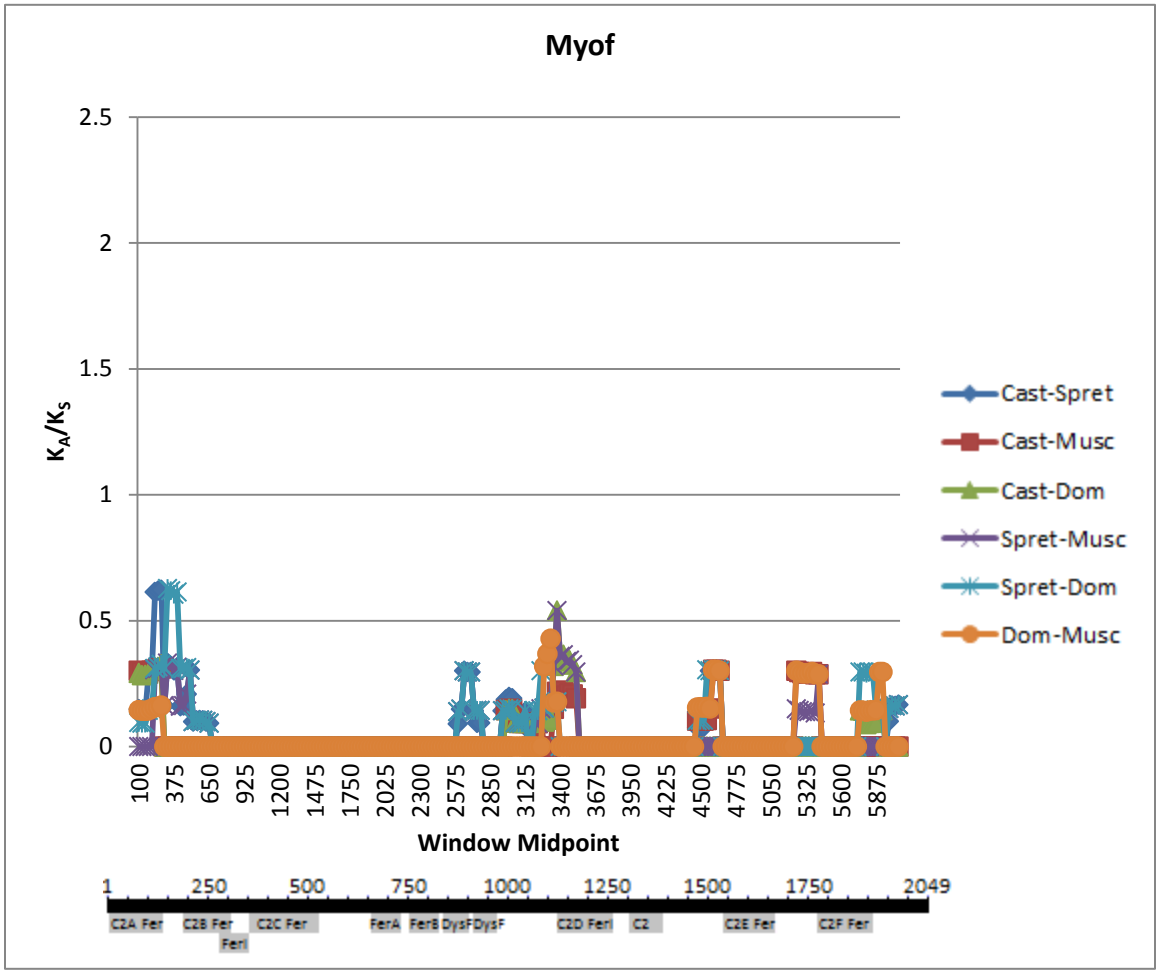

Figure 7E: Ka/Ks Sliding Window for *Myof*. Ka and Ks were calculated in a 200 bp sliding window using DnaSP 5.10. Pairwise comparisons were made between coding sequences extracted from *de novo* chromosomal assemblies for each strain. The Ks value for one synonymous substitution was added to each window to prevent divide-by-zero errors. Conserved Domain Database information is displayed below the horizontal axis.

Discussion

Comparisons of rates of synonymous and non-synonymous substitutions revealed different kinds of selection acting on genes in highly introgressing SNP regions. Calculation of whole-gene Ka/Ks ratios indicated whether genes had experienced purifying selection, positive selection, or were not significantly different from neutrality as determined by Fisher's exact test. Sliding window calculation of Ka/Ks ratios provided a closer look at synonymous and non-synonymous substitutions across the gene sequences, and how they correspond to domains within those genes.

The SNP 11.0154 region does not contain any genes. The nearest gene to the SNP 11.0154 region is V-set and transmembrane domain containing 2A (*Vstm2a*), which is 743 kb outside of the region. No tests of selection were performed on this distant gene. Whole-gene and sliding window Ka/Ks ratios were calculated for genes from the SNP 16.0587 and 19.0381 regions.

Comparisons for *Cpox* primarily produced low Ka/Ks ratios, indicating purifying selection. The Cast-Musc comparison for this gene, however, had synonymous substitutions and no non-synonymous substitutions. While this may seem to indicate positive selection, sliding window analysis reveals the non-synonymous substitutions occur at the proximal end of the gene, upstream of the Coprogen oxidase domain, where protein-changing mutations are less likely to be deleterious. This is also the case for most non-synonymous substitutions in the other pairwise comparisons. *Cpox* seems to be primarily acted on by purifying selection.

Comparisons of Spret *E330017A01RiK* with the other strains showed elevated non-synonymous substitutions. Cast-Spret and Spret-Musc comparisons showed only non-

synonymous substitutions, and Spret-Dom Ka/Ks was also elevated. Figure 7B shows that these protein-altering substitutions occur across the Ferritin domain which spans the gene sequence. Spret Ka/Ks ratios dropped in the 225 to 275 bp midpoint windows. No non-synonymous substitutions were seen in comparisons of Cast, Musc, and Dom. These tests of selection suggest that the Ferritin-like protein, *E330017A01RiK*, may have undergone positive selection in *M. spretus.*

Gm813 also has a Ferritin-like domain and Ka/Ks was elevated in comparisons of Spret to the other sequences. Figure 7C shows that Ka/Ks is elevated throughout the Ferritin-like domain in these comparisons, peaking around the 250 and 275 bp midpoints. This is contrary to what was seen in *E330017A01RiK* Spret sequences, which had Ka/Ks peaks surrounding this region of the domain but not in these windows. Cast and Dom sequences of *Gm813* are identical, providing no insight on potential selective pressures. Ka/Ks for Dom-Musc was found to be greater than both genomic and chromosomal median Ka/Ks ratios. Sliding window of Dom-Musc Ka/Ks also showed a low peak in the 200 to 300 bp midpoint windows. Ka/Ks peaks in this region occur due to a Threonine-to-Isoleucine substitution in the Musc (*M. musculus*) sequence, which is a change from a polar to a non-polar amino acid. Whether this substitution is neutral, slightly deleterious, or advantageous is unclear, but an advantageous mutation in Musc which is also successful in Dom may account for the introgression of the *M. musculus* SNP 16.0587 region into the *M. domesticus* genetic background.

Whole-gene Ka/Ks ratios for *St3gal6* are low in most comparisons and suggest the action of purifying selection. The Dom-Musc sequence comparison results show only a single non-synonymous substitution. While this may seem suggestive of positive

selection, sliding window analysis of this pairing shows that the Ka/Ks peak occurs at the proximal region of the sequence. The Dom sequence contains a Threonine-to-Isoleucine substitution. This substitution replaced a polar amino acid with one that is non-polar and may affect protein shape, but this replacement upstream of the Glycosyltransferase family 29 domain. This substitution is more likely to be neutral than if amino acid change had been in the domain.

Comparisons of *Myof* sequences resulted in significantly low Ka/Ks ratios in each pairing, indicating purifying selection. Myoferlin is a gene with a long, multi-domain sequence. The sliding window analysis of *Myof* showed several peaks in all comparisons across the coding sequence. Most peaks in these comparisons were between domains where an amino acid replacement is more likely to be neutral, but peaks were also found within domains. All sliding window comparisons showed low or moderate peaks in the C2A Ferlin domain. Comparisons with Spret sequences also showed a low Ka/Ks peak in the first Dysferlin domain.

Genes in highly introgressing SNP regions appear to be evolving under purifying selection and positive selection. Positive selection of *Gm813, St3gal6,* or *E330017A01RiK* may be the cause of the introgression of the SNP 16.0587 genomic region. The close linkage of these genes makes it difficult to determine which of these genes is the target of selection and the cause of adaptive introgression. The SNP 11.0154 region contains no genes and may be introgressing neutrally. The introgression of this region could possibly be due to positive selection of an unknown regulatory region, but investigation of this possibility is outside the scope of this study. The SNP 19.0381

region contains one highly conserved gene. The absence of any indication of positive selection suggests that this region has introgressed neutrally.

Introgression of genes under positive selection and purifying selection suggests that reproductive isolation between *M. domesticus* and *M. musculus* is limited. Introgression of these SNP genomic regions precludes the possibility of their involvement in reproductive incompatibility (Nolte et al., 2009). While hybrids are thought to have reduced fitness, heterosis has been observed in *M. domesticus* – *M. musculus* hybrids (Alibert et al., 1994; Alibert et al., 1997). It may be that heterozygosity at the loci studied here may be beneficial, resulting in the introgression of these SNP genomic regions across the hybrid zone.

Chapter V: Summary and Conclusions

Observations of gene flow and molecular evolution are important tools for studying the process of speciation. Genes which cause hybrid incompatibility reduce fitness of hybrids and act as barriers to gene flow, thereby contributing to speciation. Genes which accumulate non-synonymous mutations can have altered activity or function, disrupting coadapted gene complexes in hybrids. Alternatively, positively selected genes which do not cause hybrid incompatibilities can introgress at high frequencies. By studying these processes I have found that rapidly evolving genes in the *t*-complex may have led to the evolution of the *t*-haplotype and may contribute to reproductive isolation. Furthermore, high introgression of SNP genomic regions appears to be caused by genes under positive, neutral, and purifying selective pressures.

t-Haplotype Genotyping

Detection of the *t*-haplotype combined screening methods published by Schimenti & Hammer (1990), Morita et al. (1993), and Miller Baker (2008) for a high-confidence determination of the genotypes of the mouse DNAs used in this study. Genotyping results showed an overall frequency of 17.5% for the *t*-haplotype in the SX transect of the *M.domesticus – M. musculus* hybrid zone. The observed frequency is within the ranges described previously by Miller Baker (2008) and Ardlie & Silver (1998).

The *t*-haplotype was found exclusively in the *M. domesticus* side of the hybrid zone with the exception of one *M. musculus* individual. This individual, trapped in Kamenz, was positive for the *Tcp-1* marker but not the Hba marker. The bias of the *t*-haplotype as a *M. domesticus* haplotype in this system combined with the finding of a single partial *t*-

haplotype *M. musculus* individual suggests that there may be some element distal to the *Tcp-1* marker which is incompatible with some element of the *M. musculus* genetic background.

t-Complex Genes

Whole-gene and sliding window analyses of Ka/Ks ratios have implicated purifying selection and positive selection in genes associated with the *t*-haplotype. Sequence data of the *t*-haplotype was not available so representative wild-type genomic sequence was used. *Fgd2*, *T*, and *Tcp-1* showed evidence of purifying selection, indicating these genes are slow evolving. *Smok4a* and *Nme3* Ka/Ks ratios were elevated, and *Tagap1* Ka/Ks was high, indicating positive selection and more rapid evolution.

Fgd2, *Tagap1*, and *Nme3* are known distorter loci which impair sperm motility. The *t*-haplotype version of *Smok4a*, *Smok(Tcr)*, rescues sperm carrying the *t*-haplotype. The four inversions of the *t*-haplotype prevent recombination which allows mutations to accumulate. The presence of rapidly evolving genes in a chromosomal region which does not recombine readily likely contributed to the evolution of transmission ratio distortion in the *t*-haplotype.

Highly Introgressed SNP Regions

Whole-gene and sliding window analyses of Ka/Ks ratios have shown patterns of silent and protein-changing substitutions suggesting the action of different types of selection. Although the power of tests comparing only representative genome sequence data is more limited than what can be achieved with population data, the results of this study has

implications on the cause of the introgression of SNPs 11.0154, 16.0587, and 19.0381 across the hybrid zone.

While genes contributing to isolation in hybrid systems have reduced introgression, neutral and positively selected genes introgress with higher frequency (Barton & Hewitt, 1985; Rieseberg et al., 1999). The genomic region around SNP 11.0154 does not contain any genes and is 743 kb from the closest gene. Intergenic non-coding DNA, sometimes referred to by the misnomer "junk DNA", can contain important regulatory sequences (Birney et al., 2007; Dunham et al., 2012). Barring any strongly selected regions or sequences contributing to genetic isolation, the SNP 11.0154 region would be free to drift neutrally from *M. musculus* to *M. domesticus* because it is not closely linked to any genes.

The genomic region around SNP 16.0587 was also chosen due to a strong *M. musculus* to *M.domesticus* introgression. Comparisons of *Gm813* and *St3gal6* showed non-synonymous substitutions elevated above genome-wide and chromosome 16 median Ka/Ks. These may be evolving neutrally or under weak positive selection.
The high conservation observed in genes under purifying selection makes these genes unlikely to have interactions causing genetic isolation. Purifying selection detected in Dom-Musc Ka/Ks of *E330017A01RiK* and *Cpox* would not be expected to interfere with introgression of the SNP 16.0587. Likewise, evolution of *Myof* is governed by purifying selection, allowing neutral introgression of the SNP 19.0381 region.

Conclusions

The results of *t*-haplotype genotyping and Ka/Ks ratio tests of selection in *t*-complex genes and highly introgressing SNP regions illustrate the state of the *M. domesticus – M. musculus* hybrid zone.

It is well established that *M. domesticus* and *M. musculus* hybridize and there is gene flow between the two species. Loci which evolve neutrally or by positive selection are able to introgress across the hybrid zone. This was seen in the highly introgressed SNP regions which showed genes evolving neutrally or under weak positive selection, as well as conserved gene sequences.

Loci which result in hybrid incompatibilities act as barriers to gene flow. The *t*-haplotype does not recombine due to inversions, preventing the rapidly evolving genes in the complex from becoming separated and individually selected against. This rapidly evolving complex of genes appears to have some genetic incompatibility with the *M. musculus* genetic background. Deleterious recessive alleles in the complex prevent the *t*-haplotype from achieving fixation in wild populations despite TRD. Even at a frequency of 17.5%, the *t*-haplotype may contribute to reproductive isolation by prevent *t*-positive *M. domesticus* mice from reproducing with *M. musculus* mice.

The *M. domesticus – M. musculus* hybrid zone serves as a valuable model of speciation. This study determined the mode of selection acting on *t*-complex genes and highly-introgressing SNP regions, and near-isolation of the *t*-haplotype to *M. domesticus* in the SX transect. Gene flow of neutral and positively-evolving loci combined with the

presence of the *t*-haplotype in only the west side of the hybrid zone is consistent with incomplete genetic isolation between *M. domesticus* and *M. musculus*.

REFERENCES

Ardlie, K.G., and Silver, L.M. (1998). Low Frequency of t Haplotypes in Natural Populations of House Mice (Mus musculus domesticus). Evolution *52*, 1185–1196.

Alibert, P., Renaud, S., Dod, B., Bonhomme, F., and Auffray, J.-C. (1994). Fluctuating Asymmetry in the Mus musculus Hybrid Zone: A Heterotic Effect in Disrupted Co-Adapted Genomes. Proceedings of the Royal Society B *258*, 53–59.

Alibert, P., Fel-Clair, F., Manolakou, K., and Britton-Davidian, J. (1997). Developmental Stability, Fitness, and Trait Size in Laboratory Hybrids Between European Subspecies of the House Mouse. Evolution *51*, 1284–1295.

Barton, N.H., and G.M. Hewitt. (1985). Analysis of hybrid zones. Annu. Rev. Ecol. Syst. *16*, 113-148.

Bauer, H., Willert, J., Koschorz, B., and Herrmann, B.G. (2005). The t complex–encoded GTPase-activating protein Tagap1 acts as a transmission ratio distorter in mice. Nature Genetics *37*, 969–973.

Bauer, H., Veron, N., Willert, J., and Herrmann, B.G. (2007). The t-complex-encoded guanine nucleotide exchange factor Fgd2 reveals that two opposing signaling pathways promote transmission ratio distortion in the mouse. Genes & Development *21*, 143–147.

Bauer, H., Schindler, S., Charron, Y., Willert, J., Kusecek, B., and Herrmann, B.G. (2012). The Nucleoside Diphosphate Kinase Gene Nme3 Acts as Quantitative Trait Locus Promoting Non-Mendelian Inheritance. PLoS Genetics *8*, e1002567.

Bimova, B., Karn, R.C., and Pialek., J. (2004). The role of salivary androgen-binding protein in reproductive isolation between two subspecies of house mouse: *Mus musculus musculus* and *Mus musculus domesticus*. Biol J Linnean Soc. *84*, 439–361.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Stamatoyannopoulos, J.A., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.

Boursot, P., Auffray, J.C., Britton-Davidian, J., and F. Bonhomme. (1993). The Evolution of House Mice. Annu . Rev. Ecol. Syst. *24*, 119–152.

Britton-Davidian, J., Fel-Clair, F., Lopez, J., Alibert, P., and Boursot., P. (2005). Postzygotic isolation between the two European subspecies of the house mouse:

estimates from fertility patterns in wild and laboratory-bred hybrids. Biol J Linnean Soc. *84*, 379–393.

Clarke, G.M. (1993). The genetic basis of developmental stability. I. Relationships between stability, heterozygosity and genomic coadaptation. Genetica *89*, 15–23.

Cucchi, T., Vigne, J-D., and Auffray., J-C. (2005). First occurrence of the house mouse (*Mus musculus domesticus* Schwarz & Schwarz, 1943) in the Western Mediterranean: a zooarchaeological revision of subfossil occurrences. Biol J Linnean Soc. *84*, 429–445.

Dod, B., Jermiin, L.S., Boursot, P., Chapman, V.H., Tonnes-Nielsen, J., Bonhomme, F. (1993). Counterselection on sex chromosomes in the Mus musculus European hybrid zone. J Evol Biol. *6*, 529–546.

Dod, B., Litel, C., Makoundou, P., Orth, A., and Boursot., P. (2003). Identification and characterization of *t* haplotypes in wild mice populations using molecular markers. Genet. Res. *81*, 103–114.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Geraldes, A., Basset, P., Gibson, B., Smith, K.L., Harr, B., Yu, H-T., Bulatova, N., Ziv, Y., and Nachman., M.W. (2008). Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. Mol Ecol. *17*, 5349–5363.

Good, J.M., 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*, 2213–2228.

Harr, B., and Turner., L.M. (2010). Genome-wide analysis of alternative splicing evolution among *Mus* subspecies. Mol Ecol. *19*, 228–239.

Hayashida, K., and Kohno, S. (2009). Hybrid male sterility is caused by mitochondrial DNA deletion. Mol Biol Rep. *36*, 1365–1369.

Herrmann, B., Koschorz, B., Wertz, K., McLaughlin, J., and Kispert., A. (1999). A Protein Kinase Encoded by the t Complex Responder Gene Causes Non-mendelian Inheritance. Nature. *402*, 141–406.

Howard, C.A., Gummere, G.R., Lyon, M.F., Bennett, D., and Artzt, K. (1990). Genetic and molecular analysis of the proximal region of the mouse t-complex using new molecular probes and partial t-haplotypes. Genetics *126*, 1103.

Johnson, L.R., Pilder, S.H., Bailey, J.L., and Olds-Clarke., P. (1995). Sperm from Mice Carrying One or Two *t* Haplotypes Are Different in Investment and Oocyte Penetration. Developmental Biology *168,* 138–149.

Macholan, M., Munclinger, P., Sugerkova, M., Dufkova, P., Bimova, B., Bozikova, E., Zima, J., and Pialek., J. 2007. Genetic Analysis of Autosomal and X-Linked Markers Across a Mouse Hybrid Zone. Evolution *61,* 746–771.

Merrill, C., Bayraktaroglu, L., Kusano, A. and Ganetzky, B. (1999). Truncated RanGAP encoded by the Segregation Distorter locus of Drosophila. Science *283,* 1742–1745.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F. and Marshall, D. 2010. Tablet - next generation sequence assembly visualization. Bioinformatics *26(3),* 401-402.

Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A.G., Hosseini, S., Brandon, M., Easley, K., Chen, E., Brown, M.D., et al. (2003). Natural selection shaped regional mtDNA variation in humans. Proceedings of the National Academy of Sciences *100*, 171–176.

Munclinger, P., Bozikova, E., Sugerkova, M., Pialek, J., and Macholan., M. (2002). Genetic variation in house mice (Mus, Muridae, Rodentia) from the Czech and Slovak Republics. Folia Zool. *51*, 81–92.

Nolte, A.W., Gompert, Z., and Buerkle, C.A. (2009). Variable patterns of introgression in two sculpin hybrid zones suggest that genomic isolation differs among populations. Molecular Ecology *18*, 2615–2627.

Oka, A., Mita, A., Sakurai-Yamatani, N., Yamamoto, H., Takagi, N., Takano-Shimizu, T., Toshimori, K., Moriwaki, K., and Shiroishi., T. (2004). Hybrid Breakdown Caused by Substitution of the X Chromosome Between Two Mouse Subspecies. Genetics *166,* 913–924.

Riesberg, L.H., M.J. Kim and Seiler., G.J. (1999). Introgression between the cultivated sunflower and a sympatric wild relative, Helianthus petiolaris (Asteraceae). Int. J. Plant Sci. *160,* 102-108.

Sage, R.D., Heyneman, D., Lim, K.C., and Wilson., A.C. (1986). Wormy mice in a hybrid zone. Nature *324,* 60–63.

She, J.X., Bonhomme, F., Boursot, P., Thaler, L., and Catzeflis., F. (1990). Molecular phylogenies in the genus Mus: Comparative analysis of electrophoretic, scnDNA hybridization, and mtDNA RFLP data. Biol J Linnean Soc. *41,* 83–103.

Suzuki, H., Shimada, T., Terashima, M., Tsuchiya, K., and Aplin, K. 2004. Temporal, spatial and ecological modes of evolution of Eurasian Mus based on mitochondrial and nuclear gene sequences. Mol Phylogenet Evol. *33*, 626–646.

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 housemouse hybrid zone. Genome Res. 18, 67–76.

Teeter, K.C., Thibodeau, L.M., Gompert, Z., Buerkle, C.A., Nachman, M.W., and Tucker., P.K. 2009. The Variable Genomic Architecture of Isolation Between Hybridizing Species of House Mice. Evolution. *64*, 472–485.

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*, 1146-1163.

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*, 191–197.

Vyskocilová, M., Prazanová, G., & Piálek, J. (2009). Polymorphism in hybrid male sterility in wild-derived Mus musculus musculus strains on proximal chromosome 17. Mammalian genome. *20(2)*, 83-91. doi:10.1007/s00335-008-9164-3

Wang, L., Luzynski, K., Pool, J.E., Janousek, V., Dufkova, P., Vyskocilova, M.M., Teeter, K.C., Nachman, M.W., Munclinger, P., Macholan, M., Pialeks, J., and Tucker., P.K. (2011). Measures of linkage disequilibrium among neighbouring SNPs indicate asymmetries across the house mouse hybrid zone. Mol Ecol. *20*, 2985–3000.

Yang, H., Ding, Y.M., Hutchins, L.N., et al. (2009). A customized and versatile high-density genotyping array for the mouse. Nature Methods. *6*, 663–666.

APPENDIX A

Table 7: *t*-Haplotype Genotyping Data. Results of *t*-haplotype genotyping is shown for each mouse DNA sample.

Sample #	Locality #	Tcp-1	Tcp-Jad	Tcp-Cl	Hba	Consensus
5	13	***+/t***	***+/t***	***+/t***	***+/t***	***+/t***
7	22	***+/t***	***+/t***	***+/t***	***+/t***	***+/t***
8	21	+/+	-	-	+/+	+/+
9	10	***+/t***	***+/t***	***+/t***	***+/t***	***+/t***
10	10	***+/t***	***+/t***	***+/t***	***+/t***	***+/t***
11	10	+/+	-	-	+/+	+/+
12	10	***+/t***	***+/t***	***+/t***	***+/t***	***+/t***
13	10	***+/t***	***+/t***	***+/t***	***+/t***	***+/t***
14	10	***+/t***	***+/t***	***+/t***	***+/t***	***+/t***
15	14	+/+	-	-	+/+	+/+
16	15	***+/t***	***+/t***	***+/t***	***+/t***	***+/t***
17	15	***+/t***	***+/t***	***+/t***	***+/t***	***+/t***
18	15	***+/t***	***+/t***	***+/t***	***+/t***	***+/t***
19	15	+/+	-	-	+/+	+/+
20	15	***+/t***	***+/t***	-	+/+	Partial
21	17	+/+	-	-	+/+	+/+
22	22	+/+	+/+	-	+/+	+/+
23	17	+/+	+/+	-	+/+	+/+
24	25	+/+	-	-	+/+	+/+
25	25	+/+	+/+	-	+/+	+/+
26	25	+/+	+/+	-	+/+	+/+
27	25	+/+	+/+	-	+/+	+/+
28	25	+/+	+/+	+/+	+/+	+/+
29	25	+/+	-	-	+/+	+/+
30	25	+/+	-	-	+/+	+/+
31	25	+/+	-	-	+/+	+/+
32	25	+/+	-	-	+/+	+/+
33	25	+/+	-	-	+/+	+/+
34	25	+/+	-	-	+/+	+/+
35	25	+/+	-	-	+/+	+/+
37	25	+/+	-	-	+/+	+/+
38	25	+/+	-	-	+/+	+/+

Sample #	Locality #	Tcp-1	Tcp-Jad	Tcp-Cl	Hba	Consensus
39	25	+/+	-	-	+/+	+/+
40	25	+/+	-	-	+/+	+/+
41	25	+/+	-	-	+/+	+/+
42	25	+/+	-	-	+/+	+/+
43	25	+/+	-	-	+/+	+/+
44	25	+/+	-	-	+/+	+/+
45	25	+/+	-	-	+/+	+/+
46	25	+/+	+/+	-	+/+	+/+
47	17	+/+	-	-	+/+	+/+
48	17	+/+	-	-	+/+	+/+
49	17	***+/t***	***+/t***	***+/t***	***+/t***	***+/t***
50	17	+/+	-	-	+/+	+/+
51	25	+/+	+/+	-	+/+	+/+
52	25	+/+	+/+	-	+/+	+/+
53	11	+/+	-	-	+/+	+/+
54	11	+/+	-	-	+/+	+/+
56	4	+/+	-	-	+/+	+/+
57	7	***+/t***	***+/t***	-	+/+	Partial
60	2	+/+	-	-	+/+	+/+
61	17	+/+	-	-	+/+	+/+
62	17	+/+	-	-	+/+	+/+
63	17	+/+	-	-	+/+	+/+
64	17	-	+/+	-	+/+	+/+
65	17	-	+/+	-	+/+	+/+
66	17	+/+	-	-	+/+	+/+
67	17	+/+	-	-	+/+	+/+
68	17	+/+	-	-	+/+	+/+
69	17	-	+/+	-	+/+	+/+
70	15	+/+	-	-	+/+	+/+
71	15	-	***+/t***	***+/t***	***+/t***	***+/t***
72	15	+/+	-	-	+/+	+/+
73	15	+/+	-	-	+/+	+/+
74	15	-	***+/t***	***+/t***	***+/t***	***+/t***
75	16	+/+	-	-	+/+	+/+
76	15	+/+	-	-	+/+	+/+
78	9	-	+/+	-	+/+	+/+
80	30	+/+	-	-	+/+	+/+
81	30	+/+	-	-	+/+	+/+
82	30	+/+	-	-	+/+	+/+
83	30	+/+	-	-	+/+	+/+

Sample #	Locality #	Tcp-1	Tcp-Jad	Tcp-Cl	Hba	Consensus
84	30	+/+	-	-	+/+	+/+
85	30	+/+	-	-	+/+	+/+
86	27	+/+	-	-	+/+	+/+
88	11	+/+	-	-	+/+	+/+
89	11	+/+	-	-	+/+	+/+
90	12	-	+/+	-	+/+	+/+
91	12	-	***+/t***	***+/t***	***+/t***	***+/t***
92	15	+/+	-	-	+/+	+/+
93	15	-	***+/t***	***+/t***	***+/t***	***+/t***
94	15	-	+/+	-	+/+	+/+
95	15	-	+/+	-	+/+	+/+
96	16	-	***+/t***	-	***+/t***	***+/t***
97	12	+/+	-	-	+/+	+/+
98	12	+/+	-	-	+/+	+/+
101	12	+/+	-	-	+/+	+/+
102	12	+/+	-	-	+/+	+/+
103	12	+/+	-	-	+/+	+/+
104	12	+/+	-	-	+/+	+/+
105	12	+/+	-	-	+/+	+/+
106	12	-	+/+	-	+/+	+/+
107	3	+/+	-	-	+/+	+/+
108	3	-	+/+	-	*+/t*	+/+
109	3	***+/t***	***+/t***	***+/t***	***+/t***	***+/t***
110	3	+/+	-	-	+/+	+/+
111	16	+/+	-	-	+/+	+/+
112	13	***+/t***	***+/t***	+/+	***+/t***	***+/t***
113	13	+/+	-	-	+/+	+/+
114	13	-	***+/t***	***+/t***	***+/t***	***+/t***
115	13	+/+	-	-	+/+	+/+
116	13	-	***+/t***	***+/t***	***+/t***	***+/t***
118	16	-	***+/t***	***+/t***	***+/t***	***+/t***
119	19	-	+/+	-	+/+	+/+
120	17	+/+	-	-	+/+	+/+
121	17	-	+/+	-	+/+	+/+
122	17	+/+	-	-	+/+	+/+
123	17	-	+/+	+/+	+/+	+/+
124	17	+/+	-	-	+/+	+/+
125	4	+/+	-	-	+/+	+/+
126	12	+/+	-	-	+/+	+/+
127	12	-	***+/t***	***+/t***	***+/t***	***+/t***

Sample #	Locality #	Tcp-1	Tcp-Jad	Tcp-Cl	Hba	Consensus
128	19	-	+/+	-	+/+	+/+
129	19	+/+	-	-	+/+	+/+
130	19	+/+	-	-	+/+	+/+
131	19	-	+/+	-	+/+	+/+
132	19	+/+	-	-	+/+	+/+
133	31	-	+/+	-	+/+	+/+
134	26	-	+/+	-	+/+	+/+
135	32	-	+/+	-	+/+	+/+
136	32	-	+/+	-	+/+	+/+
137	32	-	+/+	-	+/+	+/+
138	28	***+/t***	***+/t***	***+/t***	+/+	Partial
139	17	+/+	-	-	+/+	+/+
140	17	+/+	-	-	+/+	+/+
141	17	-	+/+	-	+/+	+/+
142	17	+/+	-	-	+/+	+/+
143	17	-	+/+	-	+/+	+/+
144	15	+/+	-	-	+/+	+/+
145	15	+/+	-	-	+/+	+/+
146	15	+/+	-	-	+/+	+/+
147	15	-	***+/t***	***+/t***	***+/t***	***+/t***
148	15	+/+	-	-	+/+	+/+
149	15	+/+	-	-	+/+	+/+
150	15	+/+	-	-	+/+	+/+
151	15	+/+	-	-	+/+	+/+
152	15	+/+	-	-	+/+	+/+
153	17	+/+	-	-	+/+	+/+
154	17	+/+	-	-	+/+	+/+
155	17	+/+	-	-	+/+	+/+
156	17	+/+	-	-	+/+	+/+
157	17	+/+	-	-	+/+	+/+
158	17	+/+	-	-	+/+	+/+
159	17	+/+	-	-	+/+	+/+
160	18	+/+	-	-	+/+	+/+
161	18	+/+	-	-	+/+	+/+
162	18	+/+	-	-	+/+	+/+
163	18	-	+/+	-	+/+	+/+
164	18	+/+	-	-	+/+	+/+
165	18	+/+	-	-	+/+	+/+
166	18	+/+	-	-	+/+	+/+
167	18	+/+	-	-	+/+	+/+

Sample #	Locality #	Tcp-1	Tcp-Jad	Tcp-Cl	Hba	Consensus
168	18	-	***+/t***	***+/t***	***+/t***	***+/t***
169	18	+/+	-	-	+/+	+/+
170	18	-	***+/t***	***+/t***	***+/t***	***+/t***
171	18	+/+	-	-	+/+	+/+
172	18	+/+	-	-	+/+	+/+
173	16	+/+	-	-	+/+	+/+
174	3	-	***+/t***	***+/t***	***+/t***	***+/t***
175	5	-	+/+	-	+/+	+/+
176	15	+/+	-	-	+/+	+/+
177	15	+/+	-	-	+/+	+/+
178	15	***+/t***	***+/t***	***+/t***	***+/t***	***+/t***
179	15	+/+	-	-	+/+	+/+
180	15	+/+	-	-	+/+	+/+
181	15	+/+	-	-	+/+	+/+
182	15	+/+	-	-	+/+	+/+
183	15	+/+	-	-	+/+	+/+
184	17	-	+/+	+/+	+/+	+/+
185	17	+/+	-	-	+/+	+/+
186	13	+/+	-	-	+/+	+/+
187	13	-	+/+	-	+/+	+/+
188	13	+/+	-	-	+/+	+/+
189	13	+/+	-	-	+/+	+/+
190	17	+/+	-	-	+/+	+/+
191	8	+/+	-	-	+/+	+/+
192	8	+/+	-	-	+/+	+/+
193	8	+/+	-	-	+/+	+/+
194	8	+/+	-	-	+/+	+/+
195	35	+/+	-	-	+/+	+/+
196	34	-	+/+	-	+/+	+/+
197	34	+/+	-	-	+/+	+/+
198	33	+/+	-	-	+/+	+/+
199	33	+/+	-	-	+/+	+/+
200	33	-	+/+	-	+/+	+/+
201	34	+/+	+/+	+/+	+/+	+/+
202	34	+/+	-	-	+/+	+/+
207	35	-	+/+	-	+/+	+/+
208	35	-	+/+	-	+/+	+/+
211	1	+/+	-	-	+/+	+/+
212	1	+/+	-	-	+/+	+/+
213	1	+/+	-	-	+/+	+/+

Sample #	Locality #	Tcp-1	Tcp-Jad	Tcp-Cl	Hba	Consensus
214	1	+/+	-	-	+/+	+/+
215	1	+/+	-	-	+/+	+/+
216	1	+/+	-	-	+/+	+/+
217	1	+/+	-	-	+/+	+/+
218	1	+/+	-	-	+/+	+/+
219	1	+/+	-	-	+/+	+/+
220	1	+/+	-	-	+/+	+/+
221	1	+/+	-	-	+/+	+/+
222	1	+/+	-	-	+/+	+/+
223	1	-	+/+	-	+/+	+/+
224	1	+/+	-	-	+/+	+/+
225	1	+/+	-	-	+/+	+/+
226	1	-	+/+	-	+/+	+/+
227	1	-	+/+	-	+/+	+/+
228	1	-	+/+	-	+/+	+/+
229	1	+/+	-	-	+/+	+/+
230	1	-	+/+	-	+/+	+/+
231	1	+/+	-	-	+/+	+/+
232	1	+/+	-	-	+/+	+/+
233	1	-	+/+	-	+/+	+/+
234	1	+/+	-	-	+/+	+/+
235	1	-	+/+	-	+/+	+/+
236	1	-	+/+	-	+/+	+/+
237	1	-	+/+	-	+/+	+/+
238	1	-	+/+	-	+/+	+/+
239	1	-	+/+	-	+/+	+/+
240	1	+/+	-	-	+/+	+/+
241	1	+/+	-	-	+/+	+/+
242	1	+/+	-	-	+/+	+/+
243	1	+/+	-	-	+/+	+/+
244	1	+/+	-	-	+/+	+/+
245	1	+/+	-	-	+/+	+/+
246	4	+/+	-	-	+/+	+/+
247	9	-	+/+	-	+/+	+/+
248	6	-	+/+	-	+/+	+/+
249	17	-	+/+	-	+/+	+/+
250	17	-	+/+	-	+/+	+/+
251	17	-	***+/t***	***+/t***	***+/t***	***+/t***
252	17	-	***+/t***	***+/t***	***+/t***	***+/t***
253	24	-	+/+	-	+/+	+/+

Sample #	Locality #	Tcp-1	Tcp-Jad	Tcp-Cl	Hba	Consensus
254	24	-	+/+	-	+/+	+/+
255	24	+/+	-	-	+/+	+/+
256	24	-	+/+	-	+/+	+/+
257	24	+/+	-	-	+/+	+/+
258	24	-	+/+	-	+/+	+/+
259	24	-	+/+	-	+/+	+/+
260	24	+/+	+/+	-	+/+	+/+
261	20	-	***+/t***	-	***+/t***	***+/t***
262	17	-	+/+	-	+/+	+/+
263	23	-	+/+	-	+/+	+/+
264	23	-	+/+	-	+/+	+/+
265	23	-	+/+	-	+/+	+/+
266	23	-	***+/t***	***+/t***	***+/t***	***+/t***
267	23	+/+	-	-	+/+	+/+
268	23	+/+	-	-	+/+	+/+
269	23	+/+	-	-	+/+	+/+
270	23	+/+	-	-	+/+	+/+
271	15	-	***+/t***	***+/t***	***+/t***	***+/t***
272	15	***+/t***	***+/t***	***+/t***	***+/t***	***+/t***
273	15	-	+/+	-	+/+	+/+
274	15	-	***+/t***	***+/t***	***+/t***	***+/t***
275	24	-	+/+	-	+/+	+/+
276	24	+/+	-	-	+/+	+/+
277	24	+/+	-	-	+/+	+/+
278	24	+/+	-	-	+/+	+/+
279	24	+/+	-	-	+/+	+/+
280	24	***+/t***	***+/t***	***+/t***	+/+	Partial
281	24	+/+	-	-	+/+	+/+
282	24	+/+	-	-	+/+	+/+
283	24	+/+	-	-	+/+	+/+
284	17	-	***+/t***	***+/t***	***+/t***	***+/t***
285	17	-	+/+	-	+/+	+/+
286	17	+/+	-	-	+/+	+/+
287	18	-	***+/t***	***+/t***	***+/t***	***+/t***
288	18	-	***+/t***	***+/t***	***+/t***	***+/t***
289	18	+/+	-	-	+/+	+/+
290	18	-	***+/t***	***+/t***	***+/t***	***+/t***
291	18	+/+	-	-	+/+	+/+
292	18	-	***+/t***	***+/t***	***+/t***	***+/t***
293	18	+/+	-	-	+/+	+/+

Sample #	Locality #	Tcp-1	Tcp-Jad	Tcp-Cl	Hba	Consensus
294	18	-	+/+	-	+/+	+/+
295	18	-	***+/t***	-	***+/t***	***+/t***
296	18	+/+	-	-	+/+	+/+
297	18	+/+	-	-	+/+	+/+
298	18	-	***+/t***	***+/t***	***+/t***	***+/t***
299	18	+/+	+/+	+/+	+/+	+/+
300	18	+/+	+/+	-	+/+	+/+
301	18	+/+	+/+	-	+/+	+/+
302	18	+/+	+/+	-	+/+	+/+
303	18	+/+	+/+	+/+	+/+	+/+
304	18	+/+	-	-	+/+	+/+
305	18	-	+/+	-	+/+	+/+
306	18	+/+	+/+	-	+/+	+/+
307	18	-	***+/t***	***+/t***	***+/t***	***+/t***
308	18	-	+/+	+/+	+/+	+/+
309	18	+/+	+/+	-	+/+	+/+
310	18	-	***+/t***	***+/t***	***+/t***	***+/t***
311	18	-	***+/t***	***+/t***	***+/t***	***+/t***
312	18	-	***+/t***	***+/t***	***+/t***	***+/t***
313	18	-	***+/t***	***+/t***	***+/t***	***+/t***
314	18	+/+	-	-	+/+	+/+
315	18	-	***+/t***	***+/t***	***+/t***	***+/t***
316	18	-	+/+	-	+/+	+/+
317	18	-	+/+	-	+/+	+/+
318	18	-	***+/t***	***+/t***	***+/t***	***+/t***
319	18	+/+	-	-	+/+	+/+
320	18	-	***+/t***	***+/t***	***+/t***	***+/t***
329	18	-	+/+	+/+	+/+	+/+
330	18	-	***+/t***	***+/t***	***+/t***	***+/t***
331	18	-	+/+	-	+/+	+/+
332	18	-	+/+	-	+/+	+/+

APPENDIX B

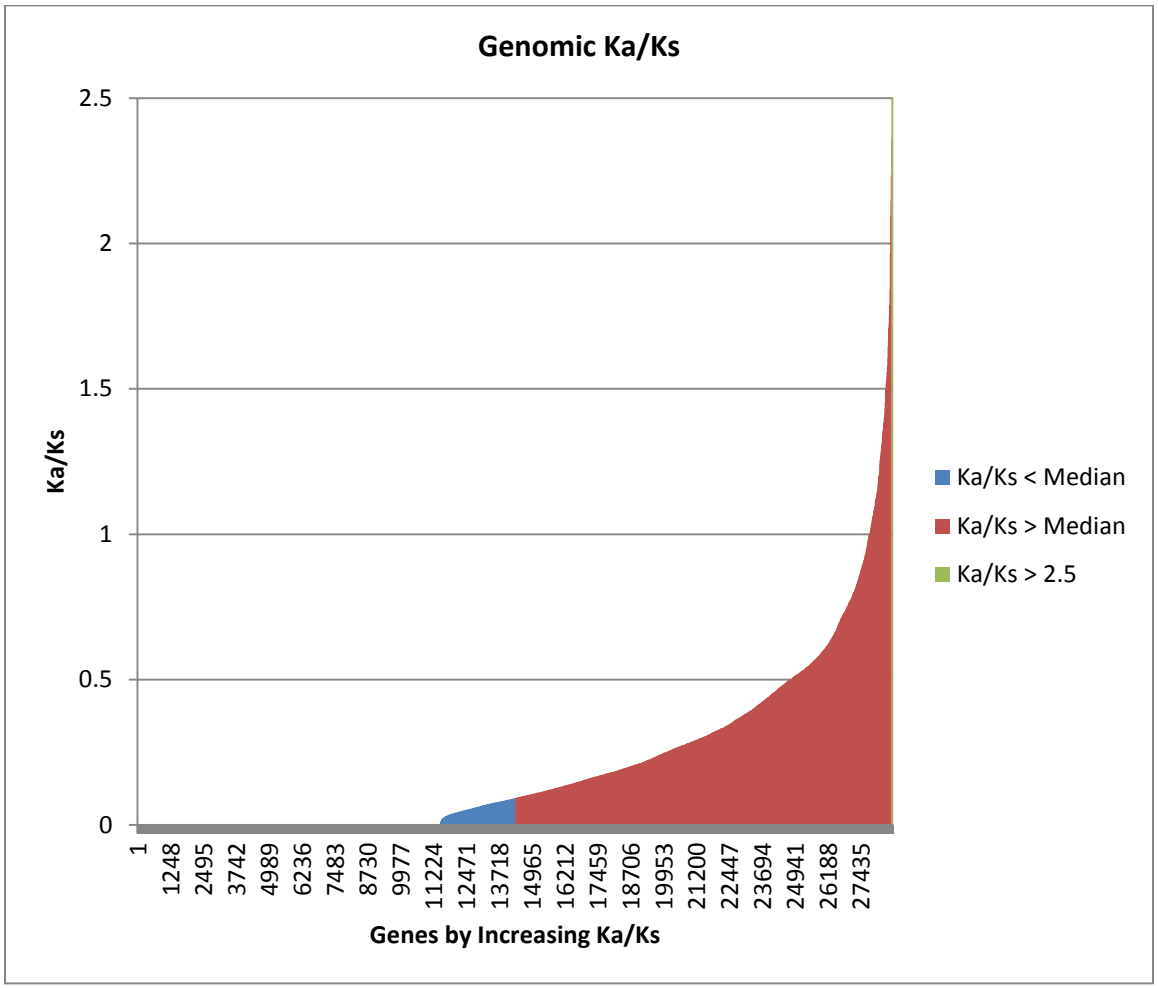

Figure 8A: Genome-wide Ka/Ks. Genome-wide values of Ka and Ks between *M. domesticus* and *M. musculus* were provided by Václav Janoušek. Ka/Ks ratios for genes across the genome are shown sorted lowest to highest. Median Ka/Ks is 0.092784.

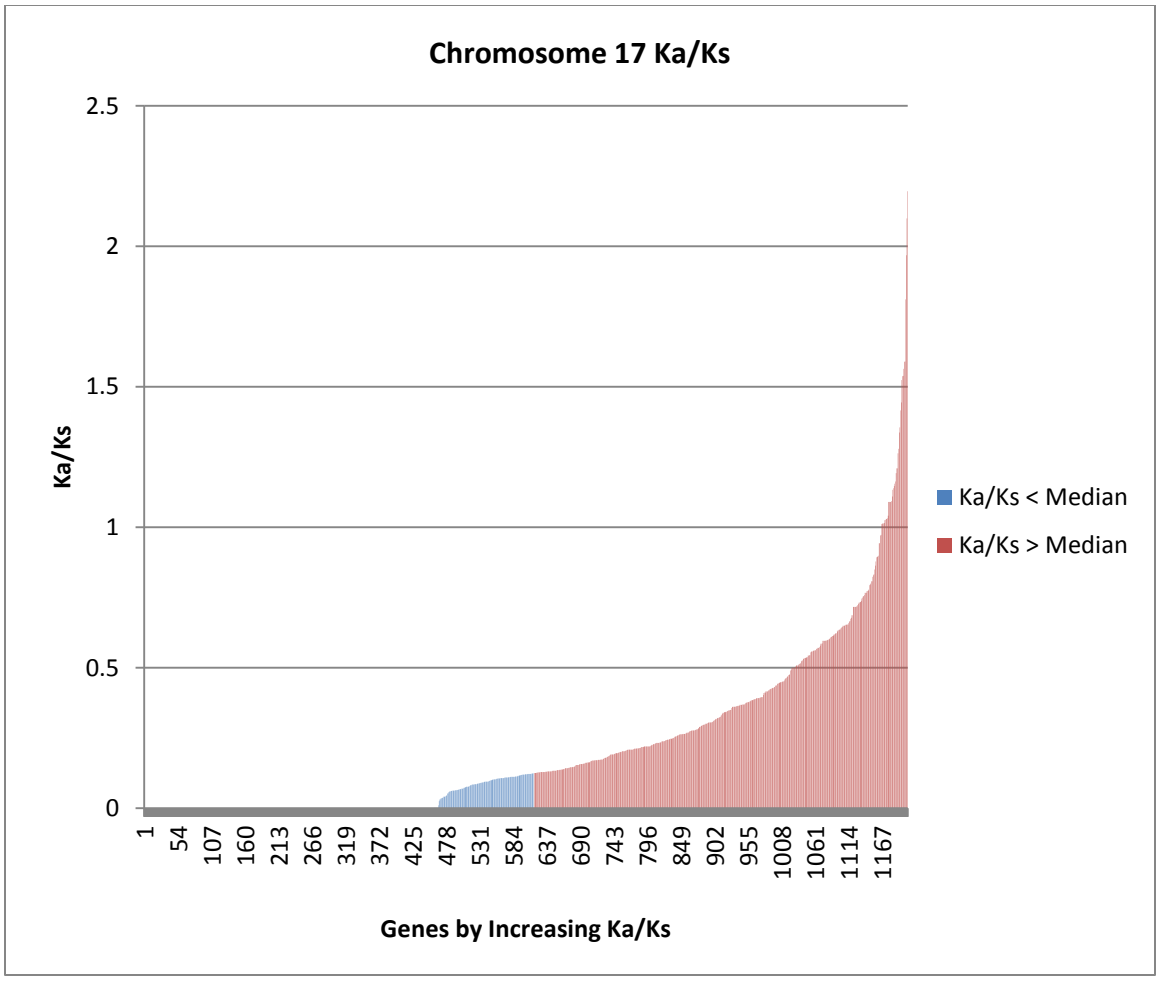

Figure 8B: Chromosome 17 Ka/Ks. Genome-wide values of Ka and Ks between *M. domesticus* and *M. musculus* were provided by Václav Janoušek. Ka/Ks ratios for genes on chromosome 17 are shown sorted lowest to highest. Median Ka/Ks is 0.120773.

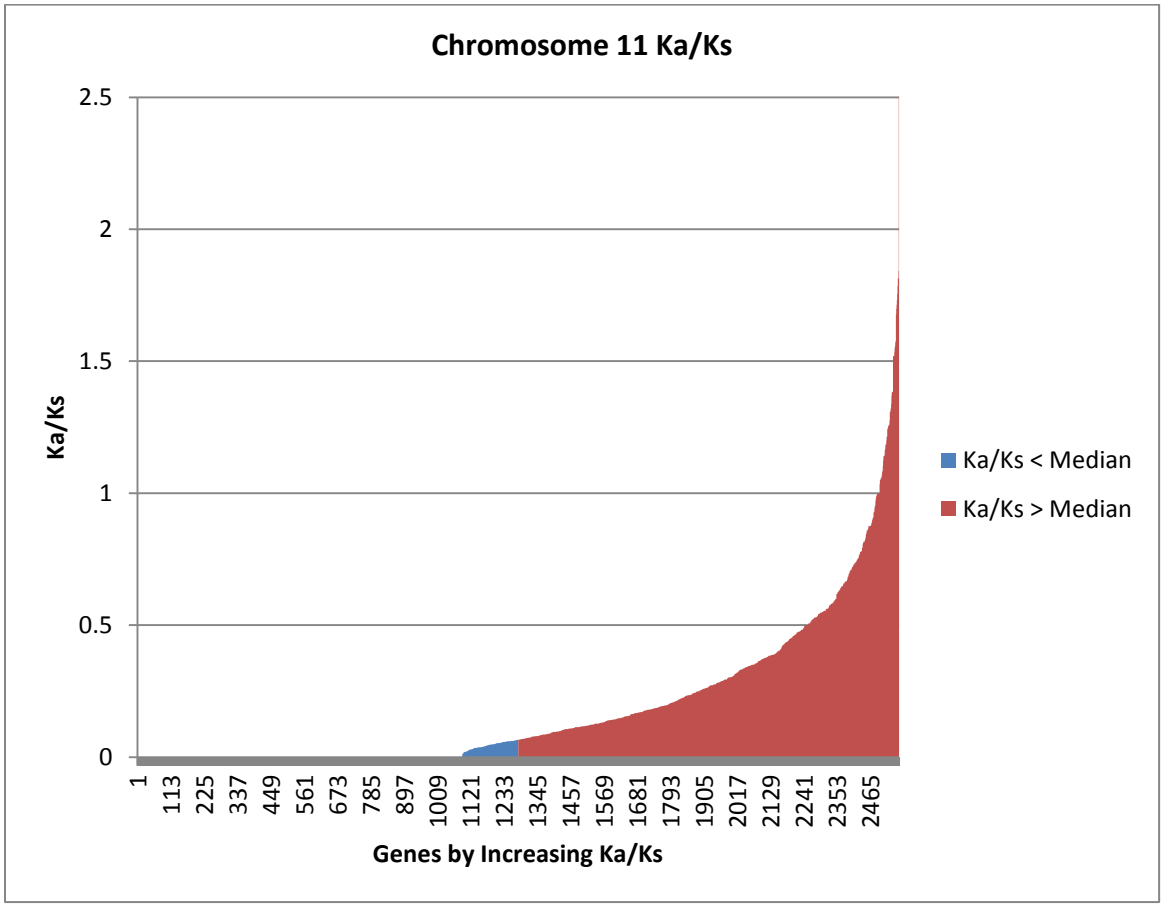

Figure 8C: Chromosome 11 Ka/Ks. Genome-wide values of Ka and Ks between *M. domesticus* and *M. musculus* were provided by Václav Janoušek. Ka/Ks ratios for genes on chromosome 11 are shown sorted lowest to highest. Median Ka/Ks is 0.065924.

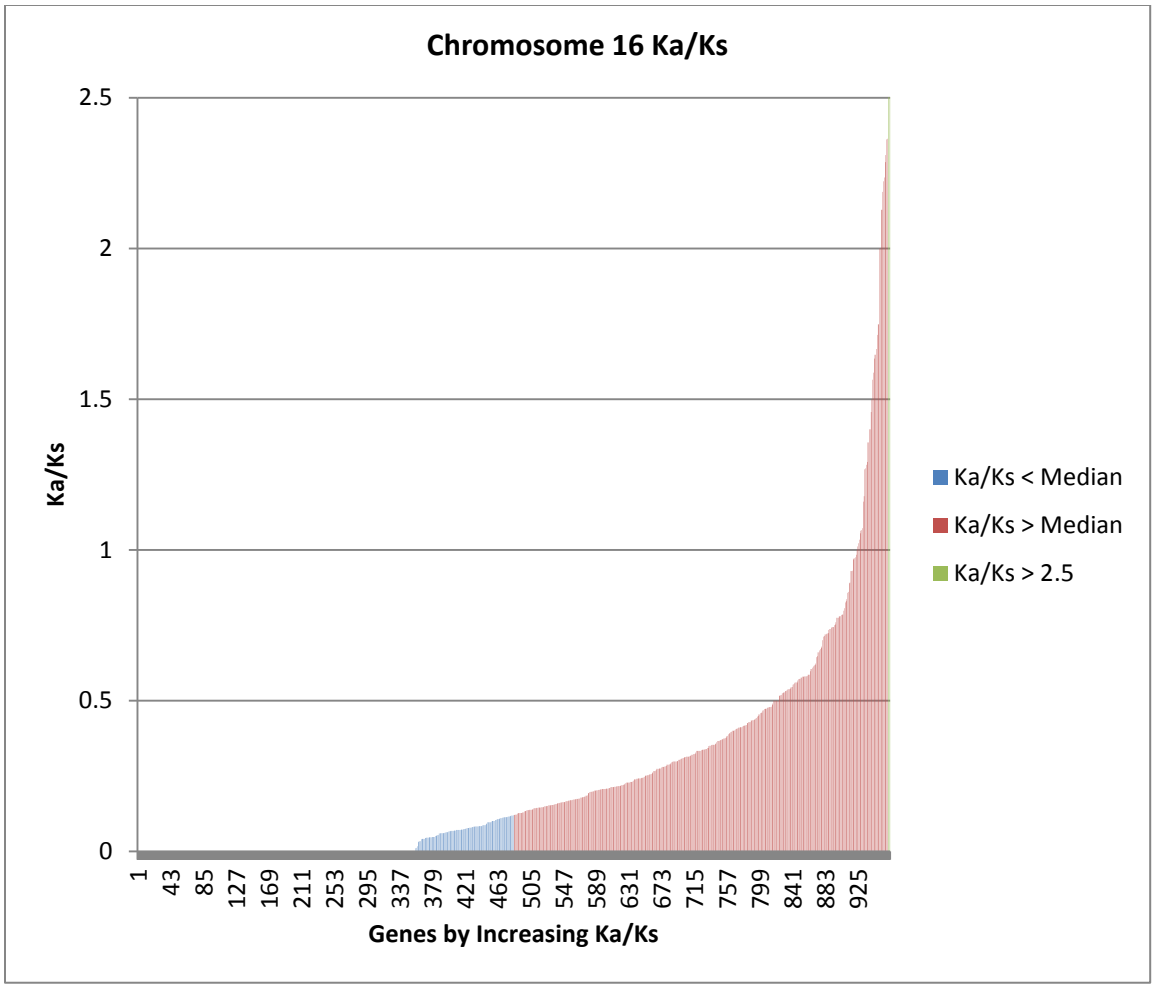

Figure 8D: Chromosome 16 Ka/Ks. Genome-wide values of Ka and Ks between *M. domesticus* and *M. musculus* were provided by Václav Janoušek. Ka/Ks ratios for genes on chromosome 16 are shown sorted lowest to highest. Median Ka/Ks is 0.120299.

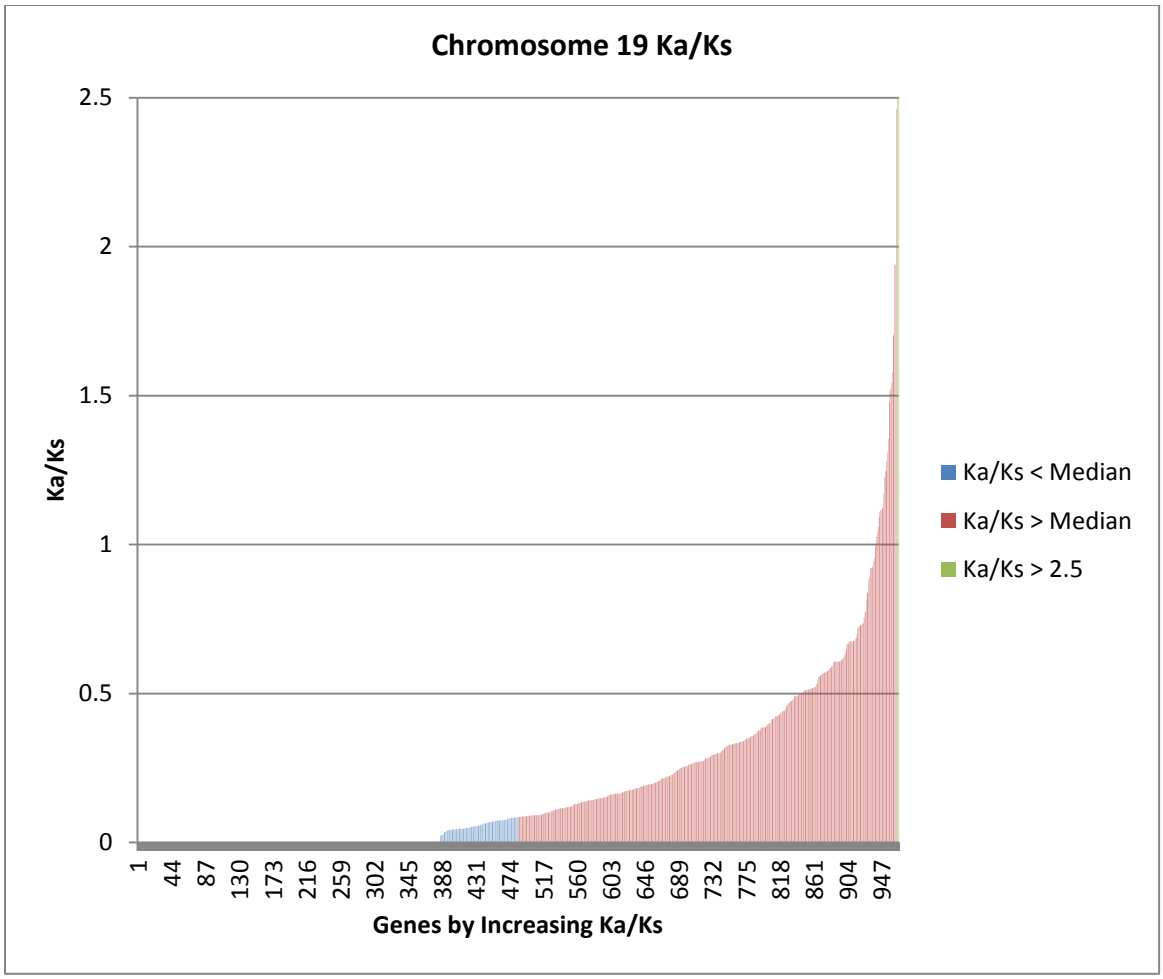

Figure 8E: Chromosome 19 Ka/Ks. Genome-wide values of Ka and Ks between *M. domesticus* and *M. musculus* were provided by Václav Janoušek. Ka/Ks ratios for genes on chromosome 19 are shown sorted lowest to highest. Median Ka/Ks is 0.084848.