Molecular Studies of House Mice in Southern New Zealand

A thesis
submitted in partial fulfilment
of the requirements for the degree
of
Master of Science in Biological Sciences
at
The University of Waikato
By

HELEN M MCCORMICK

The University of Waikato 2011

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

Abstract

The house mouse, *Mus Musculus*, was first introduced into New Zealand in significant numbers in the mid nineteenth century. Earlier research suggests that multiple introductions of the three subspecies of house mouse *M. m. domesticus, M. m. musculus and M. m. castaneus* occurred at this time. While *M. m. domesticus* is widely distributed in New Zealand with *M. m. musculus* somewhat less so, the only extant evidence found to date for *M. m. castaneus* is in hybrid mice found principally in the southern half of the South Island.
In this study, the hybrid population of mice in the South Island was examined in detail using a variety of molecular techniques. The extent of the hybrid zone was first established using a simple polymerase chain reaction (PCR) technique that enabled rapid identification of mitochondrial genomes as being *M. m. domesticus, castaneus* or *musculus* in nature. In the second part of the study, four representative mice from the hybrid zone and three *M. m. domesticus* from north of this zone were subjected to high density (600K) single nucleotide polymorphism (SNP) analyses in order to pinpoint regions of the genomes that differed in a significant manner between the two groups. In a refinement of these analyses, regions of each genome were identified as being *domesticus*-, *castaneus*- or *musculus*-like, using diagnostic SNP alleles for each sub-species.

Some 170 mice, principally collected in the Southern South Island, were screened using the rapid PCR technique. In the coastal regions, all mice further south than 44^{o}S had *M. m. castaneus* mtDNA whereas those north of 43^{o}S had *M. m. domesticus* mtDNA. Between the two, there is a 'contact zone' in which both subspecies were found, sometimes in the same building. The contact zone extends approximately 50km north to south and some 30km

inland. Classical tests with three nuclear DNA markers confirmed earlier work, namely, that the nuclear genomes of all mice appeared to be predominantly *domesticus*-like.
Although there is no obvious geographical, ecological or land-use features that characterise this contact zone it may be relevant that there is the distinct change to a wetter colder climate in the south and inland of the contact zone, especially in winter. The inability to identify obvious defining features is not surprising, given that much research over many years worldwide has failed to yield an explanation for distinct mouse hybrid zones.

The high density SNP analyses demonstrated that the nuclear genomes of all seven representative mice were very similar and largely (approximately 95%+) *domesticus*-like in nature. Despite this similarity, clusters of SNPs did reveal differences throughout the genome, often just extending over a few haplotype blocks and often encompassing much less than a million basepairs (Mb). Some of these were of biological interest, for example clusters of vomeronasal receptor genes and also genes believed to be involved in maternal-fetal conflict, which are known to vary markedly between species.

The diagnostic SNP analyses confirmed that no marked differences exist between the genomes of the hybrid and *M. m. domesticus* mice. As stated above, the genomes were predominantly *domesticus*-like but small regions (~1Mb) of *castaneus*-like and *musculus*-like genome were found scattered throughout all mice but, contrary to initial expectations, there was no preponderance of *castaneus*-like regions in the hybrid mice. The one exception to the largely identical genomic SNP patterns, were those from the mouse collected in Hamilton. This was quite bizarre, in that its genome contains large (up to 10Mb) *musculus*-like regions that correspond exactly to the same regions in a common laboratory strain of mouse, C57BL/6J.

I conclude that if pure *M. m. castaneus* mice did originally reach New Zealand, extensive 'backcrossing' with *M. m. domesticus* has virtually eliminated the *castaneus* genome, with just a few remnants remaining that may or may not confer some selective advantage, but could just as likely represent segregation of recognition factors that give rise to assortative mating. Thus, the most obvious and consistent genetic difference between the mice remains the original observation concerning their mitochondrial genomes and these should be examined in more detail.

In future research aimed at identifying potential selective advantages that have allowed the hybrid mice to exclusively populate the southern South Island factors that relate to climate should be considered. Specifically, the fact that, in humans, some mitochondrial haplotypes are believed to confer a selective advantage in colder climates immediately suggests that examining variants of key genes involved in energy metabolism such as NADH-dehydrogenase-subunit-3 (*ND*3) in the hybrid mice could be a profitable line of future research.

Acknowledgements

I would like to thank Ray Cursons and Carolyn King for their advice during my MSc research and for their assistance in preparing the final version of this thesis, as well as for being great teachers during my undergraduate studies.

Thank you to Tanya Chubb for allowing me access to her mouse samples, to Dr. Craig Herbold for several hours donated in an effort to teach me how to use 'R' and to Nick Demetras for showing me how to assemble and analyse the sequences in Geneious.

Many thanks to those people in Canterbury and Otago who collected mouse samples.

Thanks to Fernando Pardo-Manuel de Villena for access to diagnostic SNP data and other insights over the last year or so.

I received the University of Waikato Masters Research Scholarship and the LIC Patrick Shannon Masterate scholarship, and I thank these groups for their assistance.

A huge thank you to all the staff and students in the Molecular Genetics and Microbes and Proteins laboratories for making the lab such a fun and interesting place to work and for answering my numerous queries over the last few years.

Much love and thanks to my good friend and office-mate Tracey Jones for many things including offering distraction from the stressors of thesis write-up, for propping me up when required (often) and for engaging in amusing office antics with me. I will miss you and Lexi when I am several thousands of miles away putting myself through further post-graduate punishment!

Thank you so much to my friends and family for supporting me throughout undergraduate and graduate studies and being patient as I spent many years deciding what to do with my life. In particular thanks to mum for proof-reading the draft and to Justin for being there for me. I promise I won't be a student forever people, just a few more years if you don't mind.

Most importantly, to Dick Wilkins, thank you so much for teaching me so thoroughly the ways of the lab and for instilling in me the confidence to go out into the world of molecular biology. Many thanks for writing the macros that allowed me to display the SNP array data. Thank you also for being a great friend and source of much hilarity.

Ethics statement

Animal ethics approval was not necessary as all animals were trapped or poisoned during normal household and farm pest eradication by people local to the collection sites. Training was undertaken in euthanizing mice as per Standard Operating Procedure No. 9 of the University of Waikato Animal Ethics Committee, but this procedure was not used.

Table of Contents

List of Tables

List of Figures

Chapter One: Introduction

1.1 General Introduction

A central question in biology is how new species evolve. In the case of mammals, one of the most profitable avenues of research involves mice. Classic inbred laboratory strains have been extremely useful for genetic and medical research but it is now recognised that they have a significant disadvantage, namely, lack of genetic polymorphism due to derivation from crosses between a handful of founders. Thus, although the pedigrees of classic laboratory strains are well documented (Figure 1.1) they have resulted in strains with genetic makeups so unlike their wild counterparts that some suggest they should be named '*mus laboratorius*' (Guénet & Bonhomme, 2003; Pocock, Hauffe, & Searle, 2005).

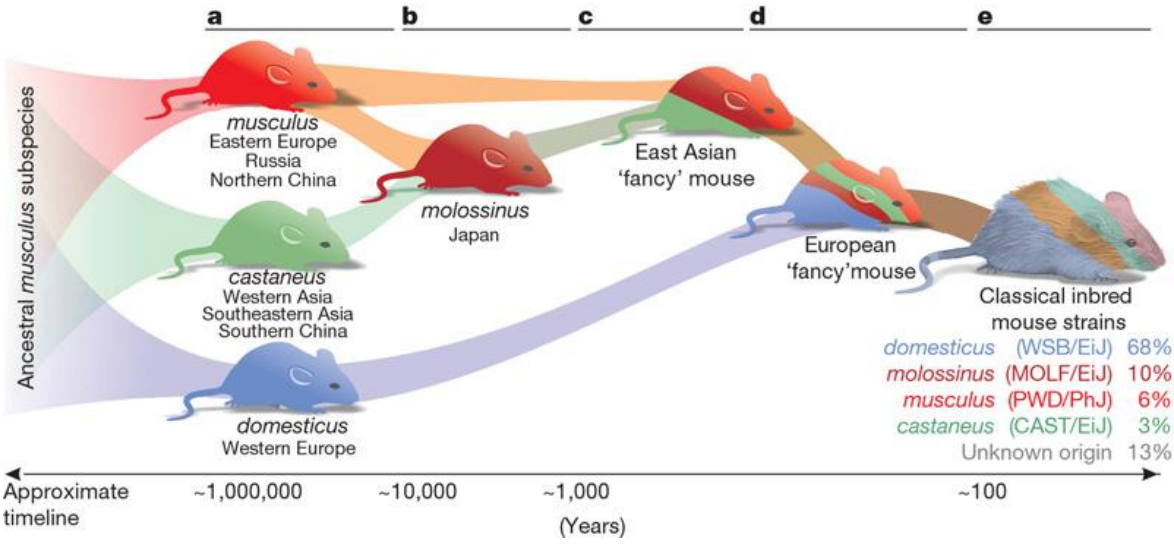


Figure 1.1: Genetic makeup of the laboratory mouse. (Frazer *et al*., 2007)

The commensal mouse has one of the most extensive world distributions of all mammals, probably second only to humans. Originating in the north of the Indian subcontinent around

0.5 million years ago, (Boursot, Auffray, Britton-Davidian, & Bonhomme, 1993; Din *et al*., 1996) they are now found on all six continents in a variety of forms. The *Mus musculus* species complex is comprised of three main varieties of commensal mice (described most often as subspecies): *Mus musculus domesticus* from Western Europe, *Mus musculus musculus* from Eastern Europe and *Mus musculus castaneus* from South East Asia. *M. m. domesticus* is the most widespread of these subspecies, having colonised every continent via European shipping in the last 500 years or so. In New Zealand, *M. m. domesticus*, *M. m. musculus* and *M. m. castaneus* subspecies are present (Searle, Jamieson *et al*., 2009). Throughout the world, different subspecies occupy distinct geographic regions and contact at hybrid zones where genetic information is shared. In some cases this has led to the emergence of distinct new sub-species, such as the Japanese house mouse *M. m. molossinus* (Terashima *et al*., 2006). *M. m. domesticus* and *M. m. musculus* hybrid zones have been extensively studied, especially in middle Europe where a narrow and seemingly very stable hybrid zone has become established (Boursot *et al*., 1993). This has provided many insights into evolutionary mechanisms, especially in the area of speciation. Hybrid zones between *M. m. domestius* and *M. m. castaneus*, the subject of this study, have not been comprehensively investigated. Possible hybrid zones have been identified at Lake Casitas in California (Orth, Adama, Din, & Bonhomme, 1998) as well as in Indonesia (Terashima *et al*., 2006) and Iran (Gunduz *et al*., 2000), but data is sketchy. Searle *et al*. (2009) also discovered such hybrid mice as part of detailed mtDNA haplotype studies of house mice in New Zealand and some of the surrounding islands. While highlighting the nature of the distribution of the subspecies (as well as origins) they did not demark the boundary of a hybrid zone.

1.2 The Middle Europe Hybrid Zone

Many theories have been proposed to explain the phenomenon of hybrid zones, in particular, that existing in middle Europe (Figure 1.2) which is remarkably narrow and appears to have remained stable over many hundreds of years. (Boursot *et al*., 1993; Raufaste *et al*., 2005).

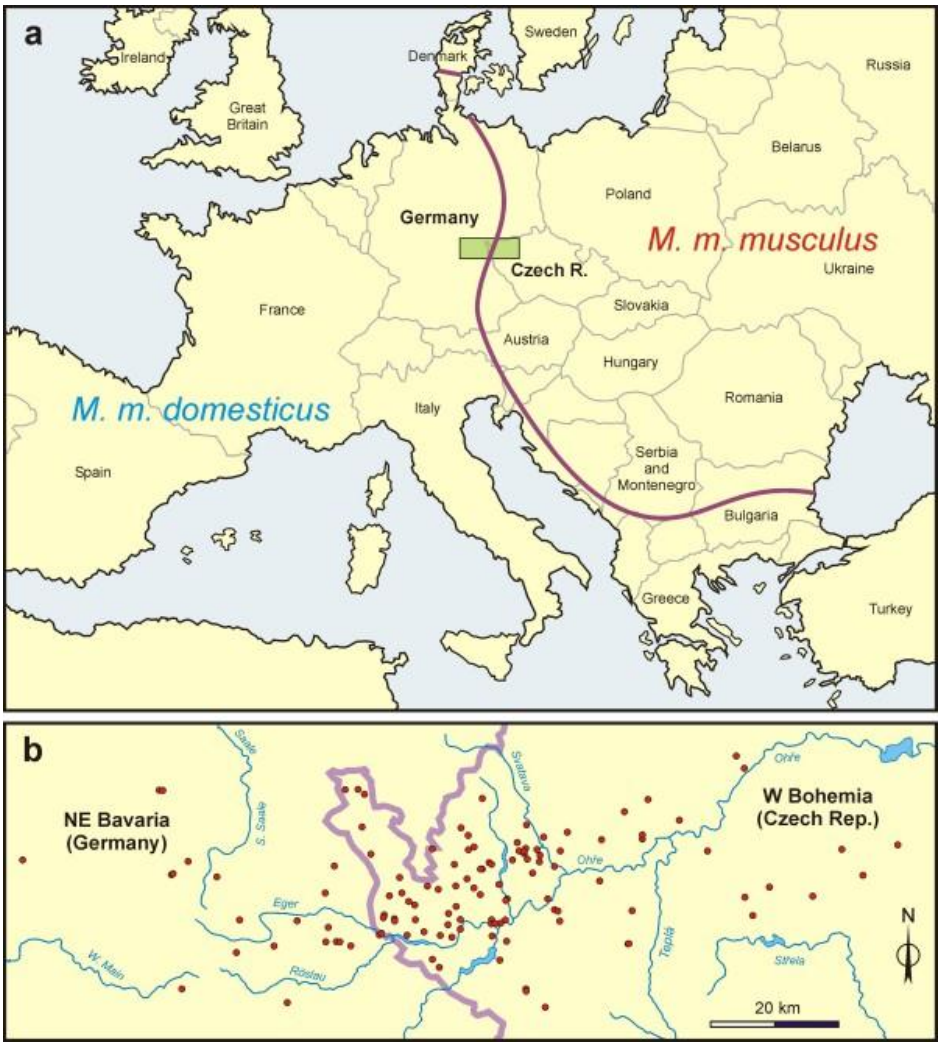


Figure 1.2 : The *M. m musculus* / *M. m. domesticus* hybrid zone in Europe (Macholan *et al*., 2008)

The most obvious explanation for the persistence of a stable hybrid zone is 'reproductive isolation' but this is a somewhat ill-defined phrase. Pheromone signatures differ greatly between different families of mice and are thought to constitute a very powerful reproductive isolating mechanism – even within one subspecies, and even within small geographic localities and between demes (a local population of interbreeding organisms that share a similar genotype) (Mudge *et al*., 2008). Even more pronounced differences in pheromone signatures exist between subspecies, for example in vomeronasal genes (Del Punta, Rothman, Rodriguez, & Mombaerts, 2000; Teeter *et al*., 2008) and MUP (major urinary protein) profiles (Mudge *et al*., 2008), and these presumably contribute to reproductive isolation. Nevertheless, hybridisation does occur. Thus, early molecular studies using nuclear markers that differentiated between *M. m. musculus* and *M. m. domesticus* demonstrated that hybrid mice exist in a narrow hybrid region in middle Europe. In most places, this hybrid zone is only 10 to 20 km wide. These early studies used a limited number of nuclear markers (Boursot *et al*., 1993). More recent studies have both extended the geographic regions studied and also the number of nuclear and mitochondrial markers (Macholán *et al*., 2011; Prager *et al*., 1993; Teeter *et al*., 2008). In general these support the earlier work, but various markers appear to have introgressed across the contact zone to varying extents, i.e. *M. m. musculus* markers to the west and *M. m. domesticus* markers to the east (BoŽÍKovÁ et al., 2005; Macholán et al., ; Munclinger et al., 2002; Teeter et al., 2008).

None of the above actually offers an explanation as to why the hybrid zone occurs where it does geographically and why it is so narrow. Although one might expect some geographic feature such as a river, mountainous territory or climate demarcations to define zone boundaries, none has been identified. Indeed, mouse hybrid zones can be 'mottled' and very small. For example, in Northern Italy, some habitats of two Robertsonian chromosomal races

of *M. m. musculus* are separated by a kilometre or less, with few hybrid mice detected in the middle (Panithanarak *et al*., 2004). In fact, 'contact zone' is a more accurate description than hybrid zone.

In general, it would appear that, historically, the various subspecies of mice have independently spread and colonised areas until they have met, somewhat akin to two armies meeting and establishing an uneasy truce, the boundary between the two being more determined by this 'history' rather than any obvious geographic or ecologic feature. This view is reinforced by the work of Prager, Sage *et al.* (1993) and more recently Jones *et al.* (2011; 2010) who have shown that in more recent times both subspecies have colonised Scandinavia and established hybrid zones therein.

Just why the hybrid zone *per se* is so narrow, and why in some regions, so few hybrid mice are found, is not clear. A possible explanation was proffered some years ago by Sage *et al*. (1986) who examined the parasitic loads of mice in southern Germany. The number of parasitic worms (pinworms and tapeworms) of hybrid mice was more than 10 times that of both *M. m. domesticus* and *M. m. musculus* on either side of the zone. Clearly, this would compromise the relative fitness of hybrid mice, and more recent investigations (Derothe, Loubes, Perriat-Sanguinet, Orth, & Moulia, 1999; Moulia *et al*., 1991) demonstrated that hybrids in the Danish zone also carry higher parasitic worm loads than those of the parental species. Derothe *et al* (1999) used experimental infection in *M. m. musculus*, *M. m. domesticus* and their hybrids and have also suggested a role for parasitism in the evolution of hybrid zones. The relevance of these observations to hybrid zone dynamics is difficult to assess, given the paucity of publications on this subject in recent years.

1.3 The New Zealand Hybrid Zone

As mentioned above, a number of additional hybrid zones have been described around the world but none of these are anywhere near as well studied as that in middle Europe. In this context, the work of Searle *et al* (2009) takes on particular importance. They surveyed mice from throughout New Zealand and, as expected, found *M. m. domesticus* and *M. m. musculus* that would have arrived from Europe on early ships. These were categorised by mtDNA (mitochondrial DNA) D-loop haplotype analyses. Somewhat unexpectedly they also found evidence of hybrid *M. m. castaneus/M. m. domesticus* mice in two regions of New Zealand, namely the South of the South Island and Wellington City in the North Island. Most of their haplotyping was based on mitochondrial D-loop sequencing, which enabled the presumed geographic origins of many of the mice to be established. The predominant *M. m. domesticus* mtDNA haplotypes (domNZ.1-5) can be traced to lineages in Britain and Germany, with domNZ.4 being identical to that of BritIsl.5 in Britain and domNZ.1 identical to the Anglo-German haplotype BritIs.1. Five other *M. m. domesticus* haplotypes are represented in New Zealand and surrounding islands. One *M. m. musculus* mtDNA haplotype was found (musNZ.1) in Wellington and is similar to lineages found in central Europe. *M. m. castaneus* mtDNA haplotypes (casNZ.1&3), found in Wellington and the Southern South Island, are similar to those in Southern Asia but their exact origin was not established. Limited nuclear marker studies were also performed using *Abpa, Btk, D11cenB2* and *Zfy2*. These are standard nuclear markers that have been used for many years to distinguish between *M. m. musculus, M. m domesticus* and *M. m. castaneus* subspecies of mice. All *M. m. musculus* and *M. m. domesticus* mice proved to be 'pure subspecies' in that both nuclear and mitochondrial markers were concordant for each subspecies. However, the mice with *M. m. castaneus* type

mitochondrial haplotypes (casNZ.1&3) all typed as *M. m. domesticus* for the nuclear markers described above. In other words, these mice are hybrids.

As mentioned in the general introduction, such *M. m. castaneus/M. m. domesticus* hybrid mice have only been documented in a few other areas of the world. Thus the existence of such mice in New Zealand is of particular interest.

Chubb (2008) confirmed and extended the work of Searle *et al* (2009) before it was published, using the same mitochondrial and nuclear marker techniques, demonstrating again that such hybrid mice exist in three localities in the south of the South Island and also reporting them for the first time between Dannevirke and Featherston in the lower North Island. She also identified additional *M. m. castaneus* mtDNA haplogroups.

Neither Searle *et al* (2009) or Chubb (2008) identified any 'pure' *M. m. castaneus* mice anywhere in New Zealand, nor for that matter did they identify any *M. m. castaneus* type alleles in any mice using the four nuclear markers.

1.4 Why *M. m. Castaneus/M. m. Domesticus* Hybrid Mice in New Zealand?

Mice were only introduced into New Zealand with the arrival of Europeans and appreciable numbers were first observed in the South Island in the 1860s (2005). Obviously, mice would have spread from port areas and, indeed, early records do describe them moving inland with time (White, 1890). New Zealand shipping also had connections with far eastern ports, mostly in the whaling and sealing trades, and Chinese miners began arriving from around

1866. *M. m. castaneus* mice would undoubtedly have populated these ships and thus reached New Zealand. This is the case at Lake Casitas in California, where Kozak and O'Neill (1987) have shown, by analysis of endogenous viral sequences, that the original *M. m. castaneus* mice leading to the *M. m. domesticus/M. m. castaneus* hybrids that currently inhabit the area very likely arrived with immigrants and cargo from China.

One can only theorise on the fate of these *M. m. castaneus* mice that arrived in southern New Zealand ports. Possibly they "survived" because, according to contemporary accounts, until the 1860s, mouse numbers were very low, especially inland. overall mouse numbers were low. This contrasts with the situation in Britain where, despite the presumed continual arrival of *M. m. musculus* (and *M. m. castaneus*) mice in returning ships, only *M. m. domesticus* mice are found on the mainland, with *M. m. musculus* restricted to outlying islands. *M. m. domesticus* is believed to have arrived in Britain some 2000 years ago and *M. m. musculus* probably arrived on outer islands via Scandinavia in more recent times (Jones *et al*., 2011; Searle, Jones *et al*., 2009). Presumably, in the latter situation, hybridization between *M. m. domesticus* and *M. m. musculus* occurred on the islands because the populations of each were small whereas on the mainland appreciable hybridization never occurred because the pre-existing *M. m. domesticus* population was so large. In New Zealand, in the mid 1850s, mouse populations were very small, according to contemporary accounts, and an analogous situation to that in the Faroe Islands could have existed with hybridisation events between immigrant *M. m. castaneus* and *M. m. domesticus* mice occurring. Subsequent population explosions in the 1860s (King, 2005) for some unknown reason appear to have favoured the *M. m. castaneus / M. m. domesticus* hybrid mice in Southern New Zealand. In other words, any *M. m. castaneus* mice that arrived in New Zealand did not establish themselves *per se*, except through hybridisation events with *M. m. domesticus*. Another possibility is that *M. m.*

domesticus and *M. m. castaneus* hybridised on the ships, en route to New Zealand, and again pure *M. m. castaneus* never established itself in New Zealand.

No matter what scenario holds, the basic fact is that no extant pure *M. m. castaneus* have been found in New Zealand. However, the hybrid mice are clearly viable and (judging by the apparent geographical spread of the hybrid population) have not been appreciably selected against, and perhaps even demonstrate some selective advantages, particularly in the southern South Island.

1.5 Genetics of Hybrid Mice

The first cross (hybridisation) of two subspecies of mice yields an F1 hybrid that will have a maternal mitochondrial genome and one autosome of each pair from each parent. Without selective pressures, further generations descending from interbreeding of F1 hybrids will continue to contain 50% contributions from, for example, each of *M. m. domesticus* and *M. m. castaneus* but these will be on smaller and smaller sized recombined chromosomal fragments until, ultimately, haplotype blocks of each parental type will be distributed randomly throughout the genome. However, in the case of the N.Z. *M. m. castaneus/M. m. domesticus* hybrids it appears, at least for the four nuclear markers examined, no *M. m. castaneus* alleles remain. This implies that after initial hybridisation between the two subspecies, extensive backcrossing has occurred between *M. m. domesticus* males and the *M. m. domesticus* nDNA/ *M. m. castaneus* mtDNA hybrid mice.

Classic backcrossing formulae can be used to get some estimate of how much *M. m. castaneus* genome could remain in the hybrid mice, bearing in mind the fact, that some 200 generations of mice are likely to have occurred since the introduction of mice into New Zealand (1-2 generations per year are believed to occur for mice in the wild). If no alleles in *M. m. castaneus* confer selective advantage the backcrossing formula $[(1/2)^N]$ gives the fraction of the genome that is *M. m. domesticus* after 'N' generations of backcrossing (see http://www.informatics.jax.org/silver/frames/frame3-2.shtml). Thus, after 4 generations, 94% of the genome would be *M. m. domesticus* and after 10 generations this would have increased to 99.8%. In other words, after 100 generations, virtually none of the genome would be expected to retain *M. m. castaneus* alleles. If a *M. m. castaneus* allele provided a strong selective advantage then the fraction of the genome retained at this locus would be approximately 200/N (see http://www.informatics.jax.org/silver/frames/frame3-2.shtml), which would imply that after 100 generations approximately 2 Centimorgan (cM) remnants of the *M. m. castaneus* genome would be retained. Thus, even in a 'best case scenario', one would only expect to find *M. m. castaneus*-like regions of DNA of a few Mb (megabases) in length.

So, even given the relatively recent introduction of house mice into New Zealand, it is quite conceivable that ongoing backcrossing of this type could give rise to hybrids with overwhelmingly *M. m. domesticus* nuclear genomes. Clearly, remnants of the *M. m. castaneus* genome are not going to be detectable using just a few nuclear markers. Rather, hundreds if not thousands of diagnostic markers spread uniformly through the genome would be required.

1.6 High Density Genome Arrays.

The nuclear markers described above were of limited utility as they were based on RFLPs (restriction fragment length polymorphisms) and only a few were developed. In the 1980s and 1990s these were supplemented by SNP (single nucleotide polymorphism) markers and reports appeared that employed tens to hundreds of these markers to give finer resolution to genome analyses. As there are 20 chromosomes in the mouse genome, encompassing 3 billion bases pairs and 3000 cM, this still on average only gave, at best, a coverage of 5 to 10 SNPs per chromosome, that is one SNP per 3 million bases which equates to 15 to 30 cM. Thus, small residual regions of the *M. m. castaneus* genome would not be detected.

In order to overcome this limitation, The Jackson Laboratory (Bar Harbor, Maine, USA) designed a high density 600K SNP array (called the 'Mouse Diversity Genotyping Array') in the mid 2000s (Frazer *et al*., 2007). Based on an Affymetrix chip, this provides one SNP, on average, every 2 to 3 kilobases (kb) throughout the genome. This resolution is sufficient to enable the makeup of the genome of the common JAX (The Jackson Laboratory) classic inbred mouse strains to be broken down into contributions from the four 'wild type' strains of mice held by the laboratory, namely WSB/EiJ (*M. m. domesticus*), PWD/PhJ (*M. m. musculus*), CAST/EiJ (*M. m. castaneus*) and MOLF/EiJ (*M. m. molossinus*) (See Figure 1.1).

1.7 Wider Applications of the high resolution 'Mouse Diversity Genotyping Array'

In theory, the 600K array can be used to analyse any mouse strain (laboratory or wild type) and determine, right down to the haplotype block level, contributions to that genome from

various ancestral mouse strains. Detailed results from such analyses are yet to be published, but an extensive database using this approach has been constructed by de Villena *et al* (http://msub.csbio.unc.edu/). Using this Molecular Phylogeny Viewer, the SNP data derived using this array for 100 classical laboratory strains, 60 wild-derived laboratory strains and 36 wild caught mice (as of March 2011) can be compared. Thus, it is possible to derive, down to the haplotype block level, the subspecific origins of any chromosomal region of a given mouse strain. For example, when the wild derived CAST/EiJ strain is analysed in this way, the analysis shows that while some 85% of the genome is of *M. m. castaneus* origin, there are actually substantial regions of *M. m. domesticus* genome (Frazer *et al*., 2007), indicating that some 'contamination' with *M. m. domesticus* strains has occurred during the breeding programme. Likewise, in other supposedly 'pure' strains, combinations of *M. m. castaneus, M. m. domesticus, M. m. molossinus* and *M. m. musculus* can be found in the genomes. Using the 600k array, any wild type mouse can be, in theory, analysed and even small regions of one sub-species origin, for example, *M. m. castaneus*, detected against a predominant *M. m. domesticus* background. Clearly, this is the method of choice for characterising the N.Z. hybrid mice.

1.8 Limitations of High Density SNP Arrays.

While the 600K SNP arrays are the method of choice for analysing the finer structure of mouse genomes, several limitations complicate the analyses. First, the design of the 600K array was not subject to the same high level of quality control as the well established human Affymetrix arrays. A particular problem arises from some ambiguous SNP probes which do not detect just one site on the genome. At least 5% of the SNPs suffer from this problem (de

Villena, personal communication). Second, although the SNPs on the array were assigned to different subspecific origins (*M. m. castaneus, M. m. musculus, M. m. molossinus, M. m. domesticus*), in hindsight, these assignments were somewhat simplistic and many are wrong (de Villena, personal communication). This arose partly because the wild-derived laboratory strains are not pure. In an effort to overcome this problem, 36 wild-trapped mice have been used to obtain 'authentic' SNPs characteristic of each wild-trapped subspecies (Yang *et al*., 2011). This subset of SNPs are referred to as 'strain specific' and 'diagnostic' and are given relative weightings of 0.05 to 1. A diagnostic SNP, given a weighting of 1, is found in 100% of *M. m. castaneus* wild-trapped mice, and not in the other two wild-trapped subspecies, whereas a SNP with a weighting of 0.05 is found in only approximately 5% of these mice. However, it is not clear if these wild type mice, which do not represent all geographical areas, are truly representative and, if for example, the *M. m. castaneus* samples are from one geographic area which is different from the Asian port areas that the *M. m. castaneus* introduced into New Zealand presumably came from.

Another limitation is that SNPs for the Y chromosome give highly problematic calls and cannot be used.

1.9 Defining the Southern Hybrid zone

Recently we have extended the work of Searle, Jamieson *et al*. (2009) and Chubb (2008) and located a hybrid zone in the south of the South Island of New Zealand (McCormick & Wilkins, 2010) which, as stated above, is of a very different nature from those described elsewhere in the world. This is a broad zone where 'pure' *M. m. domesticus* mice to the north of the hybrid zone abut mice with *M. m domesticus* nuclear DNA and *M. m. castaneus*

mitochondrial DNA to the south of the zone (again, no pure *M. m castaneus* mice were found). In an extension of earlier work, the only mice we found below 45^{o}S were *M. m. domesticus/M. m. castaneus* hybrids, but from just south of Timaru (44.40^{o}S) to Temuka (44.25^{o}S) we found a contact zone with intermixed populations of hybrid and 'pure' *M. m. domesticus* mice. What makes this hybrid zone, and particularly the contact zone, interesting is that it has become established in very recent biological time, namely in the last 150 years or so (Ruscoe & Murphy, 2005) with the colonisation of New Zealand by Europeans, (bringing with them *M. m. domesticus* and *M. m. musculus* subspecies) and the introduction of *M. m. castaneus* (presumably, rather than a pre-existing hybrid!) via the Asian shipping fleet (M. King, 2003 ; Ross, 1987; Smith, 2002). Thus, the investigator has the advantage of working with a mammalian system that is essentially evolution in action (compared with most major mammalian evolutionary events which have time scales typically of tens of thousands if not millions of years). It is possible, given the recent introduction of house mice, that this hybrid zone is still dynamic. Moreover, the three subspecies of mice are arguably really distinct species with most, if not all of the classical differences that distinguish species even for hybrids such as *M. m. molossinus* in northern Japan (Terashima *et al*., 2006), Faroe house mice (Jones *et al*., 2011) and even those in Southern New Zealand.

Much of this introduction has been concerned with the genomic rearrangements that occur during hybridisation. The clear *prima facie* evidence for recent hybridisation events in N.Z, based on initial mitochondrial and nuclear genotyping analyses, has been presented above and it has been pointed out that only high resolution SNP analyses will reveal the exact constitution of the hybrid genome. However, in addition to uncovering any remaining *M. m. castaneus* nuclear DNA, *per se*, such studies may highlight areas of the genome that are

undergoing natural selection and provide clues into the genetic factors responsible for reproductive barriers between the subspecies. This information is obviously applicable to mechanisms of mammalian evolution, especially events occurring early in divergence.

1.10 Hypotheses, aims and objectives

This study consists of two parts.

In the first part, a detailed analysis of the distribution of hybrid mice in the southern South Island is carried out in order to establish the northern limit of this zone and where it makes contact with 'pure' *M. m. domesticus* (the 'contact zone'). This enables mice to be classified into three regions, those far to the south of the contact zone, those near the contact zone where introgression of genes is most likely, and those far to the north of the contact zone which are presumably 'pure' *M. m. domesticus*.

In the second part, representative mice from the three regions are subjected to high density 600K SNP analysis so that the extent of residual *M. m. castaneus* genomic contributions can be established.

Hypotheses

I hypothesise that, despite the fact that hybrid mice in New Zealand possess nuclear genomes that are largely *M. m. domesticus* in makeup, they retain key *M. m. castaneus* alleles and

genomic regions that contribute to their survival and the exclusion of pure *M. m. domesticus* mice in the southern South Island.

Specific Aims

To ascertain the detailed genomic origins of the hybrid mice, the introgression of genes from these hybrids across the contact zone, whether or not this zone is dynamic, and the possible selective advantages these mice might have over pure *M. m. domesticus* mice to the north of the contact zone.

Specific Objectives

- To establish the location, size and nature of the contact zone between the two subspecies by mtDNA analysis and genotyping of several nuclear genes.
- To obtain high quality DNA from 7 representative wild mice spanning the study area for use in SNP microarray analysis.
- To obtain (via a JAX services contract) SNP microarray data on the selected samples in order to obtain high resolution genomic data.
- Analysis of SNP data by a number of bioinformatics and statistics programmes to establish the high resolution genomic makeup of the 7 wild-trapped mice.
- Confirm that subspecies specific regions identified in the SNP analyses are representative of those found in the population at large, by screening more mouse samples using simple PCR assays.

Chapter Two: Methods

2.1 Introduction

The ultimate aim of this research was whole genome haplotyping of hybrid mice using high resolution (600k) SNP microarrays with analysis of this data using various computer software and bioinformatics programs. In order to do this, preliminary screens using basic DNA analyses were carried out on the full sample set including genotyping with traditional nuclear markers, mitochondrial genotyping and haplotype studies using mitochondrial D-loop sequencing.

2.2 Sample collection

Mouse tissue samples were collected from 20 sites around New Zealand, extending up to the Northern limit of the hybrid zone near Timaru and from several '*domesticus*' sites north of the zone. Southern samples were gathered from Te Anau to Lincoln, with 15 sites in total. North Island samples were collected from the Hamilton area (see Appendix I for full details). Sample sites (including droppings) were isolated farmhouses or outbuildings except for those from Lincoln and Hamilton City, B3 from Dunedin City and X3-12 from Temuka township. Approximately 170 samples were collected in total, which were either 'fresh', that is mice which were immediately frozen after trapping (with traditional spring loaded traps) or dried samples which were tails from trapped animals. Carcasses from poisoned animals and faeces were also used. The majority of the samples were collected by local residents on a voluntary

basis. Ninety six previously gathered samples (Appendix V) were provided by Tanya Chubb (Chubb, 2008).

2.3 DNA extraction

Usually, a 5 to 20 mm tail tip from each animal was used for DNA extraction, with the rest of each tail being dried and archived for later reference. In some cases, ear clippings or droppings were used as a source of DNA. When very high molecular weight DNA was required, such as for SNP microarray analyses, livers were used from frozen animals. The concentration and quality of DNA samples was checked by obtaining an absorbance spectrum using a NanoDrop spectrophotomer and also 1% agarose gel electrophoresis to ensure the DNA was of high molecular weight.

2.4 DNA extraction from tails

For mouse tail tips, the DNA extraction protocol firstly involved pulverisation of the tissue with a hammer, performed aseptically in between Mylar sheets. The crushed tails were then added to 15 ml falcon tubes with 1 ml lysis solution (20 mM Tris, 50 mM EDTA, 1% SDS, 100 µg/ml proteinase K). These were rotated at 55ºC for between 2 and 12 hours with occasional mixing by hand until the tissue was completely dissolved (except hair). Then 2 µl boiled RNAse was added to each tube followed by 1 ml of 50:50 phenol:chloroform and the tube contents mixed to a complete emulsion by rotating for 10 minutes at room temperature. After centrifugation at 13,000 rev/min for 10 minutes, the upper aqueous phase was recovered and the DNA was precipitated by adding and mixing with an equal volume of

isopropanol. Thin glass rods were used to spool out the DNA from those samples that had visible precipitate; otherwise samples were spun at 13,000 rev/min for 10 minutes to pellet the DNA. The DNA was then washed twice in 70% ethanol and re-suspended in TE buffer (10 mM Tris, 1 mM EDTA, pH8.0).

2.5 DNA extraction from tails in 96-well Plates

The DNA extraction technique from mouse tail tips was simplified by reducing the SDS to 0.01% in a lysis solution consisting of standard TE pH8.0 buffer with Proteinase K to a final concentration of 100 μg/ml. This solution is suitable for straight-to-PCR use because the very low concentration of SDS does not inhibit PCR amplification; 150 μl of this solution was added to each crushed tail tip (approximately 5 to 10 mm long) in a deep well microtiter plate and the samples were shaken from 1 to 16 hours at 60°C, depending on how long it took for each tail to dissolve. The tail solutions were then diluted 1:10 into another plate with water and heated to 95°C for 10 minutes to inactivate the Proteinase K. DNA from ninety six mouse tails, most of which had been previously characterised by sequencing by Tanya Chubb (2008), were prepared in this way. For further details see McCormick and Wilkins (2010).

2.6 DNA extraction from livers

This was performed with samples from 7 animals (see table 2.3) in order to provide very high quality, high molecular weight DNA that met the Jackson Laboratory specifications for genotyping on SNP microarrays. The livers of frozen animals were subjected to standard Tris:EDTA:proteinase K: SDS treatment (as described above for DNA extraction from tails) followed by phenol:chloroform extraction and isopropanol precipitation. The visible

precipitate was spooled out with glass rods, followed by washing in 70% ethanol. After checking the absorbance spectra quality (NanoDrop) and molecular weight (agarose gels) the DNA samples were sent to The Jackson Laboratory (Maine, USA) for SNP microarray analysis.

2.7 DNA extraction from droppings

The DNA from droppings was processed using a simple protocol which involves grinding with glass beads (McCormick and Wilkins, in preparation). In summary, a single, dry pellet (dropping) was used in each preparation. If the pellets were moist they were dried out by brief incubation at 80 ° C. This is important to ensure the pellet remains intact during vortexing. Approximately 200 µl of 0.1 mm zirconia/silica beads (BioSpec Products) were added to 1.5 ml tubes along with one pellet. Tubes were then vortexed for 2-3 minutes on full power. In this way, a thin layer of the surface of the pellet is ground off, avoiding the bulk faecal matter which can contain bacteria and PCR inhibitors. The pellet was then removed with tweezers along with any visible fragments and either discarded or archived for later reference. Fifty µl of TE pH 8 buffer was added and the tubes were vortexed again in order to fully wet the beads. 500 µl of chloroform was added followed by mixing again by vortexing and then centrifugation at13,000 rev/min for 1 minute. The bubble of buffer, which forms at the top of the chloroform, was carefully collected by pipette and transferred to new tubes. Two µl of this solution was added to each 20 µl PCR reaction for mitochondrial genotyping as described below.

2.8 Mitochondrial genotyping

In order to quickly test a large numbers of samples while still in the field, and thereby establish the location and extent of the hybrid zone, a reliable and quick mitochondrial genotyping method was required. Sequencing of the mitochondrial D-loop region would have been expensive and time consuming with such a large number of samples. Hence, a multiplex restriction fragment length polymorphism (RFLP) test was designed for this purpose (see McCormick and Wilkins, 2010).

The published mtDNA restriction maps of ten *Mus musculus* subspecies (She, Bonhomme, Boursot, Thaler, & Catzeflis, 1990) were studied and we focused on those restriction enzymes with cut sites that were common to large numbers of the sub-species but which differed in *M. m. castaneus* and *M. m. domesticus* mtDNA (and, incidentally, also *M. m. musculus*). Of these, three regions and three enzymes were selected that appeared to give unique sub-species specific restriction digests, namely *Bam*H1 (which cuts *M. m. domesticus* at nt.11,177 and *M. m. castaneus* at nt.14,249), *Eco*R1 (cuts only *M. m. castaneus* at nt.8538) and *Pst*1 (cuts only *M. m. domesticus* at nt.8421). The Genbank reference sequence for nucleotide numbering was the mtDNA genome EF108344 for *M. m. domesticus*; the corresponding *M. m. castaneus* sequence that we used, EF108342, is the only complete mtDNA genome sequence of this sub-species deposited in Genbank; it is to be noted that there are slight numbering differences between these two sub-species and also other deposited mtDNA sequences. One of the motif sites, an *Eco*R1 cut at nt.8538, was based on comparison of the two deposited sequences, and was not reported by She *et al*. (1990). Primers were designed to flank each restriction site using Primer3 (Rozen & Skaletsky, 2000), selecting sequence regions homologous to both genomes, and further small adjustments made so it was possible to multiplex all three reactions.

Two microlitres of each diluted DNA sample was added to a multiplex PCR reaction of 20µl total using the standard Platinum Taq reaction mix recommended by the manufacturer with 2.5mM $MgCl_2$, 100µg/ml purified BSA (New England Biolabs), primers M1, M2, M5 and M6 each at a concentration of 420nM and primers M3 and M4 at a concentration of 170nM. The thermocycle programme consisted of an initial heating step at 95°C for 4 min then 39 cycles of 94°C for 20 s, 55°C for 30 s and 72°C for 30 s.

The PCR samples were then digested by adding 3µl of a mix that contained 5U each of *Bam*H1, *Eco*R1 and *Pst*1 and 1µ of 10X restriction buffer H (Roche) and incubating for 30 minutes at 37°C. The restriction digests were electrophoresed on 1.5% agarose gels containing ethidium bromide, using a 100bp ladder for size comparison.

2.9 Nuclear Markers

Three nuclear markers (see table 2.1) were used (together with the mtDNA markers described above) to crudely genotype selected mouse DNA samples. A standard PCR protocol was used in each case (400 nM primers, 200 µM dNTPs) with adaptations according to primer set and the addition of 1.5 mM $MgCl_2$ and purified BSA at 100 µg/ml.

Table 2.1

Nuclear markers						
Gene / marker	**Gene product**	**Primer sequences**	**Reference**	**Expected product sizes**		
				Domesticus	**Musculus**	**Castaneus**
Abp α	Androgen binding protein α subunit	AbpαF 5'GAAACAATTCAATGAAAACACTAAAG3' AbpαR 5'TGTGCCACTGCTCTGTATTC3'	Laukaitis *et al*., (2008)	192	-	-
Abp β	Androgen binding protein β subunit	AbpβF 5'ACAATTCAATGAAAACCGTGA3' AbpβR 5'AAACTTGGGCAGGGATTTAG3'		-	303	303
Zfy2	Zinc finger protein 2	ZfyF 5'CATTAAGACAGAAAAGACCACCG3' ZfyR 5'GTGAGGAAAT'ITCTTCCTGTGG3'	Boissinot and Bousot (1997)	202	184, 202	184, 202
Btk	Bruton agammaglobulinemia tyrosine kinase	BtF 5'AATGGGCTAGCGTAGTGCAG3' BtR 5'AGGGGACGTACACTCAGCTTT3'	Munclinger *et al*., (2003)	342	206	206

2.10 Mitochondrial D-loop sequencing

Sequencing of the mtDNA D-loop region (also known as the control region) was carried out to confirm the identity of some of the mouse samples and also to confirm accuracy of the genotyping methods. The primer sets L15320 & H15782 and L15735 & H00072 (Prager *et al*., 1993) were used for this, the sequences of which are displayed in table 2.2 . This form of identification was also used by Chubb (2008).

PCR reactions were carried out at a total volume of 30 µl with 0.4 µM primers, 1.5 mM $MgCl_2$ and100 µg/ml purified BSA (New England Biolabs). The thermocycle parameters consisted of an initial heating step at 95°C for 4 min then 39 cycles of 94°C for 20 s, 55°C for 30 s and72°C for 30 s. Products were then purified from agarose gels using the Zymoclean Gel DNA Recovery Kit (Zymo Research).

Sequencing was carried out by the Waikato DNA Sequencing Facility (University of Waikato, Hamilton, New Zealand) using commercial dye terminator chemistry (either GE Healthcare Life Sciences DYEnamic-ET or Applied Biosystems Big Dye v3.1. Forward and reverse sequences for each sample were aligned and manually edited using Geneious v4.8 (http://www.geneious.com) followed by a BLAST search.

Table 2.2

Mitochondrial D-loop sequencing primers (Prager *et al*., 1993)	
Primer name	Primer sequence
L15320	5’ CTTAACACCAGTCTTGTAAACC 3’
H15782	5’ CCTGAAGTAGGAACCAGATG 3’
L15735	5’ CCAATGCCCCTCTTCTCGCT 3’
H00072	5’ TATAAGGCCAGGACCAAACCT 3’

2.11 SNP micorarrays

The Jackson Laboratory was contracted to generate the raw SNP data using their in-house JAX Mouse Diversity Genotyping Arrays (see http://jaxservices.jax.org/mdarray/index.html). This service uses a Affymetrix 600k array, giving a SNP an average of every 4.3 kb across the genome. Seven samples (see table 2.3) were analysed in total, providing coverage of the contact zone as well as northern and southern extremities of the hybrid zone. These seven samples are from (north to south) the Hamilton, Lincoln (*M. m. domesticus*, north of the contact zone), Temuka, Mawaro (hybrids but in the contact zone), Waianakarua, Te Anau and **Taieri** (hybrids, well within the hybrid zone) collection sites. DNA samples used for SNP microarray analyses were quality checked by measuring both the 260/280nm absorbance ratio with the NanoDrop spectrophotometer and molecular weight (on agarose gels) to ensure they were pure and of high enough molecular weight to meet the JAX quality standards.

Table 2.3

JAX SNP analysis samples (all were males)			
Sample name	**Collection location**	**Coordinates**	**Date**
TA4	Kakapo Road, Te Anau (silo)	45°26°S 167°46°E	Aug 2009
Tr1	Schoold Road, Taieri plains (barn)	45°51°S 170°19°E	Apr 2009
J2	McKerrow Road, Waianakarua (farmhouse)	45°16°S 170°46°E	Jun 2009
W3	Greenhill Road, Mawaro (farmhouse)	44°18°S 170°52°E	Jun 2009
X1	Rangatira Valley, Temuka (farmhouse)	44°13°S 171°12°E	Apr 2009
Li2	Landcare, Lincoln (centre grounds)	43°38°S 172°30°E	Sep 2009
Ham1	East Street, Hamilton (suburban house)	37°46°S 175°8°E	Sep 2009

SNP data from the arrays was then analysed using a variety of methods, all of which essentially align the SNP data along chromosomes and enable the pinpointing of similarities and differences in genotypes amongst the various wild types and the laboratory reference strains. Some analyses utilised the statistical programme 'R' (Team, 2004), a flexible and powerful means of analysing data and displaying results from very large files. In addition, data from the UCSC (University of California, Santa Cruz) Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway?org=mouse) was used to overlay genomic features such as genes and expressed sequence tags (ESTs), in order to provide insight into possible roles that these regions may play in the evolution of the genomes. Some analyses were performed in Microsoft Excel (2007) using a range of logic statements and formulae. As well as making comparisons between the seven wild mouse samples, SNP microarray data from three inbred reference strains (CAST.EiJ, WSB.EiJ and PWD.EiJ, representing

'pure' *M. m. castaneus*, *M. m. domesticus* and *M. m. musculus* respectively) was used for comparisons.

2.12 SNP data analysis.

2.12.1 Analyses using all SNP data.

The raw data generated in the Affymetrix system essentially scores each SNP position as a 2, 0, 1 or -1 representing, respectively, a homozygous SNP call that differs from the reference mouse genome (C57B6/J), a homozygous call the same as the reference mouse genome, a heterozygous call, or unscoreable.

The data was supplied as text files which were converted into Excel or CSV (comma separated values) format that, essentially, consisted of these calls, the chromosomal locus of the SNP and the identity of the SNP (a JAX and/or a rs number).

Initially, it was believed that the four southern hybrid mice would be best analysed by pinpointing regions that were homozygous for SNPs that differed from those of the three northern '*domesticus*' mice. In order to do this, each SNP score (2, 1 or 0) was added from the four southern mice and compared with the cumulative score in the three northern mice. The assumption was that, if the difference in these scores was large, this suggested a chromosomal region of difference between the hybrid and *M. m. domesticus* mice that should be further investigated.

Initial efforts to do this analysis in R were unsuccessful as the large files proved difficult to upload. A different approach was used in which Excel files for each chromosome were generated and SNP data displayed graphically using a Mathematica 7 (2008) programme (Appendix II) as either homozygous (2 or 0) for each allele, or heterozygous (1) along with reference data for wild type mice WSB/EiJ, PWD/PhJ andCAST/EiJ. Each chromosome could be scanned manually, using the moving window tool in Mathematica, and differences between southern hybrid and northern domesticus mice pinpointed visually.

2.12.2 Analyses using 'diagnostic' SNPs.

In April 2011 Fernando Pardo-Manuel de Villena supplied a database of 'diagnostic' SNPs. These are a subset of the 600K SNPs on the Affymetrix chip that define, with probabilities of 0.05 to 1.0, the degree to which a specific SNP is associated exclusively with the *M. m. domesticus*, *M. m. musculus* or *M. m. castaneus* genome. This database was integrated with the databases for the seven wild-trapped N.Z. mice using a Micorsoft Excel (2007) macro to give a display of all diagnostic SNPs for each mouse, and each chromosome, along with those for the *domesticus* wild type Jackson Laboratory mouse WSB/EiJ. Even this subset of data presents substantial display challenges as each series in an Excel chart is limited to 36,000 data points. In order to display all the data for all eight mice for one chromosome, up to 140,000 data points have to be processed (Chr 1). To do this, data for each mouse was split into 24 series, arranged mouse by mouse into the 3 subspecies SNPs (distinguished by full RGB colourations) which were plotted onto the one chart (see Appendix III on disk for macro enabled Excel files). The pictorial displays (Appendix IV) are arranged, chromosome by chromosome, with SNP alleles scored as *domesticus* (blue), *musculus* (red) or *castaneus*

(green) with the ‘height’ of each allele (0.05 to1.0) giving its strength as a diagnostic SNP. The displays consist of two sets of data, the upper eight for regions of each chromosome of each mouse homozygous for the diagnostic allele and the lower, those that are heterozygous.

The data generated using diagnostic SNPs differs significantly from that using the Jackson Laboratory calls (0 or 2) because, in categorising the latter data, erroneous assumptions were made concerning the degree to which ‘wild type’ laboratory mice WSB/EiJ, PWD/PhJ, CAST/EiJ represented these subspecies in the wild (de Villena, personal communication).

2.13 Screening from SNP microarray results

Once SNP microarray data from the 7 samples had been analysed and regions of interest identified, a number of samples from throughout the country were screened to test for the patterns of these loci within the N.Z mouse population. This was achieved by designing simple RFLP assays (using the WatCut rpogramme at http://watcut.uwaterloo.ca/watcut/watcut/template.php) that tested for the presence or absence of the SNPs.

2.14 Statistical analyses

For the standard molecular biology procedures statistical analyses were not required. For determining significant SNP calls, the standard Affymetrix statistical programs (BRLMM-P program within the Affymetrix Power Tools software package) were used to analyse raw array data from CEL files (Affymetrix, 2007). For the main analyses, the principal question was whether or not the detected differences and similarities in genomic regions of various

wild mice are significant or just chance correlations. Essentially, this can be addressed by using a Chi-square ($\chi 2$) statistic. To do this, Megabase regions were compared between wild type, *M. m castaneus* hybrids and *M. m. domesticus* samples using the inbuilt Jackson Laboratory statistical program at: http://phenome.jax.org/db/q?rtn=snps%2Ffindreg_signif&handle=findreg_aa6hf2g83&groupA=3&groupB=15&chr=1&start=12&stop=13.

Chapter 3: Results - Characterisation of the Contact Zone.

3.1 mtDNA RFLP assay

MtDNA analyses are normally used to delineate the geographic distributions of different mouse subspecies. While the gold standard is sequencing of D-loop regions, this can become quite laborious, expensive and time consuming when surveying large areas and also suffers from the disadvantage of not being adaptable to assays in the field.

In the present study, the hybrid zone involving *M. m. castaneus* and *M. m. domesticus* was geographically large and complex with an uncertain contact zone. In other words, there was a need for extensive sampling.

I chose to utilise simple restriction fragment length polymorphic (RFLP) motifs in mtDNA. (This is a well proven method that was originally utilised in human studies by Allan Wilson and co-workers who used bulk mtDNA; Cann *et al*., 1987). Bozikova *et al*. (2005) used a very simple PCR test of this type to localise hybrid zones of *M. m. domesticus* and *M. m. musculus*, in different geographic localities based on the presence or absence of, respectively, just one *Bam*H1 cut site at nt.3565. A more sophisticated test to distinguish *M. m. castaneus* and *M. m. domesticus* mtDNA is used here, utilising three RFLPs spread throughout the genome and designed to minimise artefacts arising from point mutations and PCR inhibitors.

From the published mtDNA restriction maps of ten *M.m.* subspecies (She *et al*., 1990), one can focus on those restriction enzyme cut sites common to large numbers of the sub-species

but which differ in *M. m. castaneus* and *M. m. domesticus* mtDNA (and, incidentally, also *M. m. musculus*). Of these, three regions and three enzymes that appeared to give unique sub-species specific restriction digests, namely *Bam*H1 (which cuts domesticus at nt.11,177 and castaneus at nt.14,249), *Eco*R1 (cuts only castaneus at nt.8538) and Pst1 (cuts only domesticus at nt.8421) were chosen. The PCR primers, the product sizes and the expected restriction length fragments are shown in table 3.1 (reproduced from McCormick and Wilkins, 2010) .

Table 3.1 Primer sequences and product sizes for mtDNA RFLP assay.

Primer	Sequence	Target +		Product size (bp)	*Bam*H1 product size (bp)	*Eco*R1 product size (bp)	*Pst*1 product size (bp)
M1	TCTCCTAGGCCTTTTACCACA	8175-8669	Domesticus	493	*	*	249,244
M2	GCTCCAGTTAATGGTCATGGA						
			Castaneus	493	*	358,135	*
M3	CACATAGCACTTGTTATTGCATC	11046-11460	Domesticus	416	131,285	*	*
M4	GTGTGTGAGGGTTGGAGGTT						
			Castaneus	416	*	*	*
M5	TTCAACTGCGACCAATGAC	14089-14416	Domesticus	328	*	*	*
M6	AATATTGAGGCTCCGTTTGC						
			Castaneus	328	169,159	*	*

+ Reference sequence EF108342, * No cut site

3.1.1 Extraction of DNA From Tissue

The method of extracting DNA appeared suitable for all mouse tail samples. Some samples were from fresh tails, some from frozen mice, others from carcasses dried for months (or possibly years) after trapping or poisoning, and others from alcohol embalmed specimens. All gave strong PCR products.

3.1.2 Validation of Method

The undigested PCR products were of the predicted sizes but the 416 bp product was overly amplified if equimolar concentrations of all primers were used (data not shown); this problem was corrected by lowering the concentration of the M3/M4 primer set in the PCR reaction. After digestion, the restriction fragment patterns were clearly distinguishable on agarose gels. Thus in Figure 3.1a, the pattern for *M. m. castaneus* consists of bands at 135, 159 and 169 (appearing as one band), 358 and 416 base pairs and that for *M. m. domesticus* of bands at 131, 244 and 249 (appearing as one band), 285 and 328 base pairs. A high resolution 2% agarose gel was used to demonstrate this point. However, the three DNA banding patterns (uncut, restricted *castaneus* and restricted *domesticus*) are so distinct, that running samples in a microwell type format on a 1% agarose gel (7V/cm) for just 10 min is sufficient to distinguish all three (Figure 3.1b). At this time the DNA has only migrated ~1cm. After 30 min, all bands were resolved, but not as clearly as on a 2% gel (Figure 3.1d).

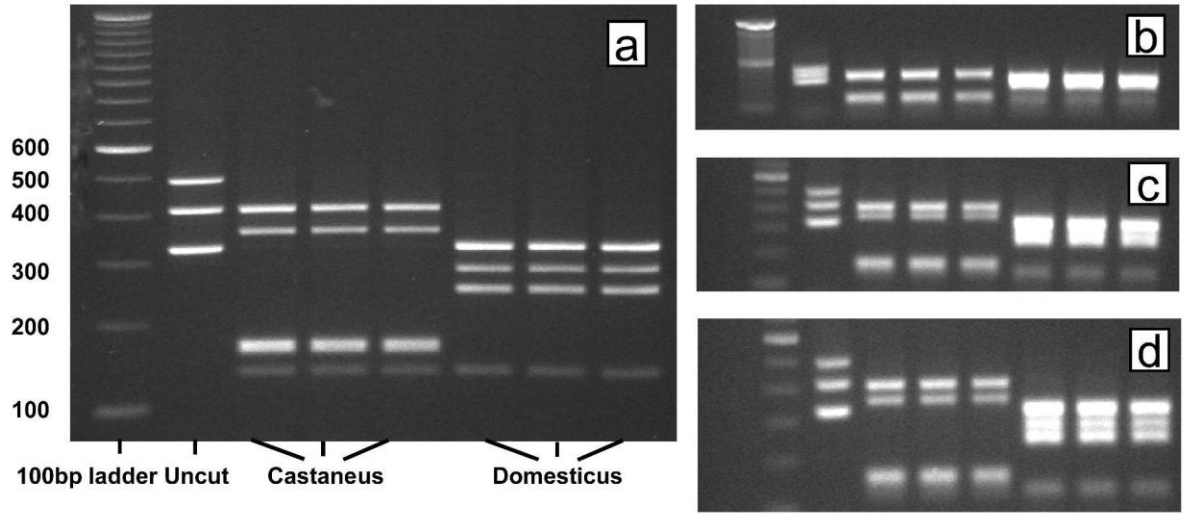


Figure 3.1: RFLP patterns of *castaneus* and *domesticus* multiplexed PCR products. Electrophoresis using (a) a 2% agarose and (b-d) a 1% agarose in a microwell format run for 10, 20 and 30 min respectively.

3.1.3 Validation of mtDNA RFLP assay using JAX Laboratory mouse reference DNA

Samples of DNA for two JAX wild type *castaneus* mouse strains CAST/EiJ (which is widely used as the *M. m. castaneus* reference strain) and CASA/RjK, as well as two *domesticus* wild type strains TIRANO/Ei and ZALENDE/Ei, were subjected to the mtDNA RFLP assay (Figure 3.2). CAST/EiJ, TIRANO/Ei and ZALENDE/Ei all gave the expected restriction digest patterns of *castaneus*, *domesticus* and *domesticus* mtDNA respectively. However, the second *M. m. castaneus* reference strain, CASA/RkJ, typed as *domesticus*! I concluded that this mouse strain is a *M. m. castaneus/M. m. domesticus* hybrid and probably reflects animal husbandry errors. CAST/EiJ and CASA/RkJ strains were bred from the same group of mice captured in Thailand in the 1970s (http://jaxmice.jax.org/jaxnotes/archive/456e.html) but were subject to different breeding programmes. Breeding errors of this nature were apparently not uncommon in the early days of the JAX laboratory (de Villena, personal communication).

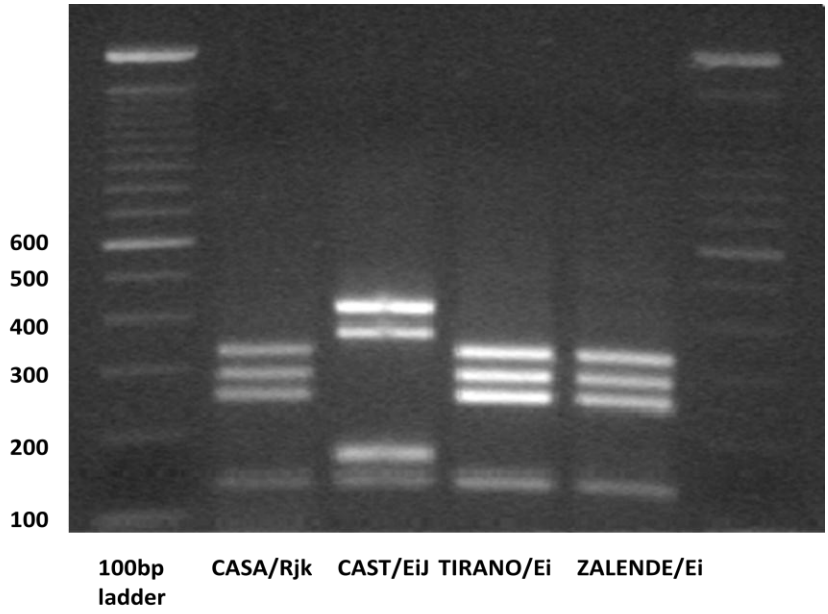


Figure 3.2: mtDNA RFLP test on JAX mice.

3.1.4 Wild-trapped mouse assays

Initially, I genotyped 96 mice using the microtiter plate method from section 2.5 (Figure 3.3 and Appendix V); 68 of these had been previously analysed (Chubb 2008) by conventional (Prager *et al*. 1993) mtDNA sequencing methods. Although some of the bands in Figure 3.3 are faint, the pattern given by each sample is still easily distinguishable. The first point to note is that all samples gave one or the other RFLP pattern; no aberrant patterns that deviated in any way (fewer bands, or bands of different sizes) were seen in any samples in this or subsequent analyses. The RFLP typings were in complete agreement with the forty samples that had been classified as *M. m. domesticus*, as deduced from the earlier sequencing of the D-loop region, and there was also agreement for the 13 *M. m. castaneus* samples from Karori (SH2-25 to SH2-36 and SH2-41) that had been typed by D-loop sequencing . The other 15 samples that had beentyped as *M. m. castaneus* by sequencing, gave RFLP patterns characteristic of *M. m. domesticus* DNA. These were all from the North Island, north of

Wellington. In order to resolve this discrepancy I resequenced two of these samples, H15 and H21, (using archived DNA stored in the microtiter tray) using the primer sets L15320 & H15782 and L15735 & H00072 (Prager *et al*. 1993) to amplify the D-loop region. The results confirmed that the RFLP classification was correct and the D-loop analyses in error, which suggests that the DNA sequence data for the other 13 anomalous samples should be re-examined.

In further analyses, two "mouse" samples (K1 and K2) gave anomalous results, namely just one band of ~328bp, which was not cut by any of the restriction enzymes. I sequenced the D-loop mtDNA of these samples and submitted them to a BLAST search; the alignment was very close to *Rattus exulans* (the Polynesian Rat, EU273711). Obviously, two young rats had been mistakenly collected as mice. (Subsquently, some 10 skeletal remains collected from farm houses also yielded a similar single band – on closer examination, it was concluded that these were the remains of young rats but presumably not *Rattus exulans*.).

It should be noted that some 170 mice (consisting of animal and feacal samples) have been sampled from throughout New Zealand, many using a field laboratory, and that only the two distinctive subspecies restriction fragment patterns have been seen. These results confirmed those reported earlier by Chubb (2008) and Searle *et al*. (2009), namely that the mtDNA is predominantly *castaneus* in the south of the South Island and *domesticus* from mid-Canterbury north (with 3 nuclear marker assays suggesting that the nuclear genomes are predominantly *domesticus* in all mice).

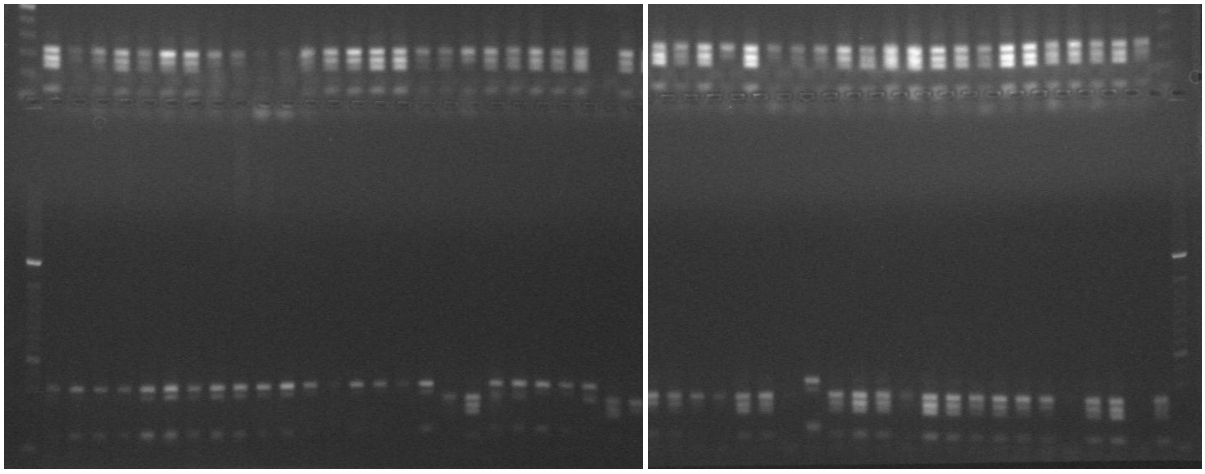


Figure 3.3: mtDNA RFLP assays from low SDS, 96 well plate method. See Appendix V for details.

3.1.5 Extension of method to faecal samples

The mtDNA RFLP assay was also used to type faecal DNA samples that were extracted according to the methods described in section 2.7. Mouse pellets (faeces) provide a useful source of DNA when extensive sampling is required and to this end a simple protocol for DNA extraction was designed (McCormick and Wilkins, in preparation). Figure 3.4 shows the amplification results using the mtDNA RFLP primers (without restriction enzyme digestion) according to the length of time they were subjected to vortexing. As faeces contain bacteria and various PCR inhibitors, the aim of the protocol is to remove only the very surface of each dropping. As can be seen in figure 3.4, results are variable, but usable DNA can be obtained for the majority of samples after 1 to 3 min of vortexing.

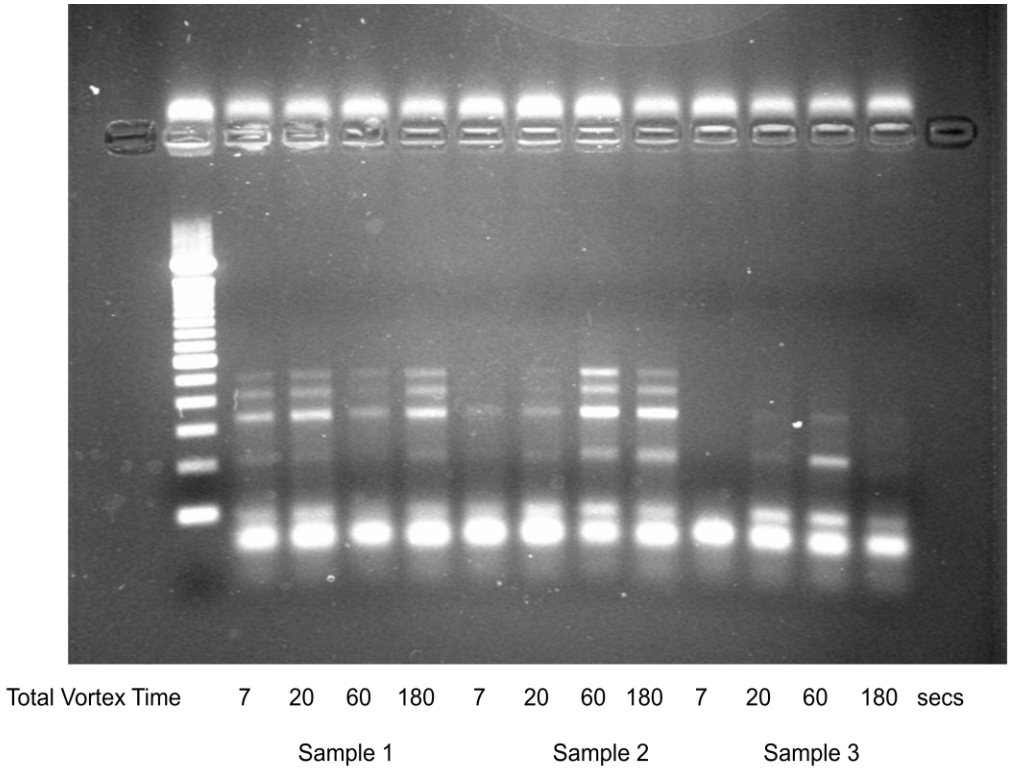


Figure 3.4: Effect of vortex time on mtDNA amplification (undigested) from three faecal samples from different mice. (Extracted material was processed from each pellet after an initial 7 sec vortex then, in succession, after a further 13, 40 and 120 sec of vortexing).

3.2 Mitochondrial DNA D-loop sequencing

Sequencing of the mtDNA D-loop region was carried out for the seven mice used for JAX SNP micorarray analysis (TA4, TR1, J2, W3, X1, Li2, Ham1) using the Prager (1993) primers shown in table 2.2. Ham1, Li2 and X1 were found to be within the haplotype groups DomNZ.1-5, originating from the British Isles, and the most common group of N.Z *domesticus* haplotypes found by Searle *et al*. (2009). The hybrid mice (W3, J2, TR1 and TA4) were all found to be of the CasNZ.1 haplotype from Asia. Edited sequences can be found in Appendix VIII on attached disk.

3.3 Defining the contact zone in the southern South Island.

All mice from Waimate (45^{o}S) and Hakataramea (45^{o}S) and further south were mtDNA genotyped as *castaneus*. From Lincoln north, all mice were mtDNA genotyped as *domesticus*. In between these two areas, from just north of Waimate, inland to Mawaro and north to Temuka the hybrid mice and *domesticus* coexisted, often on the same property (farms and houses). Altogether, 170 Mice were genotyped. Extensive sampling was not done from Ashburton north as earlier work by Searle *et al* (2009) established that northern regions of New Zealand were predominantly inhabited by *M. m. domesticus*, and they only found *M. m. castaneus* in a localised area in Wellington City. The results of initial mtDNA genotyping studies are presented in Figure 3.5.

These data clearly define a contact zone extending from approximately 45^{o}S to 44.25 ^{o}S over some 50km and extending from the sea to at least 40km inland. As with the hybrid zone in middle Europe, no distinctive geographical, ecological or climatic factors define this contact zone and, although there are rivers and low ranges in the area none are remarkable and in fact much more dramatic demarcations occur at the large braided rivers to the north and south (Rakaia and Waitaki, respectively). Nevertheless, it should be borne in mind that as one proceeds south of Waimate and west (Hakataramea Valley) into the Otago province, one moves from an 'Eastern South Island' climate zone into 'Southern New Zealand', 'Inland South Island' and 'Western South Island' zones (as defined by the National Institute of Water and Atmospheric Research, NIWA). In general, these last three zones are colder (especially in winter) with more frost days than the Eastern South Island zone (and the rest of New Zealand).

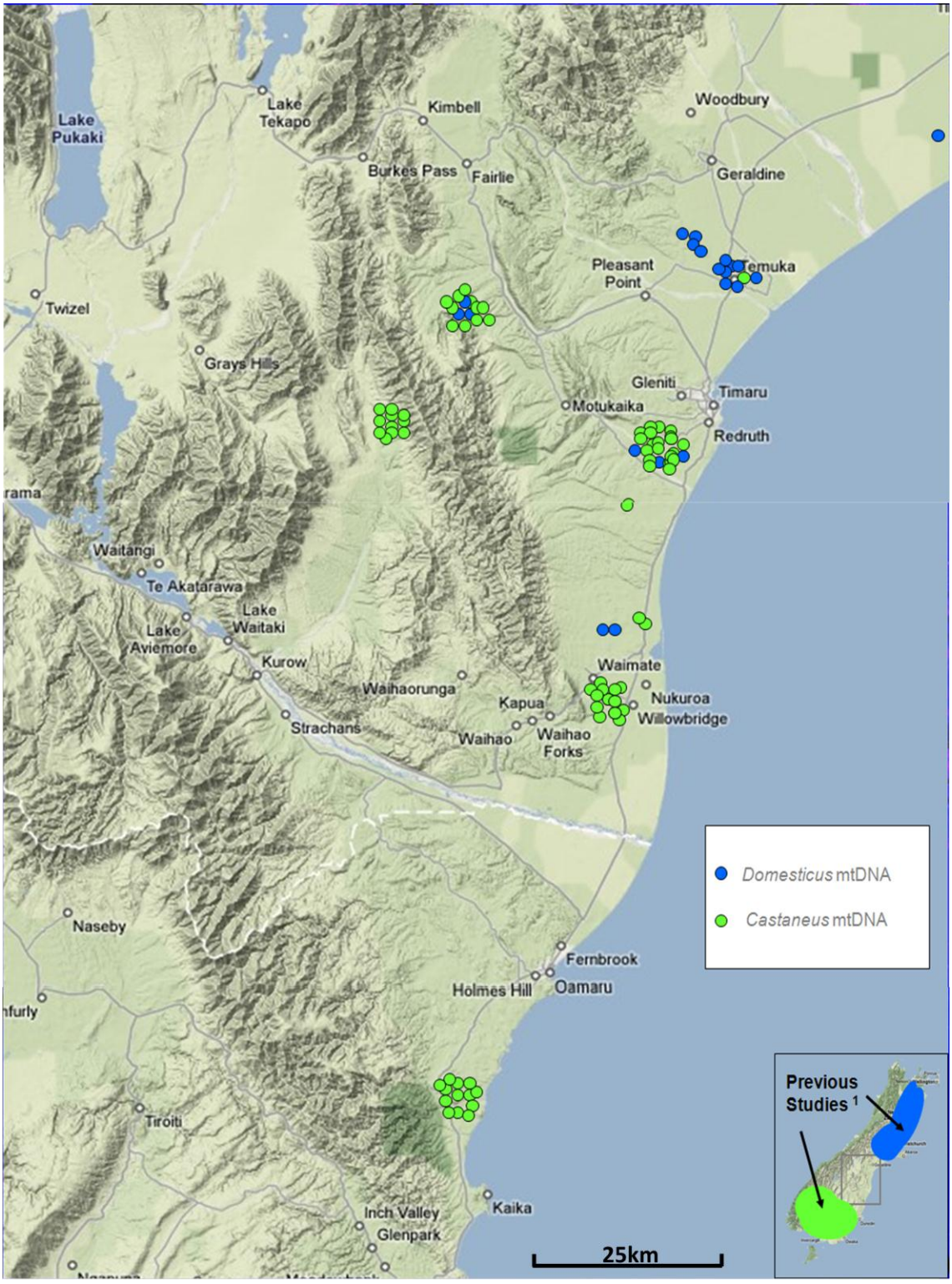


Figure 3.5: Initial field studies that defined the contact zone
[1] Searle *et al*., (2009)

3.4 Nuclear Markers

Most of the 170 mouse DNA samples were also genotyped using 3 standard nuclear markers (Table 2.1). All samples typed as *M. m. domesticus* for these markers.

Androgen binding protein (*Abp*), encoded on mouse chromosome 7, is found in the saliva of rodents and is thought to be involved with sexual identification, chemosensation and subspecies recognition in mice (Laukaitis *et al*., 2008). As shown in table 2.1, *M. m. domesticus* mice test positively for the α subunit, giving a band at 192 bp, whereas the β subunit is seen (at 303 bp) for *M. m. musculus* and *castaneus* mice. All samples typed as *M. m. domesticus* according to this marker. Examples of these results are shown in figures 3.6 and 3.7.

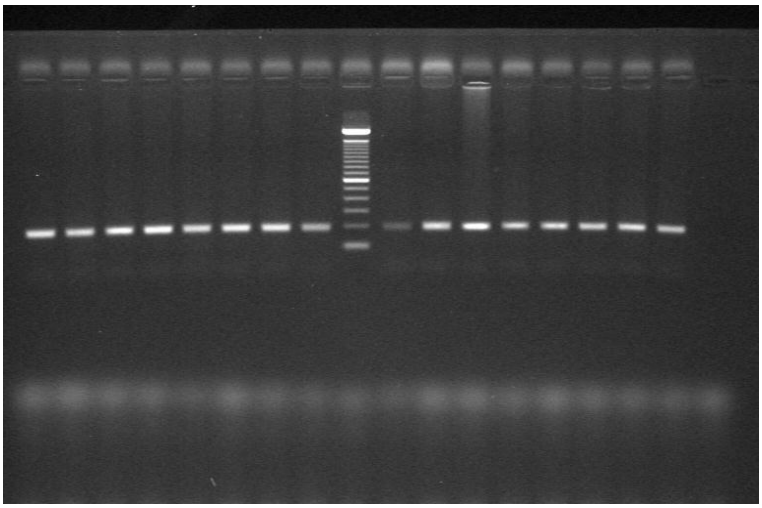


Figure 3.6: *Abpα* marker. From left to right: X7, Wn1-7,100 bp ladder, Wn8-15.

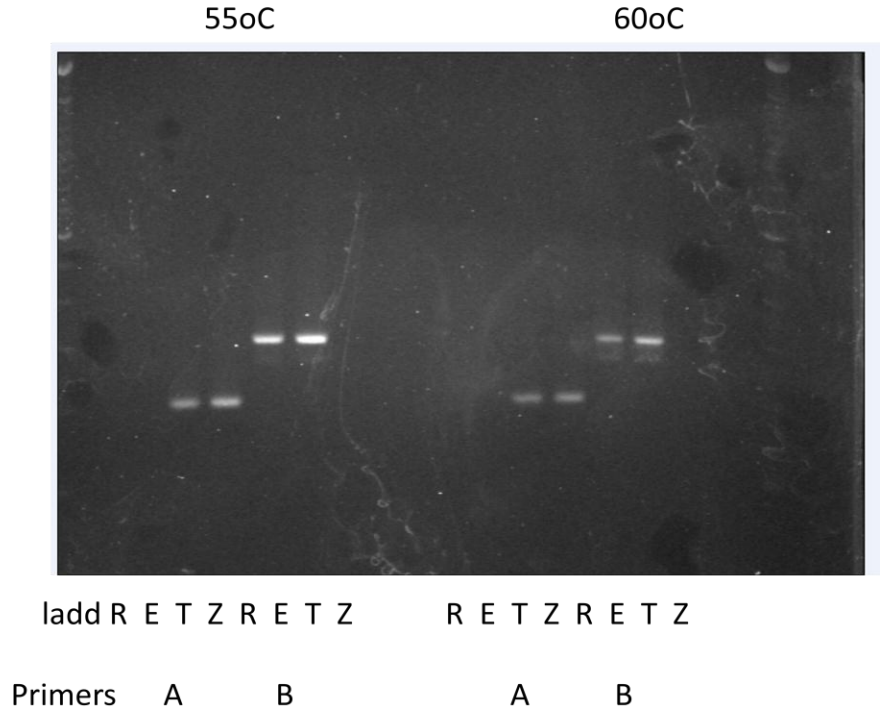


Figure 3.7: *Abpβ* and β (A and B) nuclear markers at two different annealing temperatures. R,E,T and Z are Jackson Laboratory wild type reference strains. R: CASA/RjK, E: CAST/EiJ, T: TIRANO/EiJ, Z: ZALENDE/EiJ.

The Y chromosome genes *Zfy1* and *Zfy2* (zinc finger protein 1 and 2) were used to type DNA samples according to the presence or absence of an 18 bp deletion (Boissinot & Boursot, 1997; Nagamine *et al*., 1992). The primers, shown in table 2.1, amplify *Zfy1* and *Zfy2* equally, so that *M. m. domesticus* samples yield a 202 bp product (deletion absent) whereas *M. m. castaneus* and *M. m. musculus* mice give two bands at184 and 202 bp. All samples typed as *M. m. domesticus* for these markers (examples of results shown in figure 3.8). Either a faint product, or no product is given for female mice as can be seen in lanes 6, 7, and 9.

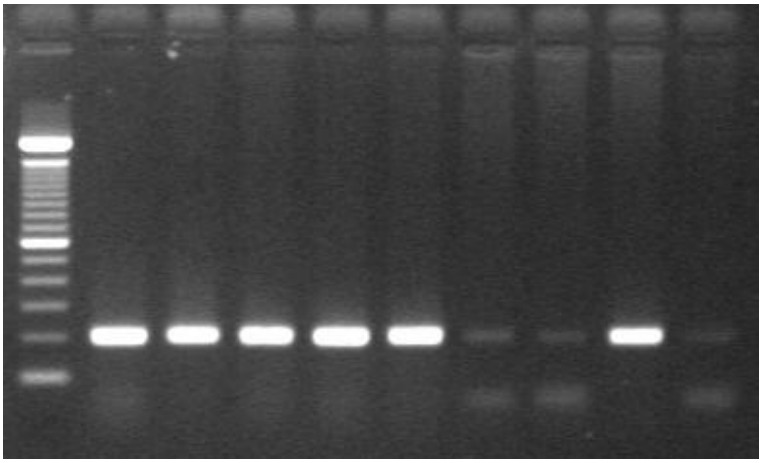


Figure 3.8: *Zfy2* marker. From left to right: 100 bp ladder, W1-9.

Btk is an X chromosome marker which tests for the B1 insertion of the Bruton agammaglobulinemia tyrosine kinase (Munclinger *et al*., 2002). *M. m. domesticus* mice carry this insertion (342 bp product) whereas *M. m. castanues* and *M. m. musculus* mice do not (206 bp product) . As for the other 3 nuclear markers, all N.Z. mice typed as *M. m. domesticus* for this marker. Figure 3.9 shows an example of these results.

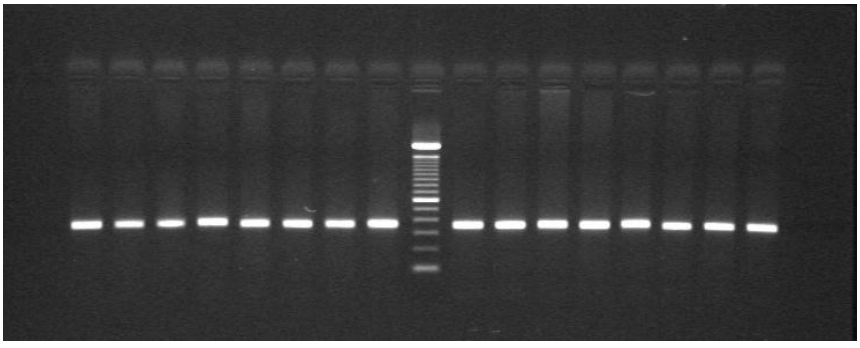


Figure 3.9: *Btk* marker. From left to right: X7, Wn1-7, 100 bp ladder, Wn8-15.

Chapter 4: Results - SNP Analysis of Hybrid Mice

4.1 Introduction

Seven representative male mice were submitted for high density 600K SNP microarray analysis using the JAX laboratory service. Because of the cost of each analysis (approximately USD1, 500 per sample) it was not possible to submit as many samples as one would wish. Thus, seven mice were chosen; two from deep within the hybrid zone (Te Anau and Taieri), one just south of the contact zone (Waianakarua), one in the contact zone (Mawaro) and three *M. m. domesticus* from the contact zone (Bluegum Park near Temuka), 50 km north of the contact zone (Lincoln) and the North Island (Hamilton). Although the number of samples is obviously limited, it was hoped that two of these would be representative of 'pure' hybrids , two of 'pure' *M. m. domesticus* and three from the contact zone that might exhibit introgression between the two. Male mice were chosen for analysis, as the the X-chromosome calls should be "homozygous" and give the unambiguous genotype of a single X-chromosome.

Once genomic areas of interest were identified from analysis of the SNP data, the intention was to analyse these small regions on a much larger set of samples using conventional PCR based SNP assays.

4.2 SNP micorarray results

4.2.1 Quality of SNP data.

The DNA for analysis was submitted in two batches; three in the first batch (September 2009) in and four in the second (February 2010).

As part of their service, the JAX laboratory provided data sheets for each analysis that gave comprehensive statistics summarising quality control parameters, call rates for SNPs and comparisons of the alleles called for the wild mice with those of a reference laboratory mouse strain C57B6/J which is largely of *domesticus* origin. The following information is from the summary sheet for the first three wild-trapped mice we submitted (the second summary sheet was virtually identical).

The APT software generates a set of summary statistics for each sample. A subset of these statistics are shown in Table 2. An overview of the calls for good-quality SNPs (581,672 total) for each sample are summarized in Table 3.

Sample	Overall Call Rate	Heterozygous Call Rate	Homozygous Call Rate
J2	98.29869	13.9997	84.29899
W3	98.16437	15.56413	82.60024
X1	98.20087	15.07504	83.12583

Table 2: Selected summary statistics for each sample.

Sample	SNPs with Allele A	SNPs with Allele B	Heterozygous SNPs	Uncalled SNPS
J2	397,488	100,150	75,533	8,501
W3	394,529	92,638	85,233	9,272
X1	396,818	93,508	82,234	9,112
Common to All Samples	321,005	45,211	12,090	480

Table 3: Number of SNPs with particular alleles among samples. The last row indicates the number of SNPs with particular alleles across all samples.

It can be seen that the overall call rates are extremely good, indicating that the quality of the submitted DNA was high. These call rates compare very favourably with those achieved by

the Jackson Laboratory themselves, using inbred laboratory strains (Frazer *et al*., 2007). An important point to note is that all three samples are near identical with respect to call rates for allele A, for homozygous call rates and for alleles that are common for all three samples. In other words, at this global analysis level, there are no striking differences between the nuclear genomes and they are predominantly *domesticus* in nature.

4.2.2 X and Ychromosome SNP data

Because all JAX submitted mice were males, the SNP calls for the X-chromosome should be called as homozygous. As can be seen in Appendix IV this is not the case and some 5-10%, depending on the region, are called as heterozygous. This occurs because of poor quality control in designing the original 600K chip (de Villena, personal communication) and indicates that hybridisation of mouse DNA from some other genomic region is also occurring on the "X" SNP specific oligonucleotides. It is probable that similar error levels also apply to autosomal SNPs, but the difference between an authentic and an artefactual heterozygote is not distinguishable.

Unfortunately, the 600K chips did not yield reliable Y-chromosome SNP calls; this is an inherent problem with the chips, both in design and in the programmes that interrogate the raw SNP signals.

4.2.3 Mitochondrial SNP calls

All seven samples were classified correctly as either *castaneus* or *domesticus* (data not shown, but D-loop sequencing results can be found in section 3.2 which confirm the mtDNA genotype of the seven Jax submitted samples).

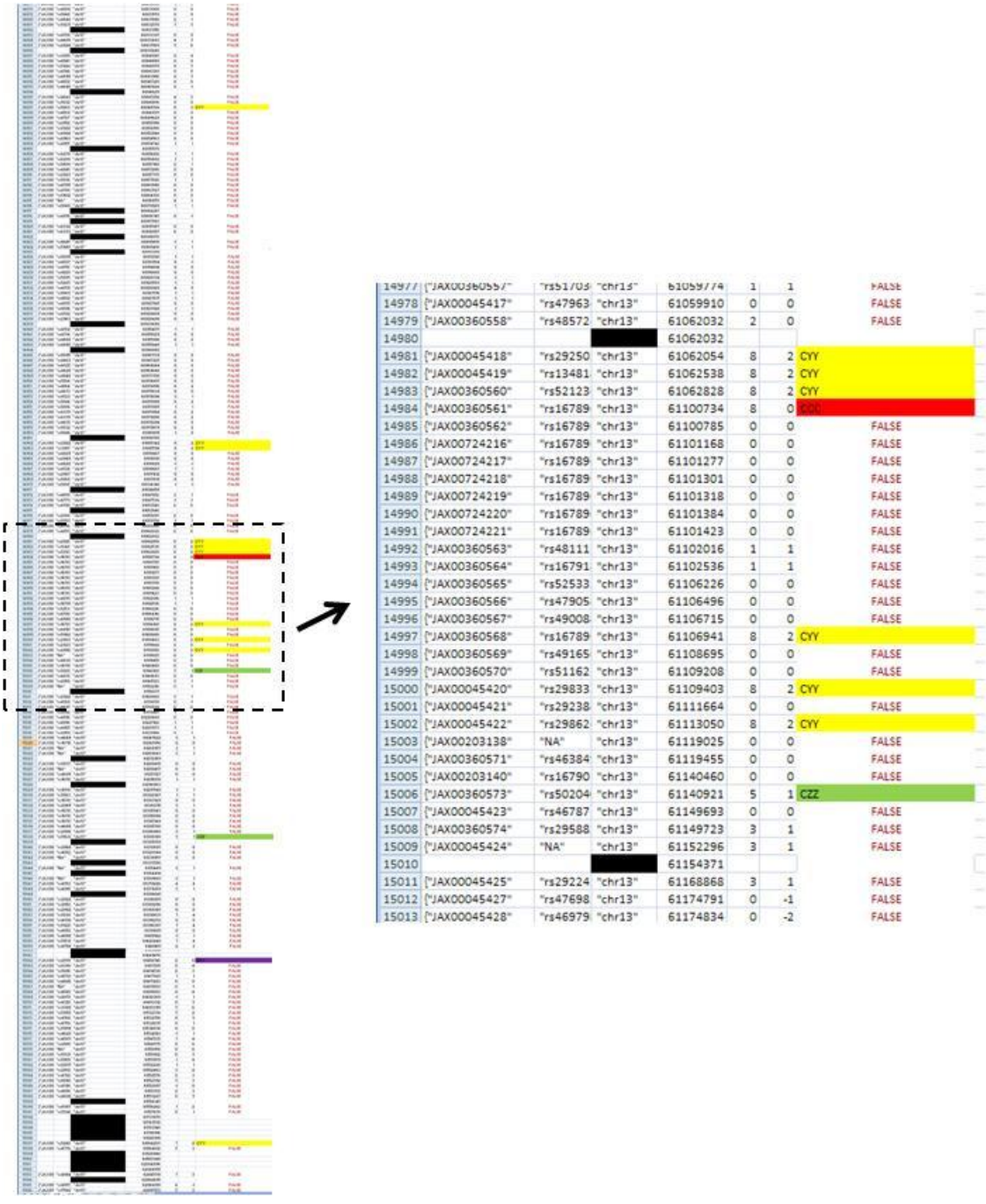
4.2.4 Detailed Analysis of SNP data chromosome by chromosome

Based on the global analysis, it was clear that any residual *castaneus*-like genome fragments were going to be small, perhaps only extending over a million base pairs (~1 CentiMorgan) or even less. Moreover, other unknowns have to be factored into this expectation. For instance, the path to the establishment of the hybrid mice is uncertain – thus while it is clear that female *M. m. castaneus* mice must have originally extensively backcrossed into *M. m. domesticus* mice, subsequent interbreeding amongst these hybrids must have occurred in the absence of pure *M. m. domesticus* mice, especially in the southern South Island. Thus it is possible that remnants of *M. m. castaneus* genomes may have remained by chance and been rendered homozygous at a later stage by this 'interbreeding' or alternatively, certain regions that gave a selective advantage to hybrids (in the south) have been retained and, again, rendered homozygous.

Because of these considerations, and to simplify the analyses, the initial analysis of the SNP data was designed to pinpoint regions of homozygosity in the hybrids that differed (normally by exhibiting allele B, but occasionally allele A) from the corresponding homozygous region in the *M. m. domesticus* wild type mouse (WSB/EiJ). Two simple scanning programmes were used to examine the data, chromosome by chromosome, one based on Mathematica (see Appendix II) and the other on adapting Excel displays of the data (Appendix III on disk). Appendix VI (on disk) is one of the initial Excel files used to interrogate the data from the first batch of samples (the formulae used for this can be seen by clicking on the relevant cells) based on homozygous SNPs, but this was later modified to that in Appendix III.

In retrospect, this approach was overly simplistic for several reasons, especially as the first analyses were based on just three animals, J2 a hybrid from some 50km south of the contact zone, W3 a hybrid in the contact zone and X1, a *domesticus* type mouse in the contact zone. Nevertheless, by simply displaying as colour bars on an Excel chart those SNPs for which J2 and W3 both had allele '2' whereas X1 had allele '0', small chromosomal clusters of allelic differences between hybrid and *domesticus* wild mice were highlighted. These often fell within just one haplotype block (visualised in the Excel charts as black horizontal bars, derived from the Perlegen database at http://mouse.cs.ucla.edu/perlegen/). By using coloured bars to represent maximum differences (red), very marked differences (orange), marked differences (yellow) and differences (green) and setting the Excel zoom on 10% it is possible to scroll down each chromosome in a few minutes, with each screen spanning 1 to 2 Mb. An example is shown in Figure 4.1 for a region of Chromosome 13. Many such clusters were found on all chromosomes, for instance some 15 on Chromosome 13.

The number of clusters is far in excess of what one would expect by chance. For example, Chromosome 13 contains almost 30,000 SNPs, of which some 440 were scored as red, orange or yellow, and if one arbitrarily divides the chromosome into 1500 blocks of 20 SNPs, one can estimate the probable number of clusters of 6 or more highlighted SNPs occurring in a block by chance using an approximation of a 'balls-in boxes' type calculation. The probability of one such SNP being in a given block is ~ 0.3, the chance of six is ~ 6.3×10^{-4}, so one would only expect one such cluster of six SNPs in one block to occur by chance amongst the 1,500 blocks in Chromosome 13. In fact, there are at least 15.



14978	"JAX00045417"	"rs47963	"chr13"	61059910	0	0		FALSE
14979	"JAX00360558"	"rs48572	"chr13"	61062032	2	0		FALSE
14980			[redacted]	61062032				
14981	"JAX00045418"	"rs29250	"chr13"	61062054	8	2	CYY	
14982	"JAX00045419"	"rs13481	"chr13"	61062538	8	2	CYY	
14983	"JAX00360560"	"rs52123	"chr13"	61062828	8	2	CYY	
14984	"JAX00360561"	"rs16789	"chr13"	61100734	8	0	CZZ	
14985	"JAX00360562"	"rs16789	"chr13"	61100785	0	0		FALSE
14986	"JAX00724216"	"rs16789	"chr13"	61101168	0	0		FALSE
14987	"JAX00724217"	"rs16789	"chr13"	61101277	0	0		FALSE
14988	"JAX00724218"	"rs16789	"chr13"	61101301	0	0		FALSE
14989	"JAX00724219"	"rs16789	"chr13"	61101318	0	0		FALSE
14990	"JAX00724220"	"rs16789	"chr13"	61101384	0	0		FALSE
14991	"JAX00724221"	"rs16789	"chr13"	61101423	0	0		FALSE
14992	"JAX00360563"	"rs48111	"chr13"	61102016	1	1		FALSE
14993	"JAX00360564"	"rs16791	"chr13"	61102536	1	1		FALSE
14994	"JAX00360565"	"rs52533	"chr13"	61106226	0	0		FALSE
14995	"JAX00360566"	"rs47905	"chr13"	61106496	0	0		FALSE
14996	"JAX00360567"	"rs49008	"chr13"	61106715	0	0		FALSE
14997	"JAX00360568"	"rs16789	"chr13"	61106941	8	2	CYY	
14998	"JAX00360569"	"rs49165	"chr13"	61108695	0	0		FALSE
14999	"JAX00360570"	"rs51162	"chr13"	61109208	0	0		FALSE
15000	"JAX00045420"	"rs29833	"chr13"	61109403	8	2	CYY	
15001	"JAX00045421"	"rs29238	"chr13"	61111664	0	0		FALSE
15002	"JAX00045422"	"rs29862	"chr13"	61113050	8	2	CYY	
15003	"JAX00203138"	"NA"	"chr13"	61119025	0	0		FALSE
15004	"JAX00360571"	"rs46384	"chr13"	61119455	0	0		FALSE
15005	"JAX00203140"	"rs16790	"chr13"	61140460	0	0		FALSE
15006	"JAX00360573"	"rs50204	"chr13"	61140921	5	1	CZZ	
15007	"JAX00045423"	"rs46787	"chr13"	61149693	0	0		FALSE
15008	"JAX00360574"	"rs29588	"chr13"	61149723	3	1		FALSE
15009	"JAX00045424"	"NA"	"chr13"	61152296	3	1		FALSE
15010			[redacted]	61154371				
15011	"JAX00045425"	"rs29224	"chr13"	61168868	3	1		FALSE
15012	"JAX00045427"	"rs47698	"chr13"	61174791	0	-1		FALSE
15013	"JAX00045428"	"rs46979	"chr13"	61174834	0	-2		FALSE

Figure 4.1 An Excel display of a 2Mb region of Chromosome 13 showing a haplotype block (insert) with a preponderance of hybrid mouse SNP alleles which differ from those of *domesticus* wild mice. Black bars are haplotype block boundaries; red, yellow and green bars represent (in decreasing order) the strength of the allelic differences. "FALSE" simply means the differences are slight.

A somewhat more sophisticated approach to scanning chromosomes was also developed using a Mathematica programme (Appendix II). Each SNP allelic call for each animal (plus reference SNPs for Jackson Laboratory wild type strains PWD/PhJ (*M. m. musculus*), CAST/EiJ (*M. m. castaneus*) and WSB/EiJ (*M. m. domesticus*) were displayed with respect to chromosome position, in three bands – homozygous calls '2', heterozygous calls '1' and homozygous calls '0'.

Examples of the Mathematica scanning displays are shown in Figure 4.2 . As with the Excel displays, each chromosome can be scanned in a few minutes in order to pinpoint regions where hybrid and *domesticus* wild mice differ. Many such regions were detected (clusters of SNPs in allelic calls between the two wild mice groups) and, although they were unlikely to occur by chance, it was difficult to ascribe any compelling phenotypic (based on the properties of genes in these regions, as listed in the Jackson phenotypic database) properties to most of these regions. However, one region of real potential interest was identified (see Figure 4.3 and also the Excel display in Figure 4.1) on Chromosome 13 which is the maternal-fetal conflict locus (see Section 4.4). This region is involved with mediating interactions between mother and offspring, and is reported to be undergoing positive selection in mice (Chuong, Tong, & Hoekstra, 2010).

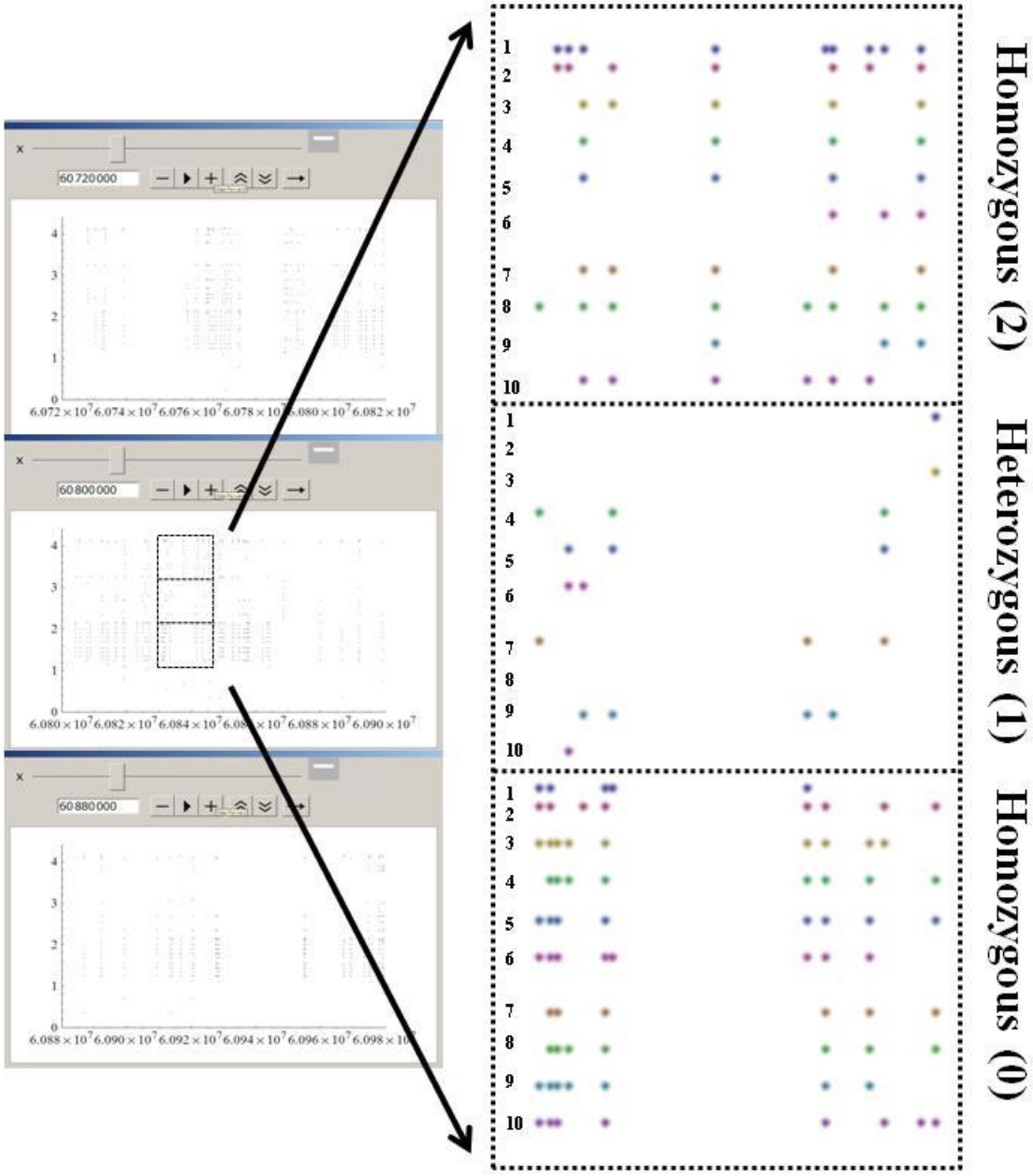


Figure 4.2 Mathematica representation of SNP calls. The three screen shots on the left show overlapping 1Mb regions of chromosome 13. A region around 60.84 Mb is expanded on the right. For each SNP, each mouse should either call as homozygous "2" (top), heterozygous "1 "(middle) or homozygous "0" (bottom). Dots are purposely made small so close SNPs can be resolved – alternatively the Mathematica scrolling cale can be expanded. (1) PWD/PhJ, (2) CAST/EiJ, (3) TA4, (4) Tr1, (5) J2, (6) W3, (7) X1, (8) Li2, (9) Ham1, (10) WSB/EiJ.

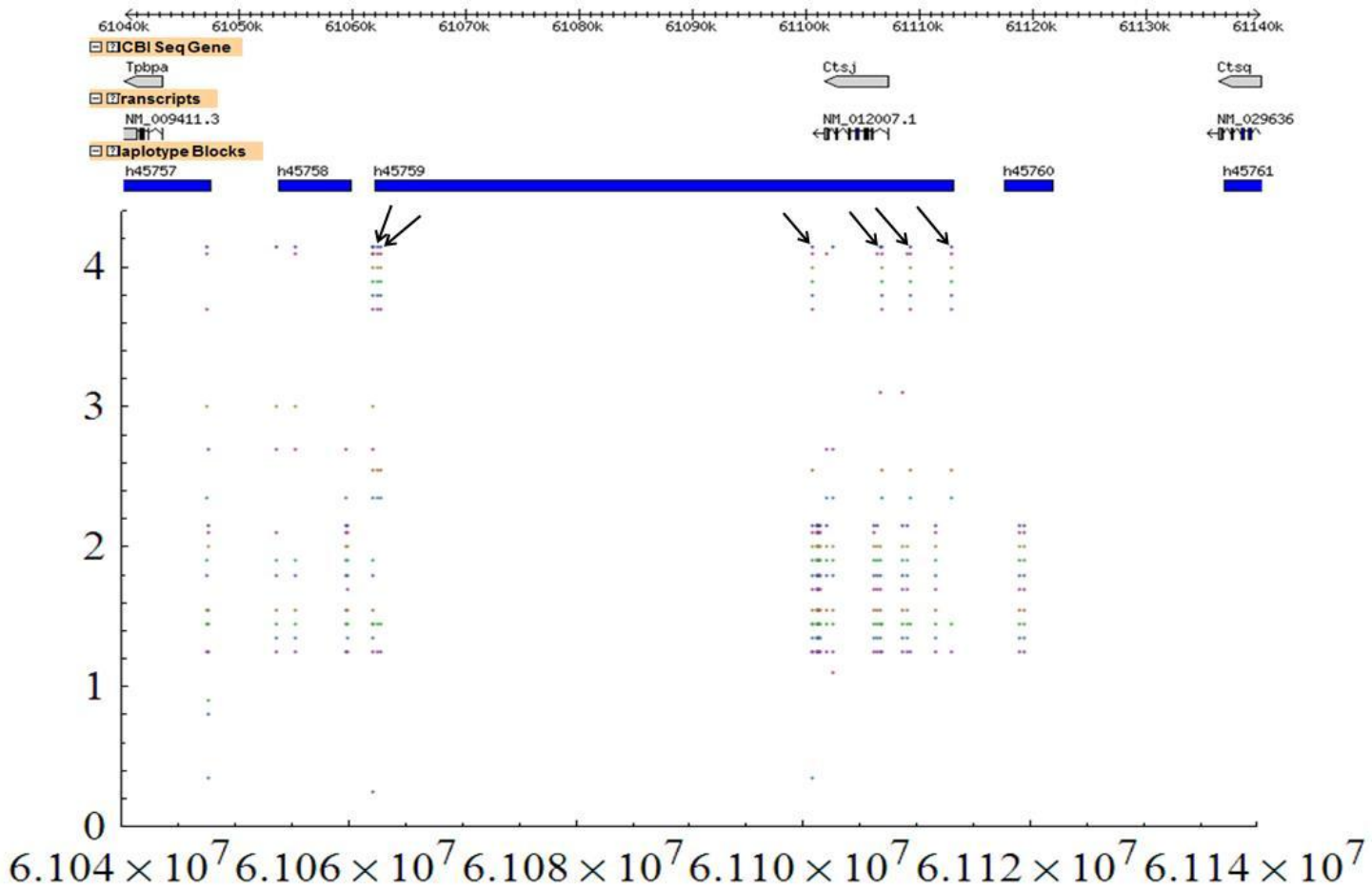


Figure 4.3: Chromosome 13 Region displaying five SNPs (arrowed) that are homozygous '2' for the four hybrid wild mice but not for either *domesticus* wild mice or WSB/EiJ. Perlegen genes and haplotype blocks are overlaid at top.

In a slightly different approach, Excel graphs were also plotted that displayed SNP regions that predominantly scored as '2' in the four hybrids (giving a maximum score of 8) or three *domesticus* (giving a maximum score of 6 – shown as negative for display purposes) wild mice. Again, obvious clusters of these homozygous regions can be seen – extending over a few Mbp in most cases (see Figure 4.4 for an example), and although the density of these regions did tend to correlate with regions of low recombination and, in some cases, gene densities, this was not always the case.

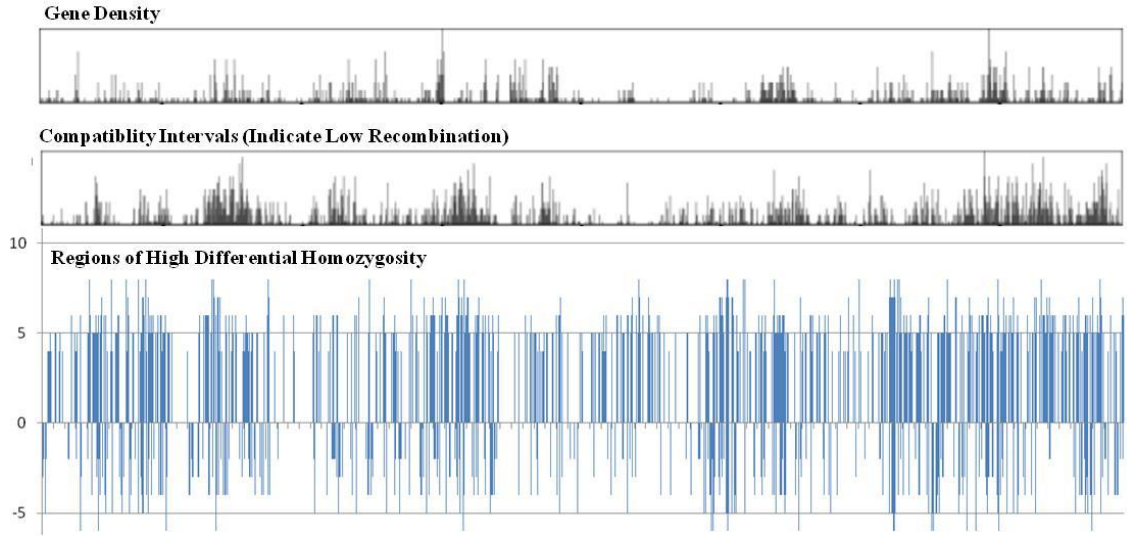


Figure 4 .4: Excel plot of Chromosome 1 displaying regions in which SNPs are highly homozygous in either the hybrid (maximum score 8) or *domesticus* (maximum score -6) wild mice. Gene density and compatibility intervals (histogram represents density of local areas showing no evidence of historical recombination) are taken from the Mouse Phylogeny Viewer.

Although analysis was also attempted by programming in R, this proved challenging because of the large SNP file sizes and subsequent problems with data input and was abandoned due to time constraints. The code used for this analysis, is provided in Appendix VII.

For the Mathematica display (Figure 4.2), the allele calls are shown in the order PWD/PhJ (*M. m. musculus* inbred wild type strain), CAST/EiJ (*M. m. castaneus* inbred wild type strain), Te Anau, Taieri, Waianakarua, Mawaro, Temuka, Lincoln, Hamilton , WSB/EiJ (*M. m. domesticus* inbred wild type strain). The SNP calls that are in the top group of 10 are those that are homozygous for allele B, the middle group represent heterozygous A/B calls and the bottom group homozygous A/A calls.

In essence, in visually scanning the chromosomes, one looks for regions in which the homozgosity in the 4 hybrid N.Z mice (Te Anau, Taieri, Waianakarua and Mawaro) is different from the 3 N.Z *domesticus* mice (Temuka, Lincoln andHamilton) and also (but not necessarily) more similar to CAST/EiJ than to WSB/EiJ.

Regions satisfying these criteria exist on chromosomes 1, 3, 9, 13, 17, 19 and X (see Table 4.1 and section 4.4). However, relatively few regions of this nature were detected – only one or two per chromosome, and most did not encompass genes with strong phenotypic properties and therefore could well represent chance SNP fluctuations. Only one region of real potential interest was identified which was the maternal-fetal conflict locus on chromosome 13 (see section 4.4).

Likewise, when the data was analysed in a similar manner using an Excel graphical display, one or two regions of potential interest were found on all chromosomes, but again, these only encompassed ~ 1 Mbp. Interestingly, when these regions are plotted against histograms of haplotype and gene block density, most occur in regions of high density. The biological relevance of this is not clear.

4.3 Validation of SNP array analysis results by wider screening

Although the biological relevance of the 'clusters' described in the last section was uncertain, it was important to validate SNP data, both with respect to microarray SNP calls and to extend data to a wider range of N.Z. mice. In order to do this, SNP assays were developed for selected chromosomal regions (Table 4.1). Twenty eight additional mice were analysed in this way for each region (see methods section 2.13).

In all cases, the RFLP SNP assays gave the same calls as the SNP microarrays; moreover, the hybrid and *M. m. domesticus* mice in the larger sample set gave the same grouping calls as did the original 7 mice (figure 4.5, figures 4.6 A and B, figures 4.7-9.)

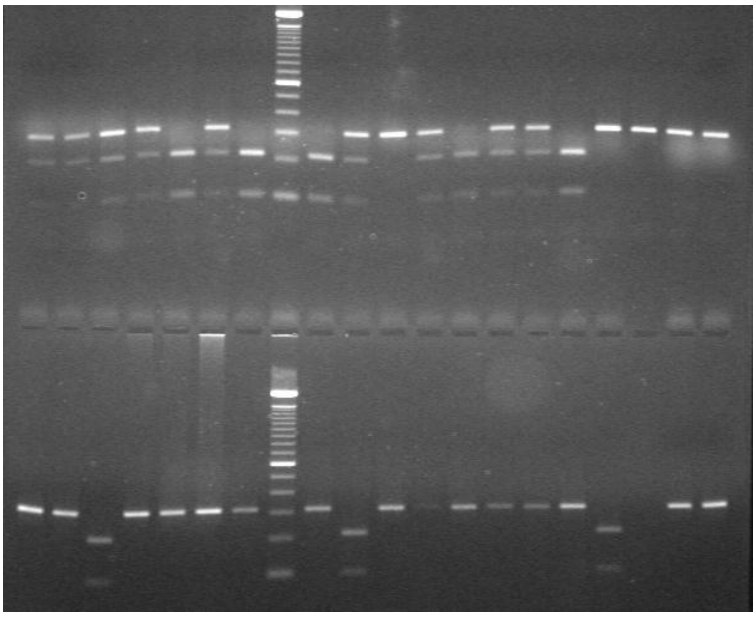


Figure 4.5: Chr3/HindII assay. Top row: Ham1, Li1-6, (100 bp ladder), X1-8, TA1-4. Bottom row: TA5 Tr1-2, J1-3, Wn1, (100 bp ladder), Wn2, Wn8, Wn15, Wn34, Wm1, Wm2, Wm5, W1, W9, X2 uncut control, X2 uncut control. *Dometicus*-like: 216 bp, 99 bp. *Castaneus*-like: 315 bp.

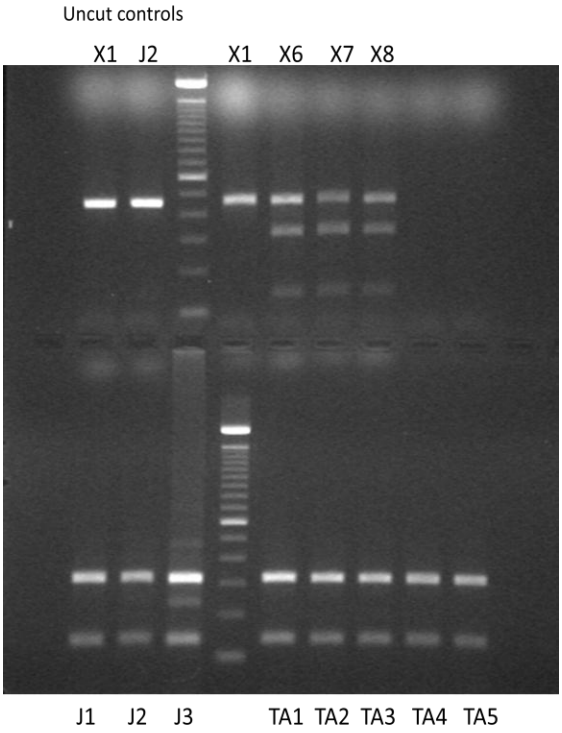


Figure 4.6 A: Chr9/*Mva*I assay. A: *Domesticus-like*: 459 bp (X1), *castaneus*-like: 317 bp, 142 bp (J1-3, Ta1-5). X6, X7 and X8 are heterozygous.

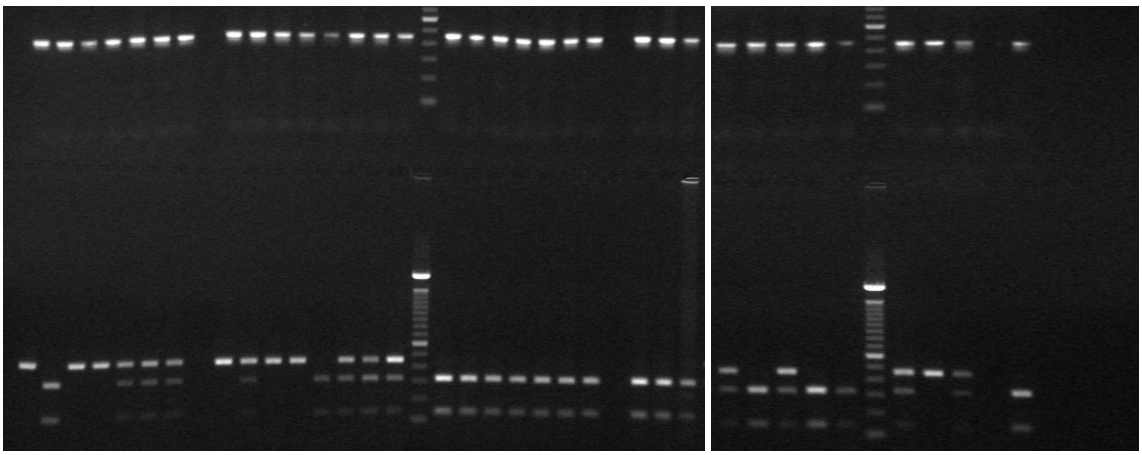


Figure 4.6 B: Chr9/*Mva*I assay on larger sample set. Top row: undigested PCR products of Ham1, Li1-6, X1-8, (100 bp ladder), TA1-5, Tr1-2, J1-3, Wn1, Wn2, Wn8 , Wn34, Wn15,(100 bp ladder), Wm1, Wm2, Wm5, W1, W9. Bottom row: Corresponding *Mva*I digests.

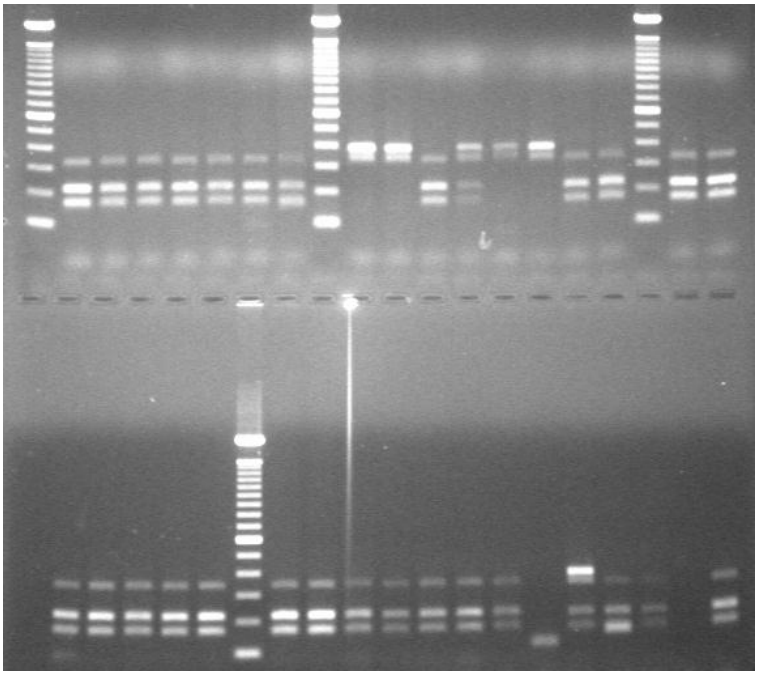


Figure 4.7: Chr17/*Hae*II assay. Top row: Ham1, Li1-6, 100 bp ladder, X1-8, 100 bp ladder, TA1, TA2. Bottom row: TA3-5, Tr1-2, 100 bp ladder, J1-3, Wn1, Wn2, Wn8, , Wn34, Wn15, Wm1, Wm2, Wm5, W9. *Domesticus*-like samples give band at 388bp, *Castaneus*-like samples give bands at 220 and 168 bp. Band at 320 bp is presumed to be single stranded DNA.

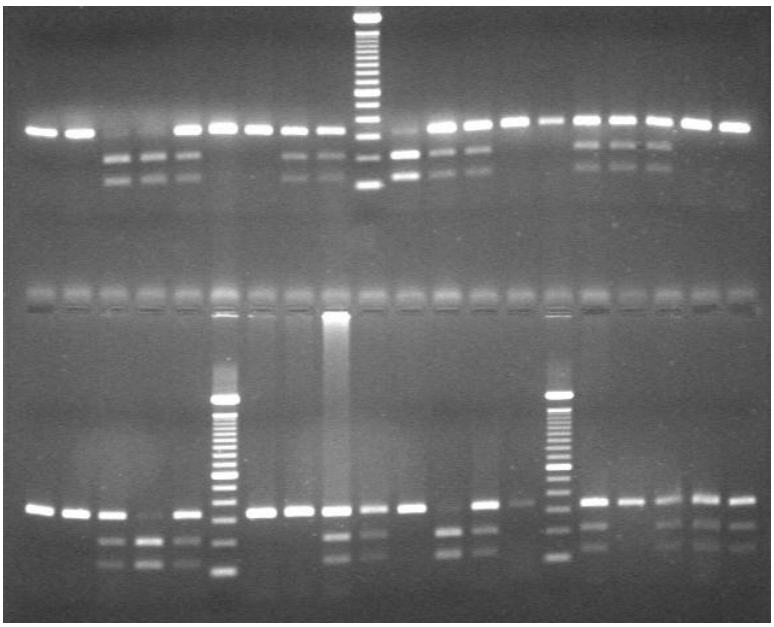


Figure 4.8: Chr19/*HinF*I assay. Top row: uncut control (2 lanes) Ham1, Li1-6, (100 bp ladder), X1-8, TA1-2. Bottom row: TA1-5, Tr1-2, (100 bp ladder), J1-3, Wn1, Wn2, Wn8, , Wn34, Wn15, (100 bp ladder), Wm1, Wm2, Wm5, W1, W9. *Domesticus*-like: 207 bp, 141 bp, *castaneus*-like: 348 bp.

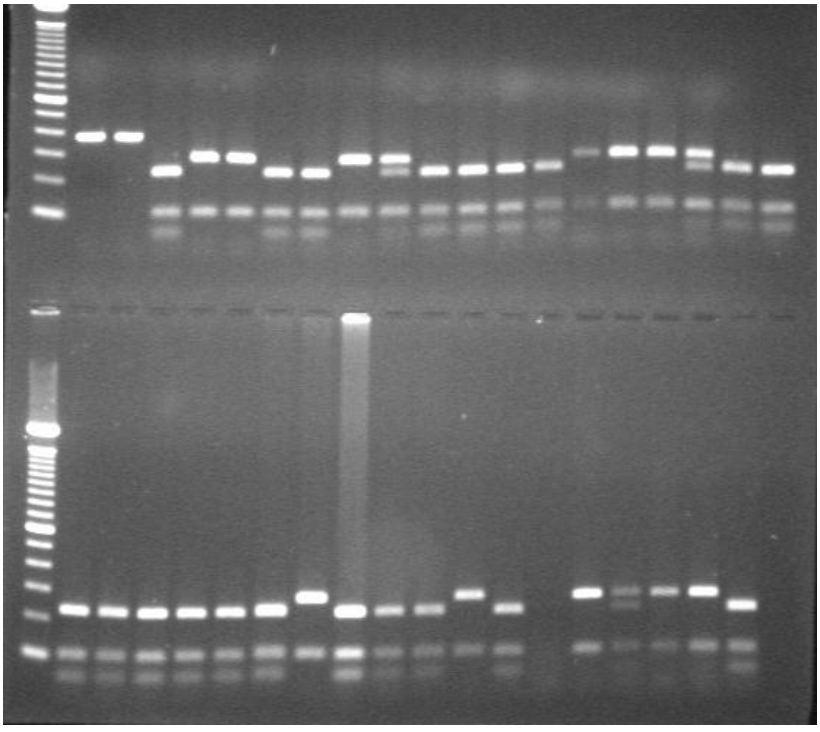


Figure 4.9: ChrX/Tru91 assay. Top row: X1 (uncut control), TA1 (uncut control), Ham1, Li1-6, X1-8, TA1, TA2. Bottom row: TA3-5, Tr1-2, J1-3, Wn1, Wn2, Wn8, Wn15, Wn34, Wm1, Wm2, Wm5, W1, W9. *Domesticus*-like: 228 bp, 102 bp, 51 bp. *Castaneus*-like: 279 bp, 102 bp.

Table 4.1

SNP RFLP screening assays						
Genomic location		**RefSNP no.**	**Character state (cast/dom)**	**PCR primers**	**Enzyme**	**Expected fragment sizes (bp)**
Chr3	nt132,212,362	rs30730244	A/C	C3F TGTTGGGTTTTCCCTTTCTG C3R GGGGCACACTATGGTATTGG	*Hind*II	dom: 99, 216 cast: 315
Chr9	nt58,783,133	rs29785650	A/G	C9F GTATCCCTGGCTGTTCCAGA C9R CTTGATGGCGTGTGTGACTT	*Mva*1	dom: 459 cast: 142, 317
Chr17	nt92,141,027	rs48139034	C/T	C17F TGCATATCAATGTGCATTTTTG C17R TGGGAAGTACAGGCAAGAGG	*Hae*III	dom: 388 cast: 220, 168
Chr19	nt13,053,064	Rs38152526	T/G	C19(2)F CCTGACCAAAATTTCCCAGA C19(2)R TCTTTCTTCCACAAGCTGTTTTT	*HinFI*	dom: 141, 207 cast:348
ChrX	nt118,188,775	rs29107743	G/A	CXF GGGGAGCACCCTCATAGAAT CXR TGAGATTTTGTCACCTCATGAAA	*Tru*91	dom: 102, 51, 228 cast: 279, 102

4.4 Chromosome 13 maternal-fetal conflict region

Particular interest focussed on the maternal-fetal conflict gene region on Chromosome 13 that contains a cluster of placental genes involved with mediating interactions between mother and fetus (see figure 4.3). Chuong *et al*. (2010) found evidence of positive selection in one of these genes, namely *Tpbpa,* within the *mus* lineage. Sequencing of this gene was carried out using the five primers described in Chuong *et al*. (2010) on the mice Li2 and TA4, from the north of and deep within the hybrid zone, respectively. Surprisingly, *Tpbpa* in Li2 was found to be identical at the nucleotide (and thus the protein) level to that of *M. m. molossinus* from Kyushu Fukuoka, Japan (NCI colony, M. Potter, GenBank accession number GU294771.1) For TA4, the translated amino acid sequence was identical to that of *M. m. musculus* from Deventz, Bulgaria (collected by Wenfei Tong, GenBank accession number GU294772.1) but the nucleotide sequence differed at 4 out of 2445 nucleotides sequenced. Edited sequence files for this gene can be found in Appendix VIII on attached disk.

Examination in the Mouse Phylogeny Viewer of the Ch13 60.9 to 61.2Mb region reveals a *M. m. castaneus*-type SNP motif in the Jackson laboratory *M. m. molossinus* wild type strain MOLF/EiJ which has been bred from M. Potter's NCI colony. The Li2 mouse is heterozygous for this same *castaneus*-like region (data not shown) but can be viewed in the relevant Excel chromosome displays in the Appendix III.

4.5 Analyses using SNPs diagnostic for mouse sub-species

Although some interesting chromosomal regions of hybrid mice were identified in the SNP analyses described above, it quickly became apparent that most SNPs on the 600K chip were not sub-species specific. This ambiguity thwarted the main aim of the study, namely, pinpointing remnants of *M. m. castaneus* genomes in hybrid mice.

There are two reasons for this ambiguity. First, the wild type reference strains maintained by the Jackson laboratory, CAST.EiJ, WSB.EiJ and PWD.EiJ are not, respectively, pure *M. m. castaneus*, *M. m. domesticus* or *M. m. musculus* , with each contaminated with appreciable portions of the other genomes (see Mouse Phylogeny Viewer: http://msub.csbio.unc.edu/). Second, and perhaps more importantly, many SNP alleles are non-diagnostic for the simple reason that the allele (a '0' or a '2') is shared by two, or all three subspecies. This latter problem is readily apparent when one peruses the Excel files (Appendices III and VI) Thus, even though the CAST.EiJ (and CASA.RiJ) genomes are predominantly *M. m. castaneus* (shaded green in the Mouse Phylogeny Viewer), many of the SNP alleles in these regions are shared by *M. m. domesticus* or *M. m. musculus,* or both.

Because of these serious limitations, contact was made with Professor Fernando Pardo-Manuel de Villena, Department of Genetics, University of North Carolina School of Medicine, one of the designers of the 600K SNP array (Frazer *et al*, 2007). He confirmed the limitations of the array, but very generously offered access to the Mouse Phylogeny Viewer before it was publicly available and, more recently, supplied a database which consists of a subset of SNPs which are diagnostic for each of the three mouse subspecies. As outlined in Chapter 2, de Villena and co-workers (Yang *et al*., 2011) analysed 36 wild trapped mice to create this database. Ten of these wild mice were *M. m. castaneus* trapped in the States of

Uttarakhand and Himachal Pradesh in northwestern India (Baines and Harr, 2007, as listed in Geraldes *et al*., 2008). The ten *M. m. domesticus* were predominantly from Italy and Spain, and the sixteen *M. m. musculus* from Slovak, Hungary and Poland (Geraldes *et al*, 2008).

The SNPs which were diagnostic were extracted from the N.Z. mouse databases and used to generate chromosome displays. Graphical representations are shown in Appendix IV; Excel data sheets, along with interactive high resolution charts that can be found in Appendix III on the included disk. An example is shown in Figure 4.10 for Chromosome 3. Data from each mouse is plotted separately, in a band one diagnostic unit in height, with the homozygous calls in a group above a group of heterozygous calls. For reference, the SNPs for the domesticus wild type strain WSB/EiJ are also plotted. Thus, the data is in two groups of eight bands, with the four hybrid mice at the top of each, the three domesticus type mice next and WSB/EiJ at the bottom. The order of the mice, top to bottom, is geographically from the south-west of N.Z. (Te Anau) to the north (Hamilton) shown on the displays as S➔N (although, strictly speaking it is SW➔N). *Domesticus* markers are blue, *castaneus* green and *musculus* red. Several important observations can be made from these data. These will be dealt with in turn.

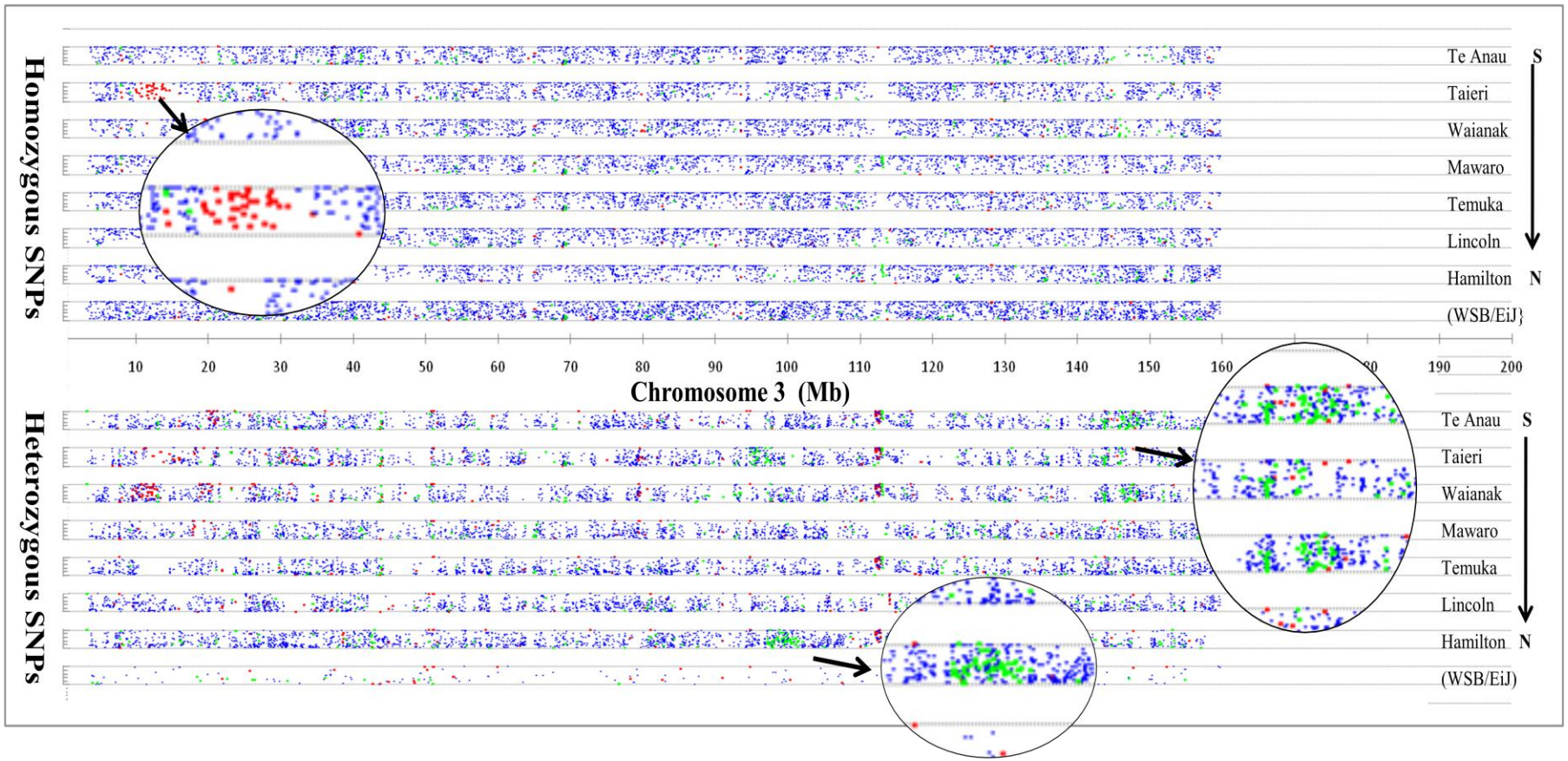


Figure 4.10 : Example of SNP analysis results in Excel, based on a Mathematica programme (Appendix II). Blue: *domesticus*, red: *musculus*, green: *castaneus*. Magnified regions are possible haplotype blocks.

4.5.1 All seven N.Z. mouse genomes are predominantly '*domesticus*'

Only small regions of 'non-*domesticus*' genome were observed in any of the N.Z. mice chromosomes. It was estimated that, overall, at least 95% of the genomes were *domesticus* in origin. The exact percentage is difficult to compute because many more *domesticus* than *musculus* and *castaneus* diagnostic alleles are in the database. Thus, for example, each of the Chromosome 3 displays in Figure 4.10 consists of only 14,400 SNPs as the remaining 25,000 SNPs on the 600K array are not diagnostic. Moreover 56% of these SNPs are diagnostic for *domesticus*, 26% for *musculus* and only 18% for *castaneus*. This tends to distort the display in favour of *domesticus* (partially allowed for by making the blue markers smaller in size than the red for *musculus* and green for *castaneus*). It also means that the resolutions are much lower than for the original full set of SNPs, namely around one SNP per 20 kb for *domesticus*, one per 42 kb for *musculus* and one per 61 kb for *castaneus* diagnostic SNPs. This resolution is further degraded in some heterozygous regions (see 4.5.2) where the SNPs may be 'shared' between the two alleles.

4.5.2 Heterozygosity varies markedly amongst the seven N.Z. mice

Only the Jackson wild type mouse WSB/EiJ is completely inbred, thus displaying zero heterozygosity (no heterozygous SNP calls) at all loci on all chromosomes (the low levels of noise in the heterozgosity band arise from artefactual erroneous calls for some SNPs on the 600K array – see 4.5.4). Amongst the wild mice, only 'Temuka' displays a high level of inbreeding with 10 of the 19 autosomes homozygous over 10 to 50% of their length. Some homozygous regions were also detected in Te Anau, Mawaro and Lincoln mice, but only on three or four chromosomes in each case. The Hamilton mouse displayed a completely

different pattern of homozygosity/heterozygosity, thus while only one overt region of homozygosity was observed (approximately 10% of Chr 4), it is obvious that all of Chr 6 to 17 display low overall levels of heterozygosity.

Overall, it would appear that six of the wild mice come from reasonably heterozygous populations (data comparable to that for wild mice in the Mouse Phylogeny Viewer) with the Temuka mouse obviously inbred. The Hamilton mouse, although not highly inbred (only one small homozygous region) comes from a population that is significantly more homozygous than that in the other six geographical regions.

4.5.3 Many small chromosomal regions are '*castaneus*' or '*musculus*'-like in composition, but somewhat randomly distributed

Surprisingly, virtually all chromosomes of all mice have distinct regions that are *castaneus* or *musculus*–like in nature. Many of these are small (just a few Mb or less) but a few extend over 10 Mb. Just as surprising, and counter to my original hypothesis, there was no preponderance of *castaneus* regions in the hybrid mice. Just as many were seen in the northern mice and, indeed, an appreciable number of *musculus* regions were seen in the hybrids. For example, in Figure 4.10, an expanded region around 10 to 14 Mb of Chr 3 for the Taieri mouse is shown which appears near homozygous for *musculus*; the same region is heterozygous for this *musculus* region in the Waianakarua mouse (the other 'allele' in this regions is *domesticus*). There is no evidence of this region in other mice. Also highlighted in Figure 4.10 is a *castaneus* region around 145 Mb that is heterozygous in the Te Anau, Taieri and Waianakarua mice, and another *castaneus* region around 100 Mb in just the Hamilton mouse.

There is also a narrow (~1 Mb) region at 103 Mb that is *musculus*-like, an allele that is present in five of the seven mice.

4.5.4 Some large chromosomal regions appear to have complex origins

Contrary to expectations, Ham1 appears to be different from all other six wild type mice both in exhibiting somewhat lower heterozygosity overall, and also in containing numerous large chromosomal regions of non-*domesticus* origin. Most of these regions were homozygous and extended over 10 Mb or more. Notable examples are in Chromosomes 6, 8, 11, 14 and 16. On closer examination, the structure of these regions is quite complex and, for example, the sub-species origins of the region at the distal end of Chr6 in Ham1, from 133 to 150 Mb (Figure 4.11) contains tracts that can be labelled as *domesticus*, *castaneus* or *musculus*. Interestingly, when this region is lined up with a range of Jackson laboratory strains, striking similarities are apparent. Virtually all of the laboratory strains have 'acquired' the 142.46 to 143.43 Mb castaneus region and although the WSB/EiJ and PWD/PhJ wild laboratory strains (along with all the wild trapped *M. m. musculus* and *domesticus* mice – not shown) remain true to sub-species origins, there are also considerable rearrangements of CAST/EiJ, CASA/RkJ and MOLF/EiJ chromosomes in this region.

Taken together, the boundaries between the various coloured segments of the mice shown in Figure 4.11 define the haplotype blocks in this region; clearly, interbreeding between *M. m. domesticus, castaneus and musculus* wild mice has led to the haplotype block pattern in

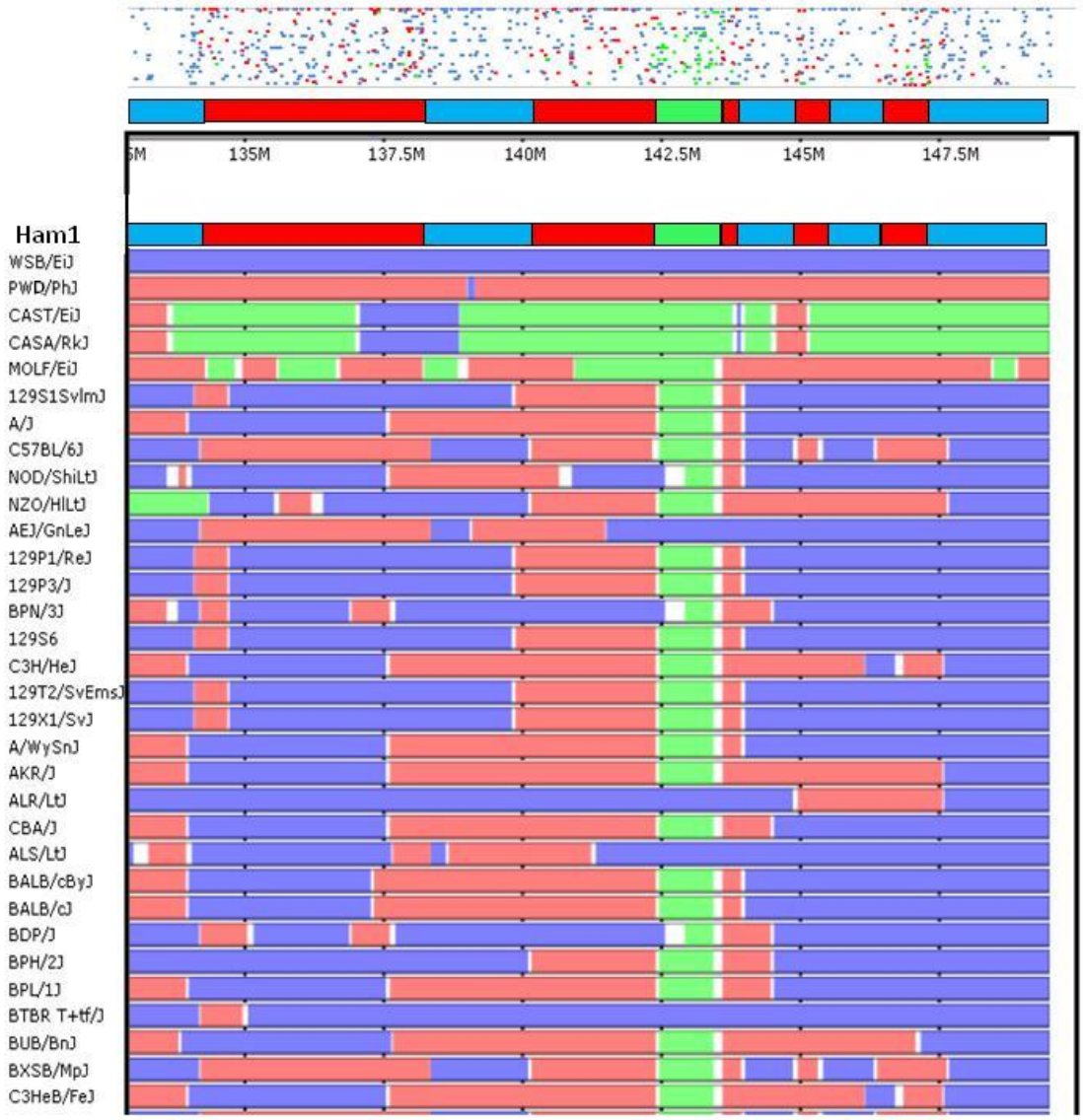


Figure 4.11: Comparison of subspecies origin of distal region of Chr6 in Ham1 and various Jackson laboratory strains. Homozygous diagnostic SNP markers for Ham1 are displayed at top, along with coloured bars showing the predominant subspecies origin of various segments. Ham1 and the reference lab strain C57BL/6J are very similar.

Ham1. Intriguingly, the actual composition of this region is very similar to that in many reference or Jackson Lab mouse strains, notably C57BL/6J. There is no obvious explanation for this similarity, but it would seem highly unlikely that two independent breeding scenarios (one in laboratories in the Northern hemisphere, the other in the wild in New Zealand),

would give such similar patterns without some selective pressures. It is relevant to point out that all of the laboratory mice in the Jackson laboratory collection with this Chr 6 pattern were originally bred from fancy mice, and that these partly had their origins in East Asia, including contributions from *M. m. molossinus* mice (Figure 4.12). If the Chr 6 region in the laboratory strain MOLF/EiJ (Figure 4.11) is representative of *M. m. molossinus*, it is conceivable that this region originated in wild mice in mainland Asia (*M. m. molossinus* is a hybrid of *M. m. castaneus* and *M. m. musculus*) and such mice may have been transported to Australasia with early shipping.

Additionally, other complex origin regions of Ham1 chromosomes also have striking similarities to corresponding regions in laboratory mouse strains, in particular C57BL/6J. Thus, two regions of Chr 14 (46 to 53 Mb and 88 to 101 Mb) and three regions of Chr 8 (30.6 to 44.4; 70.7 to 74.8 and 90.5 to 94.9 Mb) can be aligned with many of the mice strains displayed in Figure 4.11.

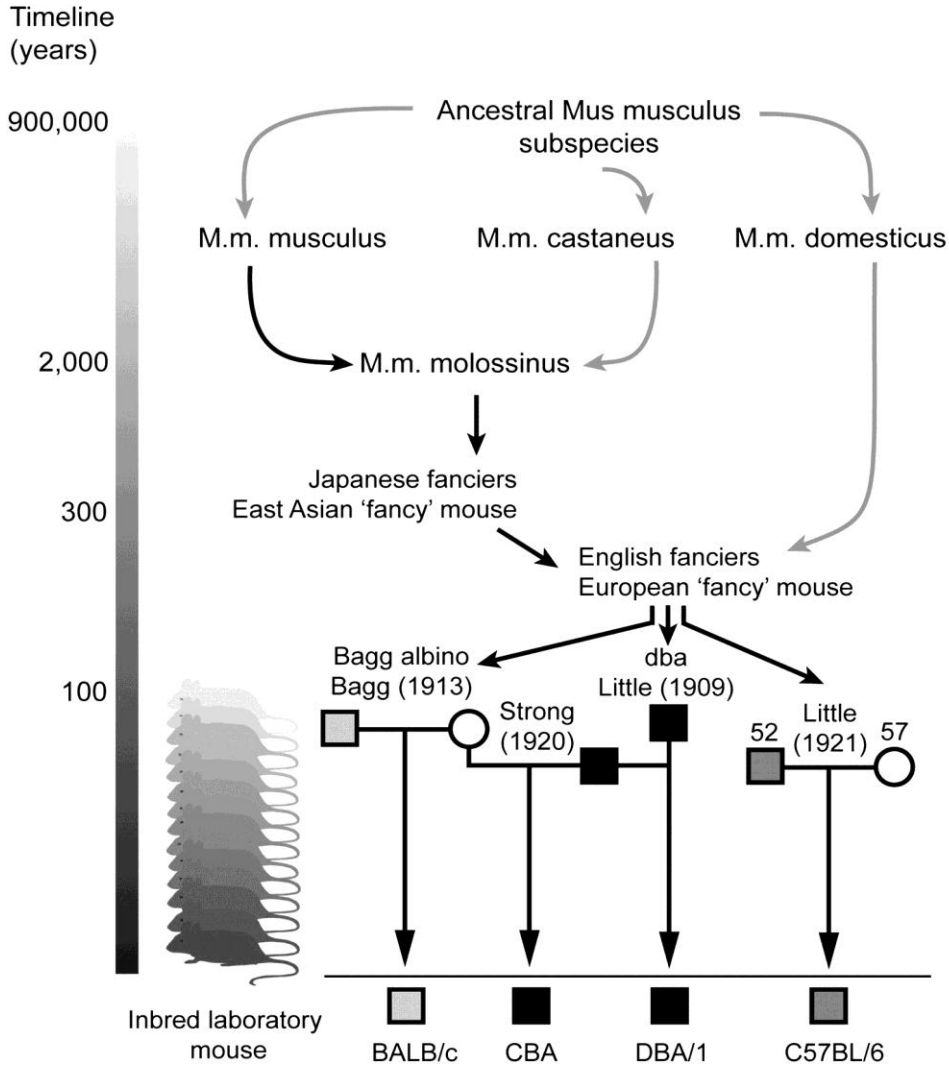


Figure 4.12: Origins of inbred laboratory mice. Most were bred from fancy mice that had both ill-defined East Asian and European origins. (Pertile, Graham, Choo, & Kalitsis, 2009).

4.5.5 'Candidate genes' cannot be categorically linked to specific *castaneus* and *musculus* chromosomal regions

Making correlations between chromosomal regions that have a sub-species signature and specific genes proved difficult. Thus, even in the distinctive *castaneus* region from 142.46 to 143.43 Mb of Chr6, although there are two genes that are listed as having phenotypic associations in the Jackson Informatics database, *Abcc*9 and *St8sia*1, the distal half of the

region is devoid of genes. Again, the majority of the distinctive *musculus* region of Chr3 in the Taieri mouse (Figure 4.10), from 10.5 to 13.5 Mb is an utter gene desert. Possibly, smaller regions of chromosomes, somewhat less than a Mb in size, encompassing just one haplotype block and a few genes, common to several wild mice, would be displayed in a more definitive way if more diagnostic SNPs were available, but the 600K array does not have this resolution.

Despite these limitations, it is clear that the hybrid mice chromosomes only contain very small remnants of the *castaneus* genomes, possibly even less than the *domesticus* wild type mice in New Zealand.

4.5.6 Appreciable levels of erroneous SNP calls are inherent to the 600K array

As all seven wild mice were males, one would expect the Chromosome X display to have zero heterozygous calls. In fact, some 10% of the calls are heterozygous (Appendix IV, see last page), which means that an appreciable level of noise is inherent in all of the chromosome displays. While this does not affect the overall picture, it does degrade the finer resolution of specific regions.

Chapter Five: Discussion

The original aim of this thesis was to establish the geographic extent of the *M .m. castaneus/domesticus* hybrids in the southern South Island and then, by the use of high density SNP arrays, characterise the nuclear genomes of these mice in order to identify potential retained genes or genomic regions of *M. m. castaneus* origin that might provide a selective advantage of the hybrids over domesticus.

5.1 Hybrid and Contact Zones

The rapid mtDNA RFLP assay proved very successful in enabling the extent of the hybrid zone to be quickly established. As expected, and in accordance with earlier work (Chubb, 2008; Searle, Jamieson *et al*., 2009) all mouse samples from the deep south were hybrids. In the current survey, hybrid mice were found as far north as 44^{o}S on the coast (Waimate) and 43^{o}S inland (Hakataramea). Conversely, all mice collected in the South Island north of 42^{o}S were *M. m. domesticus*. Between these two areas, hybrid mice coexisted with *M. m. domesticus* in a contact zone with mixed samples from Hook in the south to Otaio, Mawaro, Kingsdown and Temuka in the north. This distribution is depicted in Figure 3.1.

The most striking example of contact between the two mice strains was found at Kingsdown where, in one farmhouse and the adjacent barn, of 33 mouse carcasses assayed, 26 were hybrids and 7 domesticus. Mixtures of the two types of mice were also found in one house in Mawaro and in Temuka. Whether or not these mice cohabited is not known, as live mice were not trapped.

Chubb (2008) described a similar phenomenon on one farm in Ekatahuna in the lower North Island where of 14 mice trapped in one shed, approximate 50% were of each type. Obviously, these mice were cohabitating the shed.

The contact zone in South Canterbury is surprisingly small, approximately 50km north to south and 25km east to west. As stated earlier, there are no obvious geographic, climatic or ecological features that characterise the contact zone. While there are a number of small rivers that run from hills and mountains in the west some 20 to 40km east to the sea, much larger rivers with extensive braided beds lie to the north and south.

As one move south, the climate does change from typical Canterbury conditions with dry winters and hot summers into wetter, colder conditions characteristic of North coastal Otago, but this change is gradual. Whether or not the colder southern conditions have some selective influence is uncertain, but it is relevant to note that in Japan, the *M. m. molossinus* hybrids (*musculus* nuclear/*castaneus* mtDNA) are found in the northern colder regions (Terashima *et al*., 2006). Possibly, the *castaneus* mtDNA could exert some selective advantage similar to that proposed for humans (Balloux, Handley, Jombart, Liu, & Manica, 2009) though whether or not different mtDNA haplotypes exert selective effects in mammals is a matter of some controversy.

The other possibility is that the contact zone, unlike that in middle Europe, is dynamic and slowly moving either to the north or south, and no particular importance should be attached to its current location. Unfortunately, historical mouse specimens, which might answer this possibility were not available.

Finally, it is to be emphasised that, because no geographical area has been found in either this or earlier work (Searle, 2009; Chubb, 2008) that is populated by 'pure' *M. m. castaneus* mice, one cannot carry out the classical and highly informative experiments, such as have been done in Europe, in which the introgression of specific gene alleles in both directions from one subspecies across the contact zone into the zone populated by the other subspecies is analysed. This limitation will be discussed in more detail later.

5.2 Why *M. m. castaneus* mtDNA/*M. m. domesticus* nuclear DNA and not vice versa?

An obvious question that arises is, if originally *M. m. castaneus* and *M. m. domesticus* interbred, why is only one type of hybrid found today? Clearly, some 150 years ago when mice were first introduced in numbers into New Zealand, *M. m. domesticus* males must have bred for an extensive period with *M. m. castaneus* females and, in turn, with the subsequent offspring. In effect a backcross situation that largely eliminated the *M. m. castaneus* genome before, more recently, the hybrids interbred between themselves.

It is probable that the opposite crosses also occurred but these do not appear to have survived. While one can imagine a scenario involving more aggressive behaviour of *M. m. domesticus* males or *M. m. castaneus* females, a more plausible explanation involves the characteristics of crosses involving *M. m. castaneus*.

Some important clues come from the experiences of the Jackson Laboratory in attempting to cross wild-derived mouse strains. They state that fighting is particularly prevalent in progeny of any crosses involving *M. m. castaneus* (http://jaxmice.jax.org/type/wild/wild_care.html).

Moreover, and serendipitously, they have created the 'opposite' hybrid cross to that found in southern New Zealand, namely between *M. m. castaneus* males and *M. m. domesticus* females to create the strain CASA/RjK – see Chapter 3. Interestingly, this cross has proved problematic to maintain because of breeding problems and now have been phased out of the active breeding programme (http://jaxmice.jax.org/jaxnotes/archive/456e.html) (Hammer & Wilson, 1987). Thus, it is quite conceivable, if early hybrids of the same phenotype occurred in early New Zealand, they would have been at a severe selective disadvantage and would not survive. Severe breeding problems have also been noted with crosses of another *M. m. castaneus* strain, CasA (Hammer & Wilson, 1987).

5.3 Historic Origins of the *castaneus/domesticus* strains in New Zealand

Mitochondrial DNA haplotypes found in this study agree with those mentioned in Searle *et al*. (2009), with *domesticus* mice belonging to haplotypes DomNZ.1-5. These mice originate from the Bristish Isles and would have arrived with immigrants and shipping from this area in the 1800s. Hybrid N.Z *castaneus* mice are of the CastNZ.1 haplotype and, again in agreement with Searle *et al*., (2009), presumably arrived with shipping from the Far East.

It is a possibility that *M. m. castaneus* and *M. m. domesticus* mice may have hybridised *en route* to N.Z, rather than after arriving in the country. It is important to note, however, that *M. m. castaneus* never established themselves in mainland Europe, presumably because of the presence of *M. m. doemsticus* mice, and that this may also apply in the case of pure *M. m. castaneus* in N.Z It is also relevant to note that, if there is a reservoir of hybrids of this type that evolved, prior to their introduction into N.Z., in some other locality such as Asia, or

possibly Australia, they do not appear to have colonised other areas to an appreciable extent (McCormick and Wilkins, 2010)

5.4 SNP analyses of nuclear genomes

Although the 600K SNP arrays yielded large amounts of high quality data, significant challenges were encountered using this to analyse the characteristics of the hybrid mice in a meaningful way. The chief problem was that no zone of 'pure' *M. m. castaneus* mice exists in the southern South Island which can be used as a reference. This would not be so problematical if large portions of the hybrid genomes were *castaneus* in nature, but this is not the case.

Two approaches were used in an attempt to overcome this problem. In the first, regions that were highly homozygous in the four hybrid mice and different from the corresponding regions in '*domesticus*' type mice were pinpointed. In the second approach, subsets of SNPs diagnostic of *castaneus*, *domesticus* and *musculus* mice were used. As will be seen below, both approaches have their advantages and disadvantages.

5.4.1 Regions of genome difference identified by simple homozgous scores

As discussed in Chapter 4, many more chromosomal regions than would be expected by pure chance exhibited big differences in homozygosity between the 4 hybrid and the 3 '*domesticus'* wild mice. These regions typically encompassed no more than one or two Mb and most could not be correlated with gene(s) of high phenotypic interest. Others were more interesting, for instance the maternal-fetal conflict region on Chromosome 13 (see section 4.4).

In some cases, when one took the opposite approach and examined gene regions that contained genes of predicted interest, for example, ‘social’ genes, to see if homozygous SNP correlations occurred, some interesting clusters are seen. Thus, for the 200 plus vomeronasal receptor genes, while the SNPs within most clusters do not appear appreciably different, one region on Chr7 from approximately, 24.2 to 24.8Mb, which encompasses the eighteen Vmn1r167 to Vmn1r185 genes, stands out (figure 5.1). Other researchers have discovered differences in Vmn gene clusters that are characteristic of different *Mus musculus* subspecies, but not in this particular cluster (Karn, Young, & Laukaitis, 2010).

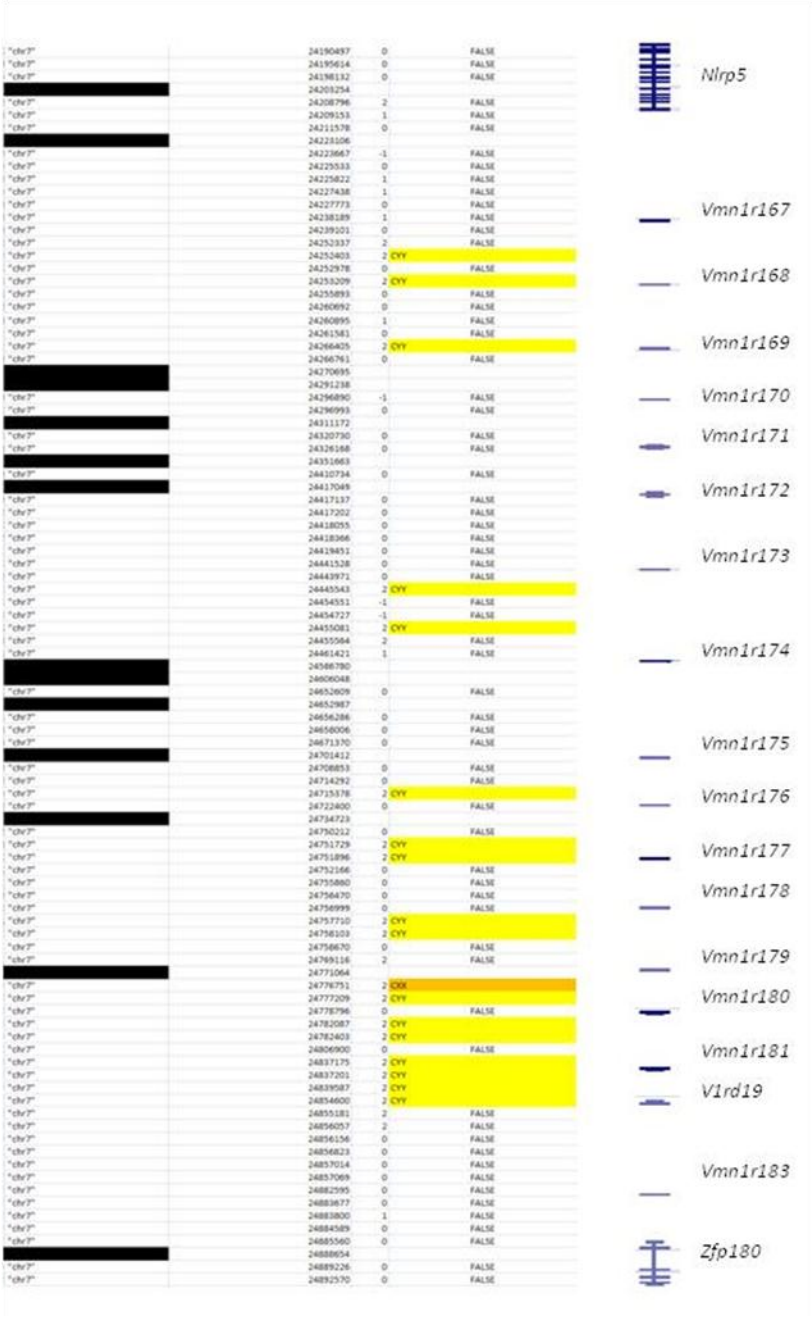


Figure 5.1: A 700 kb region of Chromosome 7 that contains a cluster of 17 vomeronasal receptor genes that coincide with 17 SNPs with alleles preferentially occurring in hybrid mice (yellow and orange). Black bars on left are haplotype block boundaries and approximate gene positions shown on right Although these genomic regions are interesting and suggestive of biologically meaningful differences between the hybrid and *M. m. domesticus* wild mice, and worthy of further study, they do not specifically enable remnants of the castaneus genome to be identified in the hybrid mice.

5.4.2 Analyses using diagnostic SNPs

As discussed earlier, notwithstanding the assumptions made in the original publication (Frazer *et al.,* 2007), the definition of what is a *castaneus* versus a *domesticus* SNP (or haplotype block) is not clearcut because the original genome types were defined by assuming the two wild type laboratory strains, CAST/EiJ and WSB/EiJ were of pure origin. In fact, both contain significant 'contamination' both with each other and with *M. m. musculus* genome fragments (as demonstrated in the Mouse Phylogeny Viewer).

The availability of a subset of cardinal SNPs that are definitive for *M. m. castaneus*, *domesticus* and *musculus* (Mouse Phylogeny Viewer; (Yang *et al*., 2011) is an invaluable advance that largely overcomes the above limitations, even though these subsets are based on analysis of relatively small subsets of wild type mice from very limited localities.

The diagnostic SNP analyses presented in Chapter 4 clearly show that the vast majority (95% plus, or even a higher percentage if Ham1 is treated as aberrant) of the hybrid genomes are *domesticus* in nature. Scanning the SNP data, chromosome by chromosome, does yield some regions suggestive of *castaneus* remnants, and also a few that are *musculus* like. Despite the predictions made in Chapter 1, even these small regions of *castaneus*-like genome are not consistently found in all four hybrids, nor are they consistently absent in the three wild *M. m. domesticus* mice. Indeed, their presence appears somewhat randomly distributed through all seven wild mice. To a lesser extent, *musculus* type regions are also likewise randomly distributed. Very surprisingly, the Ham1 mouse was an obvious exception to all of this, exhibiting large (5-10Mb) complex regions of *musculus* type genome in approximately 50% of the chromosomes. This paradox is discussed in section 5.6.

It is instructive to understand why the hybrid genomes are overwhelmingly *domesticus* in nature, given that originally wild mouse hybrid genomes must have existed which were 50% *castaneus*/50% *domesticus*. As discussed earlier, extensive backcrossing must have occurred with male *M. m. domesticus* mice and early hybrids, preserving the *M. m. castaneus* (maternal) mtDNA while the nuclear genome was 'converted' to *domesticus*. This has many similarities to an inbreeding situation in which one keeps crossing generations of two subspecies and ultimately gets descendents which are homozygous (fixed) at most loci, except in this case we are taking *domesticus* type SNPs as the end point. If we take, as an approximation, the classic inbreeding equations (Green, 1981), after 20 generations of inbreeding, 98.7% of the loci in the genome of each animal should be homozygous. This is the operational definition of the cross being rendered 'inbred'. At each subsequent generation, the level of heterozygosity will fall off by 19.1%, so that at 30 generations, 99.8% of the genome will be homozygous and at 40 generations, 99.98% will be homozygous (figure 5.2, taken from http://www.informatics.jax.org/silver/frames/frame3-3.shtml).

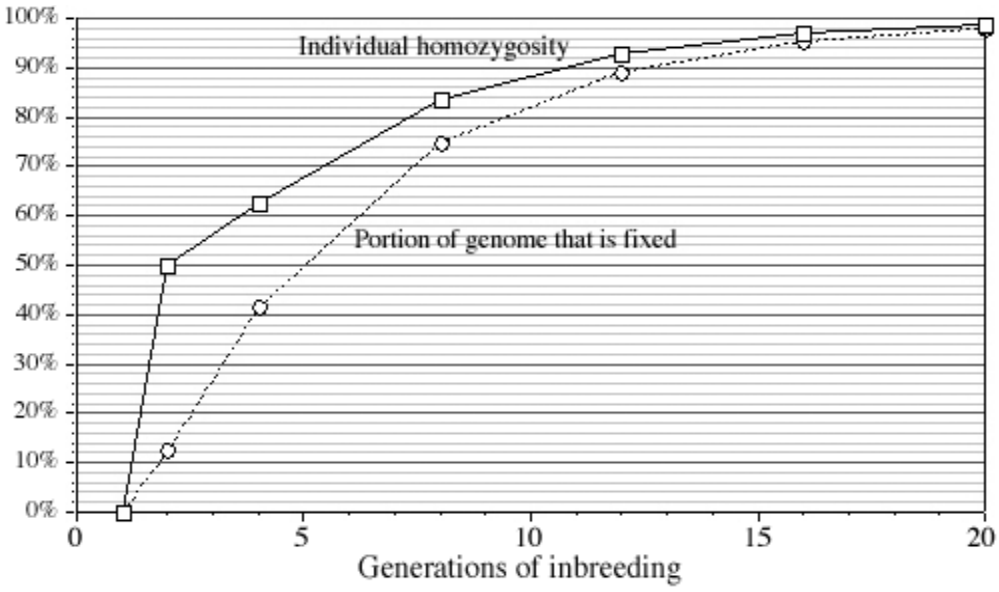


Figure 5.2: Consequences of inbreeding up to 20 generations. Points on the solid line indicate the portion of the genome that will be homozygous in any individual animal at each generation. Points on the dashed line indicate the portion of the genome that will be fixed identically in the two animals chosen to generate the following generation of animals.

Although this is not exactly the situation pertaining in the wild in New Zealand, it does demonstrate that, for all intents and purposes, after 100 to 200 generations (see below) of early hybrids subsequently breeding with an excess of (male) *M. m. domesticus* mice, somewhat less than 0.1% of the genome would retain *castaneus* alleles in the absence of some selective pressure. These retained alleles would reside in haplotype blocks, of which there are some 40,000 to 70,000 in the mouse genome (Perlegen database; (Zhang *et al*., 2005), so one would only expect 40 to 70 at the most to be retained, which equates to just a few per chromosome, and these would only be 40 to a few 100 kb in size. This indeed appears to be the case.

Even if we assume that certain *M. m. castaneus* alleles are retained because they have a strong selective advantage, the size of the 'retained' regions is still likely to be relatively small. Thus, if we were to take the classic inbreeding equations in which a particular allelic region is selected at each generation and backcrossing only occurs with those individuals (i.e. extremely strong selection) one would expect the selected region to be surrounded by a hitchhiked region of 200/n centiMorgans where n is the number of generations of backcrossing. Thus, after 200 generations one might expect ~1 CentiM or ~1Mb of '*castaneus*' type genome to surround a positively selected region.

An important parameter in the above discussion is the number of generations of (hybrid) mice that have been generated in the last 150 years in New Zealand. Estimates of the number of generations that occur in the wild for *M. musculus* vary widely. Nachman and Searle (1995) estimate 4 generations per year, while Salcedo *et al*. (2007) suggest one or two. Given this, it is probably reasonable to assume that 2 generations occur per year for mice living under commensal conditions in the South Island.

Thus one would only expect regions encompassing a few haplotype blocks to be retained after 300 generations (150 years) and these could be quite difficult to detect, given the limited resolution of the diagnostic SNPs and the uncertainty pertaining to the geographic source of those specific for castaneus.

On the other hand, *a priori* considerations (e.g. in the case of the maternal-fetal conflict region and certain vomeronasal receptor genes) do lead to the identification of "candidate" regions encompassing just a few haplotype blocks.

5.5 Chromosome 13 maternal-fetal conflict region

Sequencing of the *Tpbpa* gene clearly showed that there are 2 distinct forms of this protein amongst N.Z mice. Further screening would be useful to test this in the wider N.Z. mouse population, perhaps by use of a PCR-based SNP assay, as the amount of sequencing required (5 primers per sample) may be prohibitive. Somewhat surprisingly, *Tpbpa* in a sample from Lincoln (North of the hybrid zone) typed as that of a *M. m. molossinus* mouse from Japan and a sample from Te Anau (deep within the hybrid zone) typed as that of *M. m. musculus* from Bulgaria. This region of the genome is important for maternal-fetal co-adaptation and is thought to be undergoing positive selection (Chuong *et al*., 2010), which may be encouraging assortative mating.

5.6 The paradox of Ham1

There is no obvious explanation of the presence of large tracts of homozygous '*musculus*-like' chromosome regions in Ham1. It is particularly intriguing that, these regions, where they occur, correspond to regions that are *musculus*-like in the laboratory strain C57BL/6J, even incorporating, in some cases, additional small *castaneus*-like inserts. These exact patterns are not found in other inbred laboratory strains that are not C57 related. Some Ham1 chromosomes did not mimic C57BL/6J to an appreciable extent (Chromosomes 1-5, 15, 19 and X), but the others all contained some C57BL/6J *musculus*-like regions, and a few were virtually identical, for example, Chromosome 8 (figure 5. 3). However, when examined at high-resolution, it is clear that, whereas in the C57BL/6J mouse itself, these regions are purely *musculus,* the corresponding regions in Ham1 contain large groupings of *domesticus* SNPs, indicating that, if indeed, the two mice had the same ancestors, considerable recombination has since occurred in Ham1 (the expanded 15Mb region in Figure 5.3 contains at least 50 major haplotype blocks).

If one assumes the genetic distance encompassed by 15Mb of genome is approximately 10cM, there is a 0.1 probability of a recombination in this interval per meiosis so, after 20 and 40 generations, to a first approximation, there would be at least one recombination event in, respectively 90 and 99% of chromosomes in this interval. Thus, it is conceivable that a small group of semi-inbred mice (such as a deme) with the C57BL/6J genotype, could over a few tens of generations of limited interbreeding with wild type *domesticus* mice, give rise to the SNP patterns such as those depicted in Figure 5.3.

One possible source of C57BL/6J-like N.Z. wild mice is ‘fancy mice’ originating in East Asia and being transported to New Zealand in the 1800s. But it would seem improbable they would maintain such distinct *musculus*-like regions after a hundred plus generations of interbreeding with other, largely *domesticus* wild mice (Chubb, 2008; Searle *et al*., 2009). Possibly in more recent times, local pet-shop mice with ‘fancy mouse’ backgrounds could have escaped and bred with wild mice. However, in both cases, it is difficult to explain why the *musculus*-like regions in Ham1 would correspond only to regions in C57BL/6J and not those from other laboratory strains (these are also of Asian origin but have quite distinct musculus regions). The alternative scenario, namely that C57BL/6J mice escaping from a local animal house some 20 years ago have interbred in this area with wild mice seems more likely. A poorly maintained animal house, at the Ruakura Research Centre housed C57BL/6J mice through the 1980s until 1993 and was only 1.2km over open ground from where Ham1 was trapped. Such a scenario would explain both the extent of recombination observed and also the concordance with chromosomal regions of this specific mouse strain. In this respect, it might also be relevant that the overall heterozygosity in Ham1 is lower than in the other six N.Z. wild mice.

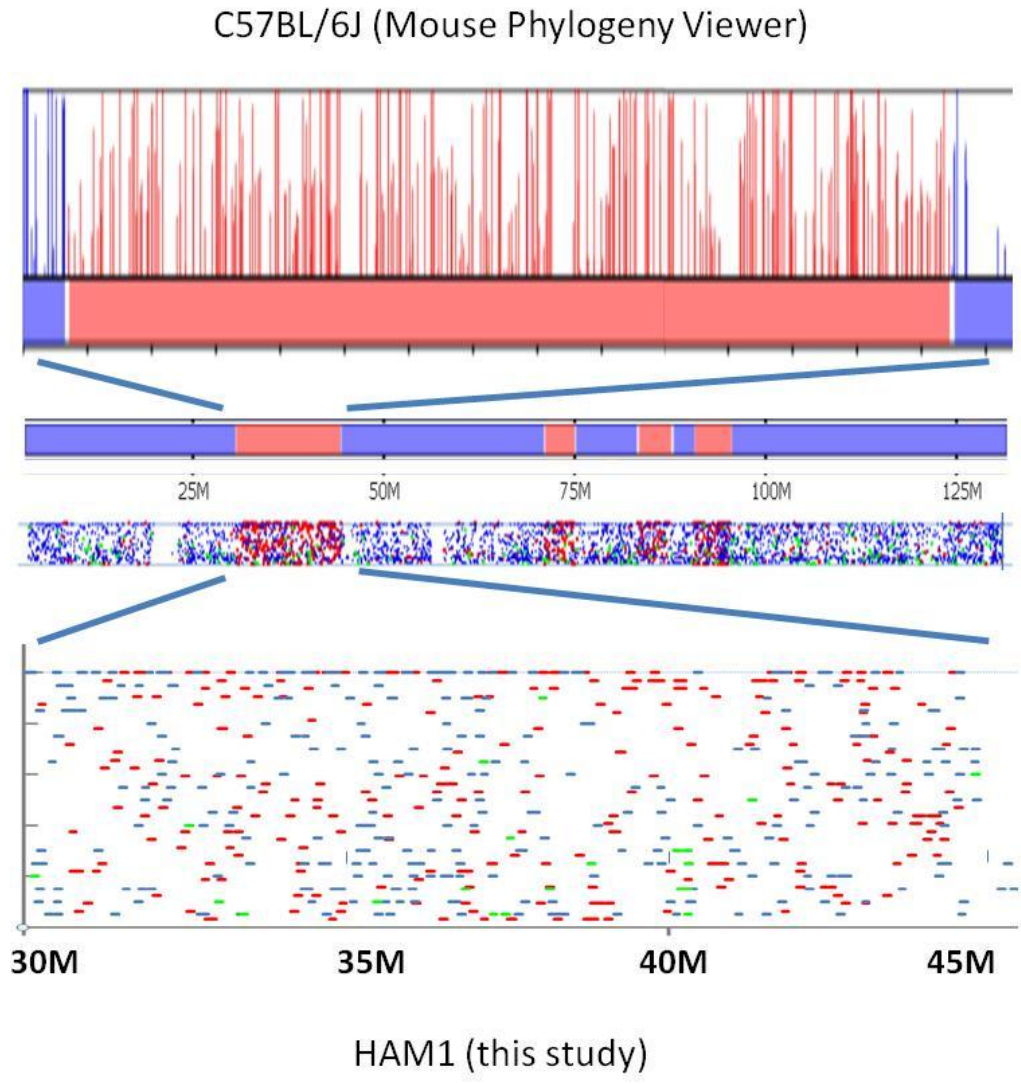


Figure 5.3: Chromosome 8 comparisons of C57BL/6J and Ham1. Diagnostic SNPs reveal four *M.m.musculus* regions in both mice (red) with identical boundaries. The expanded view of one of the regions reveals that, whereas the laboratory mice SNPs are purely "musculus-like", approximately 50% of the SNPs are "*domesticus*-like "in the same Ham1 region. (Ham1 homozygous SNPs are shown – few SNPs are heterozygous).

Obviously, this unexpected finding makes comparisons between Ham1 and the other N.Z. wild mice difficult. Nevertheless, some non-*domesticus* SNP regions unrelated to C57BL/6J also occur in these mice and can be compared and contrasted with the other wild mice (Appendix IV chromosome displays).

Regardless of its origins, the unusual chromosome composition of the Ham1 mouse is an important, if unexpected discovery. No other reports of a similar phenomenon were found in the literature and it is possible that mice in this geographic region represent a unique resource, exhibiting recent interbreeding between two mouse 'strains' with clearly distinguishable genotypes. A survey seeking similar mice in this region could be very informative and could possibly go some way to dispelling the widespread presumption that laboratory strains of mice (and interbred descendants) are quite unsuited to survival in the wild. Interestingly, Fonio *et al*., (2006) have recently questioned this presumption.

Chapter 6: Conclusions

When this project was started, there were two major aims. First, to establish the northern geographic boundary of the hybrid zone (defined as a 'contact zone') in the southern South Island, and second, pinpoint residual c*astaneus* chromosomal regions that were characteristic of the hybrid mice and would distinguish them from their northern *domesticus* wild mice counterparts. While the first aim proved relatively easy to achieve, the second was much more complicated and inconclusive.

Six major conclusions can be made:

6.1: The hybrid zone is probably dictated by climate and geography

As noted earlier, the region where hybrid mice are found – south of 43^{o}S and inland of the Hunter Hills, has an appreciably colder climate, especially in winter, than the southern range of wild type *domesticus* mice, and the actual contact zone, in the Timaru region and inland approximately 30km is climatically in an intermediate region.

6.2 Hybrid mice have a selective advantage in Southern New Zealand

Though no specific genetic characteristic has been identified in either this, or earlier work (Chubb, 2008; Searle *et al.*, 2009) it would seem clear that the absence of pure bred *M. m. castaneus* and *M. m. domesticus* wild mice from this whole region, despite their presumed

presence at early settlement times, strongly argues for some selective advantage being conferred on the hybrid mice .

6.3 Less than 5% of the genomes of the hybrid mice remain *castaneus*-like

The diagnostic SNPs clearly show that very little of the hybrid nuclear genome retains *castaneus*-like characteristics; the exact residual amount is difficult to quantitate and could be considerably less than 5% if one was to consider the genomes one by one, rather than the cumulative castaneus like fragments over all four hybrid genomes. Those areas that remain are almost all very small fragments - much less than 1 Mb and probably encompass just a few haplotype blocks, most are not consistently found in all hybrid wild mice tested, and the three N.Z. *domesticus* wild type mice also contain *castaneus*-like regions. There are also obvious regions of *musculus*-like genome in hybrids, notably the large homozygous region in Chromosome 3 of the Taieri mouse.

6.4 Small *castaneus* (non-*domesticus*) chromosomal regions exist in hybrid mice

Regions encompassing just a few haplotype blocks that are biologically relevant and differ between the hybrid and *domesticus* wild mice were revealed in the SNP analyses. Two in particular were highlighted in this work. One encompassed the maternal-fetal conflict region (Chr13) and another a group of vomeronasal receptor genes (Chr7). The existence and persistence of such small regions is entirely consistent with what one would predict from genetic calculations assuming backcrossing over a hundred plus generations were preserving

regions containing genes undergoing positive selection. There well may be other such regions in the genomes of the hybrid animals, but the number of diagnostic castaneus SNPs is insufficient to give definitive resolution down to the haplotype block level. However, one cannot rule out the possibility that such regions just represent the 'fixation' of *domesticus* alleles from early *M. m. domesticus* mice populating the southern region of New Zealand, especially as founder effects could have profound effects in the small mouse populations that first arrived in early shipping (Pocock *et al*., 2005)

6.5. The Ham1 mouse is abnormal but a valuable resource for future research

The interesting genetic makeup of this mouse has been discussed in detail above and, although it detracts from the main aim of this thesis, namely the comparison of *domesticus* and hybrid wild mice in New Zealand, it may well offer important and unique insights into how two distinct mice populations interbreed in the wild, with the potential to accurately follow genetic recombination events using diagnostic SNPs.

6.6 Mitochondrial DNA could be a major selective factor

The uncertainties surrounding the introduction and spread of mice in southern New Zealand have been discussed in the earlier Chapters of this thesis. There are so many unknowns that any theories aimed at explaining the exclusive colonisation of the southern South Island by hybrid mice can only be speculative at best. It is known however, that mouse numbers in this area expanded dramatically from a very low population base around coastal areas in the 1860s and moved inland into areas where they were previously unreported, in very large

numbers, sometimes in plague like numbers over the next 10 to 20 years. This dispersal involved several phenomena such as the "vacuum effect", immigration, founder effects and, no doubt, colonisation, extinction and recolonisation in some areas (Pocock, Hauffe and Searle, 2005). While founder effects alone might give rise to genetically distinct (hybrid) mice that existed in reproductive isolation, it would seem much more likely that some kind of selective advantage (possibly with reproductive isolation being secondary) would operate to create such a large hybrid zone. In this regard, it is important to note that, unlike most other areas of the world, mice in New Zealand cannot be categorised as being purely commensal or feral (Pocock, Hauffe and Searle, 2005), which could possibly render climatic factors, such as the more severe winters in southern New Zealand a more important selective factor than it would be elsewhere.

Given all of this, it is worth recalling that the most obvious and consistent difference between the hybrid and *M. m. domesticus* wild mice is the mitochondrial DNA. Moreover, most of the hybrid mice in southern N.Z. fall in just two closely related *castaneus* haplogroups, namely CasNZ.1 and 3 (Searle *et al*., 2009). In the current and earlier work (Searle *et al*., 2009; Chubb, 2008) only very general mtDNA classifications have been made, not proceeding beyond D-loop sequencing. In retrospect, this could be a serious oversight as evidence is accumulating that it is wrong to assume that all mitochondrial haplogroups are selectively equivalent. Thus, in humans, preliminary studies indicate that some alleles of some mitochondrial genes may be advantageous in colder climates (Balloux *et al*., 2009). One such candidate gene identified in humans is NADH-dehydrogenase-subunit-3 (ND3) which is involved in energy metabolism. It is highly relevant to note that this gene is very polymorphic in mice, with many of the substitutions being non-synonymous (Nachman *et al*., 1994a; Nachman *et al*., 1994b). To date, the phenotypic effects of these (and other mitochondrial gene) variants on mice housed at low temperatures are unknown. It is also

possible that other mitochondrial gene variants are confering a selective advantage to hybrid mice in the southern South Island, such as those related to immunity and mate choice. Yu *et a.l* (2009) reported an mtDNA polymorphism that increases the supseceptibility of mice to several autoimmune diseases as well as impairing the reproductive success of females. Thus there is a possibility that selective factors attributed to the mtDNA genome other than those related to energy metabolism are at play in the N.Z hybrid mouse population, but this requires a great deal of further study.

6.7 Future Directions

The current work clearly defines the future directions that research into N.Z. hybrid mice should take.

First, now that the southern contact zone is clearly defined many more samples should be collected across this zone extending some 100km in each direction, employing classic transect sampling methods.

Second, it is clear that high density SNP array analysis is the method of choice in future nuclear genome work, but with a much larger number required to get a robust picture of the nature of genetic variation in wild mice throughout New Zealand and to identify consistent differences at the haplotype block level. Such analyses present several challenges. The most obvious amongst these is the cost – currently ~$NZ1500 per sample. Another is the limitations of the chips (not explicitly mentioned by Yang *et a.l*, 2011). These include 5 to 10% false calls and, currently, an inability to analyse Y chromosomes. Hopefully, these problems can be resolved. Another limitation is the bias towards *M. m. domesticus*, and a

paucity of *M. m. castaneus*, diagnostic SNPs. This problem would be overcome to some extent by access to the software that interrogates SNP chip data for so-called VINO genetic variants. In theory, the current data obtained for this thesis could be reanalysed to give VINO data.

Third, it would be advantageous to obtain SNP data from wild *M. m. castaneus* in the regions of the Far East that the New Zealand *M. m. castaneus* are presumed to have been transported from in the nineteenth century, possibly the trading port of Canton. The diagnostic alleles in the Mouse Phylogeny Viewer are all based on a group of *M. m. castaneus* mice from northwest India.

Fourth, the recovery of zooarchaeological mouse samples, especially from the southern South Island could possibly provide definitive evidence concerning the early presence of pure *M. m. castaneus* in New Zealand. Reasonably preserved remains would be easily assayed for nuclear markers using modern molecular techniques.

Fifth, and perhaps most importantly, complete sequencing of mitochondrial genomes from a representative number (10 of each?) hybrid and *M. m. domesticus* wild mice would allow possible mitochondrial gene variations to be established.

Extending the work described in the current thesis in the five directions indicated above would hopefully give a definitive answer to why southern New Zealand has such a large hybrid mouse population, one that extends over a large geographical area.

As a final comment, there is a very real possibility that this New Zealand hybrid mouse will prove to be a unique resource and we are observing the early steps on the way to speciation in much the same way as has occurred over the last several hundred years in northern Japan with *M. m. molossinus.*

References

Baines, J. F., & Harr, B. (2007). Reduced X-Linked Diversity in Derived Populations of House Mice. *Genetics, 175*(4), 1911-1921. 10.1534/genetics.106.069419

Balloux, F. o., Handley, L.-J. L., Jombart, T., Liu, H., & Manica, A. (2009). Climate shaped the worldwide distribution of human mitochondrial DNA sequence variation. *Proceedings of the Royal Society B: Biological Sciences, 276*(1672), 3447-3455. 10.1098/rspb.2009.0752

Boissinot, S., & Boursot, P. (1997). Discordant Phylogeographic Patterns Between the Y Chromosome and Mitochondrial DNA in the House Mouse: Selection on the Y Chromosome? *Genetics, 146*(3), 1019-1034.

Boursot, P., Auffray, J. C., Britton-Davidian, J., & Bonhomme, F. (1993). The Evolution of House Mice. *Annual Review of Ecology and Systematics, 24*, 119-152.

BoŽÍKovÁ, E. V. A., Munclinger, P., Teeter, K. C., Tucker, P. K., MacholÁN, M., & PiÁLek, J. (2005). Mitochondrial DNA in the hybrid zone between Mus musculus musculus and Mus musculus domesticus: a comparison of two transects. *Biological Journal of the Linnean Society, 84*(3), 363-378. 10.1111/j.1095-8312.2005.00440.x

Cann, R. L., Stoneking, M., & Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature, 325*(6099), 31-36.

Chubb, T. L. A. (2008). Phylogeography and Hybridisation of the New Zealand House Mouse: The University of Waikato.

Chuong, E. B., Tong, W., & Hoekstra, H. E. (2010). Maternal-Fetal Conflict: Rapidly Evolving Proteins in the Rodent Placenta. *Molecular Biology and Evolution, 27*(6), 1221-1225. 10.1093/molbev/msq034

Del Punta, K., Rothman, A., Rodriguez, I., & Mombaerts, P. (2000). Sequence Diversity and Genomic Organization of Vomeronasal Receptor Genes in the Mouse. *Genome Research, 10*(12), 1958-1967. 10.1101/gr.140600

Derothe, J.-M., Loubes, C., Perriat-Sanguinet, M., Orth, A., & Moulia, C. (1999). Experimental trypanosomiasis of natural hybrids between house mouse subspecies. *International Journal for Parasitology, 29*(7), 1011-1016.

Din, W., Anand, R., Boursot, P., Darviche, D., Dod, B., Jouvin-Marche, E., et al. (1996). Origin and radiation of the house mouse: clues from nuclear genes. *Journal of Evolutionary Biology, 9*(5), 519-539. 10.1046/j.1420-9101.1996.9050519.x

Fonio, E., Benjamini, Y., Sakov, A., & Golani, I. (2006). Wild mouse open field behavior is embedded within the multidimensional data space spanned by laboratory inbred strains. *Genes, Brain and Behavior, 5*(5), 380-388. 10.1111/j.1601-183X.2005.00170.x

Frazer, K. A., Eskin, E., Kang, H. M., Bogue, M. A., Hinds, D. A., Beilharz, E. J., et al. (2007). A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature, 448*(7157), 1050-1053.

Geraldes, A., Basset, P., Gibson, B., Smith, K. L., Harr, B., Yu, H.-T., et al. (2008). Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. *Molecular Ecology, 17*(24), 5349-5363. 10.1111/j.1365-294X.2008.04005.x

Green, E. L. (1981). *Genetics and Probability in Animal Breeding Experiments.*: Oxford University Press, New York.

Guénet, J.-L., & Bonhomme, F. (2003). Wild mice: an ever-increasing contribution to a popular mammalian model. *Trends in genetics : TIG, 19*(1), 24-31.

Gunduz, İ., Tez, C., Malikov, V., Vaziri, A., Polyakov, A. V., & Searle, J. B. (2000). Mitochondrial DNA and chromosomal studies of wild mice (Mus) from Turkey and Iran. *Heredity, 84*(4), 458-467. 10.1046/j.1365-2540.2000.00694.x

Hammer, M. F., & Wilson, A. C. (1987). Regulatory and Structural Genes for Lysozymes of Mice. *Genetics, 115*(3), 521-533.

Jones, E. P., Jensen, J.-K., Magnussen, E., Gregersen, N., Hansen, H. S., & Searle, J. B. (2011). A molecular characterization of the charismatic Faroe house mouse. *Biological Journal of the Linnean Society, 102*(3), 471-482. 10.1111/j.1095-8312.2010.01597.x

Jones, E. P., Van Der Kooij, J., Solheim, R., & Searle, J. B. (2010). Norwegian house mice (Mus musculus musculus/domesticus): distributions, routes of colonization and patterns of hybridization. *Molecular Ecology, 19*(23), 5252-5264. 10.1111/j.1365-294X.2010.04874.x

Karn, R., Young, J., & Laukaitis, C. (2010). A candidate subspecies discrimination system involving a vomeronasal receptor gene with different alleles fixed in M. m. domesticus and M. m. musculus. *Genome Biology, 11*(Suppl 1), P22.

King, C. M. (Ed.). (2005). *The Handbook of New Zealand Mammals.* (2nd ed.): Oxford University Press. USA.

King, M. (2003). *The Penguin history of New Zealand.* New Zealand:: Penguin Books.

Kozak, C. A., & O'Neill, R. R. (1987). Diverse wild mouse origins of xenotropic, mink cell focus-forming, and two types of ecotropic proviral genes. *J. Virol., 61*(10), 3082-3088.

Laukaitis, C., Heger, A., Blakley, T., Munclinger, P., Ponting, C., & Karn, R. (2008). Rapid bursts of androgen-binding protein (Abp) gene duplication occurred independently in diverse mammals. *BMC Evolutionary Biology, 8*(1), 46.

Macholan, M., Baird, S., Munclinger, P., Dufkova, P., Bimova, B., & Pialek, J. (2008). Genetic conflict outweighs heterogametic incompatibility in the mouse hybrid zone? *BMC Evolutionary Biology, 8*(1), 271.

Macholán, M., Baird, S. J. E., Dufková, P., Munclinger, P., Bímová, B. V., & Piálek, J. ASSESSING MULTILOCUS INTROGRESSION PATTERNS: A CASE STUDY ON THE MOUSE X CHROMOSOME IN CENTRAL EUROPE. *Evolution*, no-no. 10.1111/j.1558-5646.2011.01228.x

Macholán, M., Baird, S. J. E., Dufková, P., Munclinger, P., Bímová, B. V., & Piálek, J. (2011). ASSESSING MULTILOCUS INTROGRESSION PATTERNS: A CASE STUDY ON THE MOUSE X CHROMOSOME IN CENTRAL EUROPE. *Evolution, 65*(5), 1428-1446.

McCormick, H. M., & Wilkins, R. J. (2010). Rapid, large-scale and inexpensive genotype differentiation of Mus musculus castaneus and domesticus sub-species. *Molecular Ecology Resources, 10*(1), 218-221. 10.1111/j.1755-0998.2009.02740.x

Moulia, C, Aussel, J, P., Bonhomme, F, et al. (1991). *Wormy mice in a hybrid zone : a genetic control of susceptibility to parasite infection* (Vol. 4). Basel, SUISSE: Birkhäuser.

Mudge, J., Armstrong, S., McLaren, K., Beynon, R., Hurst, J., Nicholson, C., et al. (2008). Dynamic instability of the major urinary protein gene family revealed by genomic and phenotypic comparisons between C57 and 129 strain mice. *Genome Biology, 9*(5), R91.

Munclinger, Pavel, Bozikova, Eva, Sugerkova, Monika, et al. (2002). Genetic variation in house mice (Mus, Muridae, Rodentia) from the Czech and Slovak Republics. *Folia Zoologica, 51*, 12.

Munclinger, P., Boursot, P., & Dod, B. (2003). B1 insertions as easy markers for mouse population studies. *Mammalian Genome, 14*(6), 359-366. 10.1007/s00335-002-3065-7

Nachman, M. W., Boyer, S. N., & Aquadro, C. F. (1994a). Nonneutral evolution at the mitochondrial NADH dehydrogenase subunit 3 gene in mice. *Proceedings of the National Academy of Sciences, 91*(14), 6364-6368.

Nachman, M. W., Boyer, S. N., Searle, J. B., & Aquadro, C. F. (1994b). Mitochondrial DNA Variation and the Evolution of Robertsonian Chromosomal Races of House Mice, Mus domesticus. *Genetics, 136*(3), 1105-1120.

Nachman, M. W., & Searle, J. B. (1995). Why is the house mouse karyotype so variable? *Trends in Ecology & Evolution, 10*(10), 397-402.

Nagamine, C. M., Nishioka, Y., Moriwaki, K., Boursot, P., Bonhomme, F., & Lau, Y. F. C. (1992). The musculus-type Y Chromosome of the laboratory mouse is of Asian origin. *Mammalian Genome, 3*(2), 84-91. 10.1007/bf00431251

Orth, A., Adama, T., Din, W., & Bonhomme, F. (1998). Hybridation naturelle entre deux sous-especes de souris domestiques, Mus musculus domesticius et Mus musculus castaneus, pres de lac Casitas (Californie). *Genome, 41*(1), 104.

Panithanarak, T., Hauffe, H. C., Dallas, J. F., Glover, A., Ward, R. G., & Searle, J. B. (2004). LINKAGE-DEPENDENT GENE FLOW IN A HOUSE MOUSE CHROMOSOMAL HYBRID ZONE. *Evolution, 58*(1), 184-192. 10.1111/j.0014-3820.2004.tb01585.x

Pertile, M. D., Graham, A. N., Choo, K. H. A., & Kalitsis, P. (2009). Rapid evolution of mouse Y centromere repeat DNA belies recent sequence stability. *Genome Research*. 10.1101/gr.092080.109

Pocock, M. J. O., Hauffe, H. C., & Searle, J. B. (2005). Dispersal in house mice. *Biological Journal of the Linnean Society, 84*(3), 565-583. 10.1111/j.1095-8312.2005.00455.x

Prager, E. M., Sage, R. D., Gyllensten, U., Thomas, W. K., Hübner, R., Jones, C. S., et al. (1993). Mitochondrial DNA sequence diversity and the colonization of Scandinavia by house mice from East Holstein. *Biological Journal of the Linnean Society, 50*(2), 85-122.

Raufaste, N., Orth, A., Belkhir, K., Senet, D., Smadja, C., Baird, S. J. E., et al. (2005). Inferences of selection and migration in the Danish house mouse hybrid zone. *Biological Journal of the Linnean Society, 84*(3), 593-616. 10.1111/j.1095-8312.2005.00457.x

Ross, J. O. C. (1987). William Stewart: sealing captain, trader and speculator. *Canberra, Australia: Roebuck Society.*

Rozen, S., & Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol, 132*, 365-386.

Ruscoe, W. A., & Murphy, E. C. (2005). *House Mouse.* . Auckland, New Zealand: Oxford University Press.

Sage, R. D., Heyneman, D., Lim, K.-C., & Wilson, A. C. (1986). Wormy mice in a hybrid zone. *Nature, 324*(6092), 60-63.

Salcedo, T., Geraldes, A., & Nachman, M. W. (2007). Nucleotide Variation in Wild and Inbred Mice. *Genetics, 177*(4), 2277-2291. 10.1534/genetics.107.079988

Searle, J. B., Jamieson, P. M., Gunduz, I., Stevens, M. I., Jones, E. P., Gemmill, C. E. C., et al. (2009). The diverse origins of New Zealand house mice. *Proceedings of the Royal Society B: Biological Sciences, 276*(1655), 209-217. 10.1098/rspb.2008.0959

Searle, J. B., Jones, C. S., Gunduz, I., Scascitelli, M., Jones, E. P., Herman, J. S., et al. (2009). Of mice and (Viking?) men: phylogeography of British and Irish house mice. *Proceedings of the Royal Society B: Biological Sciences, 276*(1655), 201-207. 10.1098/rspb.2008.0958

She, J. X., Bonhomme, F., Boursot, P., Thaler, L., & Catzeflis, F. (1990). Molecular phylogenies in the genus Mus: Comparative analysis of electrophoretic, scnDNA hybridization, and mtDNA RFLP data. *Biological Journal of the Linnean Society, 41*(1-3), 83-103. 10.1111/j.1095-8312.1990.tb00823.x

Smith, I. W. G. (2002). The New Zealand sealing industry: history, archaeology, and heritage management. *Wellington, New Zealand: Department of Conservation.*

Team, R. C. D. (2004). R: A language and environment for statistical computing: Vienna: R Foundation for Statistical Computing. Available: http://www.R-project.org. Accessed 28 September 2005. .

Teeter, K. C., Payseur, B. A., Harris, L. W., Bakewell, M. A., Thibodeau, L. M., Oâ€™Brien, J. E., et al. (2008). Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome Research, 18*(1), 67-76. 10.1101/gr.6757907

Terashima, M., Furusawa, S., Hanzawa, N., Tsuchiya, K., Suyanto, A., Moriwaki, K., et al. (2006). Phylogeographic origin of Hokkaido house mice (Mus musculus) as indicated by genetic markers with maternal, paternal and biparental inheritance. *Heredity, 96*(2), 128-138.

White, T. (1890). On Rats and Mice. *Transactions and Proceedings of the Royal Society of New Zealand 1868-1961, 23*, 195-201.

Yang, H., Wang, J. R., Didion, J. P., Buus, R. J., Bell, T. A., Welsh, C. E., et al. (2011). Subspecific origin and haplotype diversity in the laboratory mouse. *Nat Genet, advance online publication*.

Yu, X., Wester-Rosenlof, L., Gimsa, U., Holzhueter, S.-A., Marques, A., Jonas, L., et al. (2009). The mtDNA nt7778 G/T polymorphism affects autoimmune diseases and reproductive performance in the mouse. *Human Molecular Genetics, 18*(24), 4689-4698. 10.1093/hmg/ddp432

Zhang, J., Hunter, K. W., Gandolph, M., Rowe, W. L., Finney, R. P., Kelley, J. M., et al. (2005). A high-resolution multistrain haplotype analysis of laboratory mouse genome reveals three distinctive genetic variation patterns. *Genome Research, 15*(2), 241-249. 10.1101/gr.2901705

Appendices

Appendix I: Samples and locations

Sample Names	Collection location	Collection* Date	Sample types
B1-2 [c]	Hook	April 2009	Tails
B3 [c]	Dunedin	Unknown	Tails
B4-6 [c]	Dunedin	June-August 2009	Tails
G1-4 [c]	Gore	January 2010	Tails
Ham1 [d]	Hamilton	September 2009	Livers
J1-3 [c]	Waianakarua	June 2009	Livers
J4-10 [c]	Waianakarua	July 2009	Tails
J11-16 [c]	Waianakarua	July – Oct 2009	Tails
K1-20 [1]	Kingsdown	Pre- April2009	Tail from Remains
K21-34 [1]	Kingsdown	April – Oct 2009	Tail from Remains
Li1-10 [d]	Lincoln	(Sept) 2009	Livers
TA1-5 [c]	Te Anau	September 2009	Liver
Tr1-2 [c]	Taieri	April 2009	Liver
W1-9 [2]	Mawaro	April-June 2009	Liver
Wm1-2 [c]	Waimate	April 2009	Tails
Wm3-17 [c]	Waimate	May-August 2009	Tails
Wm18-21 [c]	Waimate	Sept-Oct 2009	Tails
Wm22-34 [c]	Waianakarua	March 2009	Tails
X1-7 [3]	Temuka	April 2009	Tails
X8-12 [3]	Temuka	Sept-Nov 2009	Tail

[c] All Cast mtDNA, [d] All Dom, [1] (27 Cast, 7 Dom), [2] (6 Cast, 3 Dom), [3] (3 Cast, 9 Dom)

*Skeletal and carcass remains were of indeterminate age – some animals obviously only died days to weeks before collection, others could be years old. Tails were often collected over several months by volunteers. Additionally, mouse droppings were collected from South Canterbury between Temuka and Mawaro in the west to Waimate in the south, and also from Hakataramea Downs. Some of these droppings were recent (2008-2009) while others were probably much older. The mtDNA in the Hakataramea samples (8) were all typed as Cast. Of the approx. 30 South Canterbury samples tested, half were typed as Cast and half as Dom, often with both types at one location.

Appendix II: Mathematica programme

(supplied by R J Wilkins)

```
ClearAll["Global`*"]
```

```
(*****************************
  ClearAll at the top clears any interfering junk before you start.
    This programme imports Chromosome data from two files in my JAX folder -  the two SNP datafiles for each chromosome e.g.Chr19 plus Chr19B, consist of, respectively, four wild type which also includes the three reference mouse strains, and in the second file the three wild type that were done originally.  In order, from South to North these are Tea (TeAnau) Tai (Taieri) Wai  (Waianakarua) Maw (Mawaro) Blu (Bluegum Park near Temuka) Lin (Lincoln) and Ham (Hamiliton) . You need to set the Imported Files for each chromosome by changing the TWO numbers in the first part of the programme, i.e. 1 to 19 and X for each chromosome.

   Then make sure your cursor is somwhere within this area and the vertical bar on the extreme right -  start the programme by pushing BOTH Enter and Shift! - the vertical bar will go dark until the files are loaded and the new graph appears at the bottom - should take no more than 10 seconds.

   The graphed output has the chromosome location along the bottom and the SNP genotypes as the ordinate.  The B allele ("2"
     in the original JAX notation is the upper set grouped at ~3.2/4.2.  There are 9 data points with CAST at the top and DOM at the bottom encompassing the seven wild type in the middle with Tea at the top and Ham at bottom (South to North). There is a small gap between the 4 hybrid and 3 dom wild types.  The heterozygotes ("1" originally) are grouped below these, then below these again the "0" homozygotes which constitute the bulk of the SNPs.  The very small group at the bottom are the Ns which did not give reliable calls. The display moves along the chromosome by slider the slider - more control is got by pushing the "+" and then using the controls that open.
      It is best to put the screen on 300% for looking at the graph.  To change the window size you need to go into the PlotRange parameters - (x + 1000000) gives a window size of a million bases which can obviously be made a lot smaller or bigger. If you do this, you need to change the increments on the controller.  Look at the last bracket, the second to last number is the upper limit - 200 million is convenient and does not need changing but the last number is how far the slider moves when the incremental control is used - 900000 for the million base window - this increment should be a bit less than the window size you set to ensure you get overlap.
      To copy the screen, go on the 300% and just PrintScreen on your computer.

   **********************************************)
```

```
TestAll=Import["C:\Documents and Settings\wilkinsd\My Documents\\JAX Feb 10\Chr13.csv"];
TestBAll=Import["C:\Documents and Settings\wilkinsd\My Documents\\JAX Feb 10\Chr13B.csv"];
fancyb[{a_,b_,c_,d_,e_,f_,g_,h_,i_,j_,k_,l_,m_}]:= {m,(b+1.5)};
fancyc[{a_,b_,c_,d_,e_,f_,g_,h_,i_,j_,k_,l_,m_}]:= {m,(c+1.5)};
fancyd[{a_,b_,c_,d_,e_,f_,g_,h_,i_,j_,k_,l_,m_}]:= {m,(d+2.15)};fancye[{a_,b_,c_,d_,e_,f_,g_,h_,i_,j_,k_,l_,m_}]:= {m,(e+1.5)};
fancyf[{a_,b_,c_,d_,e_,f_,g_,h_,i_,j_,k_,l_,m_}]:= {m,(f+1.5)};
fancyg[{a_,b_,c_,d_,e_,f_,g_,h_,i_,j_,k_,l_,m_}]:= {m,(g+1.5)};
fancyh[{a_,b_,c_,d_,e_,f_,g_,h_,i_,j_,k_,l_,m_}]:= {m,(h+1.5)};
fancybb[{aa_,bb_,cc_,dd_,ee_,ff_,gg_}]:= {gg,(bb+1.5)};
fancycc[{aa_,bb_,cc_,dd_,ee_,ff_,gg_}]:= {gg,(cc+1.5)};
fancydd[{aa_,bb_,cc_,dd_,ee_,ff_,gg_}]:= {gg,(dd+1.5)};
fancyCast[_]=fancyb/@TestAll;
fancyDom[_]=fancyc/@TestAll;
fancyMus[_]=fancyd/@TestAll;
```

```
fancyTai[_]=fancye/@TestAll;
fancyLin[_]=fancyf/@TestAll;
fancyHam[_]=fancyg/@TestAll;
fancyTea[_]=fancyh/@TestAll;
fancyWai[_]=fancybb/@TestBAll;
fancyMaw[_]=fancycc/@TestBAll;
fancyBlu[_]=fancydd/@TestBAll;

Manipulate[ListPlot[{fancyMus[_],fancyCast[_],fancyTea[_],fanc
yTai[_],fancyWai[_],fancyMaw[_],fancyBlu[_],fancyLin[_],fancyH
am[_],fancyDom[_]},PlotStyle→PointSize[.006], PlotRange→ {
{x,(x + 1000000)},{0,4.4}}],{x,0,200000000,200000}]
```

Appendix IV: Diagnostic SNPs Excel displays

(Macros supplied by RJ Wilkins)

Appendix IV: Available as a supplementary file

Appendix V: Samples used for Figure 3.3

All samples listed below were obtained from Tanya Chubb. See Chubb (2008) for details.

Bottom row, from left to right: SH2 29, SH2 37, SH2 27, SH2 35, SH2 36, DN1, DN2, DN3, DN4, DN5, DN6, SH2 26, SH2 28, SH2 39, SH2 41, SH2 25, SH2 38, SH2 46, SH2 42, SH2 34, SH2 30, SH2 32, SH2 31, SH2 33, SH2 48, SH2 49, SH2 50, SH2 55, SH2 71, SH2 84, SH2 73, SH2 59, SH2 40, SH2 58, SH2 61, SH2 68, SH2 60, SH2 80, SH2 83, SH2 67, SH2 64, SH2 56, SH2 54, SH2 65, SH2 66, SH2 44, SH2 51, SH2 70

Top row, from left to right: SH2 82, SH2 52, SH2 53, SH2 74, SH2 47, SH2 63, SH2 79, SH2 77, SH2 78, SH2 75, SH2 62, SH2 43, SH2 69, SH2 57, SH2 72, SH2 81, SH2 76, H4, H13, H6, H9, H5, H7, H8, H12, H18, H10, H11, H15, H14, H22, H19, H20, H32, H25, H26, H23, H24, H27, H29, H30, H31, H28, H47, H21, H40, H37, H39

Appendix VII: R script

```
data <- read.table(file="/Users/Helen/Documents/Jax SNP files/JAX Feb 10/CGD_genotypes_dwilkins_All.txt",
sep="\t", fill=TRUE,skip=1)
#V2 is  CAST.EiJ
#V3 is WSB.EiJ
#V4 is PWD.PhJ
#V5 is TR1
#V6 is L_2
#V7 is Ham_1
#V8 is TA_4
#V9 is Allele.A
#V10 is Allele.B

CAST.EiJ <- array('N',length(data[,1]))
WSB.EiJ <- array('N',length(data[,1]))
PWD.PhJ <- array('N',length(data[,1]))
TR1 <- array('N',length(data[,1]))
L_2 <- array('N',length(data[,1]))
Ham_1 <- array('N',length(data[,1]))
TA_4 <- array('N',length(data[,1]))


for(i in 1:length(data[,1])){
        if(data$V2[i]==0) CAST.EiJ[i]=as.character(data$V9[i])
        if(data$V2[i]==2) CAST.EiJ[i]=as.character(data$V10[i])
#       if((data$V2[i]==0)&&(data$V12=="chr5")) CAST.EiJ[i]=as.character(data$V9[i])


        if(data$V3[i]==0) WSB.EiJ[i]=as.character(data$V9[i])
        if(data$V3[i]==2) WSB.EiJ[i]=as.character(data$V10[i])

        if(data$V4[i]==0) PWD.PhJ[i]=as.character(data$V9[i])
        if(data$V4[i]==2) PWD.PhJ[i]=as.character(data$V10[i])

        if(data$V5[i]==0) TR1[i]=as.character(data$V9[i])
        if(data$V5[i]==2) TR1[i]=as.character(data$V10[i])

        if(data$V6[i]==0) L_2[i]=as.character(data$V9[i])'
        if(data$V6[i]==2) L_2[i]=as.character(data$V10[i])

        if(data$V7[i]==0) Ham_1[i]=as.character(data$V9[i])
        if(data$V7[i]==2) Ham_1[i]=as.character(data$V10[i])

        if(data$V8[i]==0) TA_4[i]=as.character(data$V9[i])
        if(data$V8[i]==2) TA_4[i]=as.character(data$V10[i])


}

plot(0, xlim=c(69000000,7000000), ylim=c(-2,3), pch=16)
for(i in 1:length(data[,1])){
        if(data$V12[i]=="chr5") points(data$V13[i],data$V3[i], pch=16, cex=0.2, col="blue")
        if(data$V12[i]=="chr5") points(data$V13[i],data$V5[i], pch=16, cex=0.2, col="red")
}
```

```
outputMatrix<-cbind(CAST.EiJ, WSB.EiJ, PWD.PhJ, TR1, L_2, Ham_1, TA_4)
write.table(outputMatrix, file="/Users/Helen/Documents/Jax SNP files/JAX Feb
10/CGD_genotypes_dwilkins_All_filtered.txt", sep="\t")

plottingPoints <- array(0,length(data[,1]))
for(i in 1:length(data[,1])){
if((L_2[i]==Ham_1[i])&&(TR1[i]==TA_4[i])&&(L_2[i]!=TA_4[i])&&(CAST.EiJ[i]==TA_4[i])&&(Ham_1[i]
==WSB.EiJ[i])) plottingPoints[i]=1
}
Haplengths <- array(-1,length(data[,1]))
temp=0
for(i in 1:length(plottingPoints)){
if(plottingPoints[i]==1){
temp=temp+1
}
else{
Haplengths[i]=temp
temp=0
}
```