Mus musculus helgolandicus: insights into their origin
A study based on genetic and morphometrics analysis

Dissertation

in fulfilment of the requirements for the degree

Doctor rerum naturalium

of the faculty of Mathematics and Natural Sciences

at Kiel University

submitted by

Hiba Mohammed Ali Babiker

International Max-Planck Research School (IMPRS)

Evolutionary Genetics

Max-Planck Institute for Evolutionary Biology

Plön, July, 2014

First referee: Prof. Dr. Diethard Tautz
Second referee: Prof. Dr. Hinrich Schulenburg

Date of the oral examination: 29.09.2014
Date of approval: 29.09.2014
Prof. Dr. Wolfgang Duschl, Dean

To the memory of my father Mohammed Ali Ahmed (1948-2008).
To the memory of my brother Waleed Mohammed Ali (1982-2005).
To my beloved mother Kawther Abubaker.

who valued education above all.

Contents

Acknowledgements

Over the past three years I have been pursuing this thesis, which has been a life time project; I would not have been able to complete it without the help and support of many people. The following deserve special appreciation and gratitude: my supervisor, Prof. Dr. Diethard Tautz for his dedicated supervision, critical questioning, continuing support and enthusiasm for my work. My thesis committee, Prof. Dr. Bernhard Haubold for his invaluable suggestions; Prof. Dr. John Baines for his support. I would like to express my deep gratitude to Dr. Anja Schunke for her precious help introducing me to the geometric morphometrics approach and for her time discussing analysis and results.

I am particularly grateful to Heike Harre for her assistance during our trip to Heligoland for mouse collection. I am indebted to you for your help in planning for the trip, equipment preparation and patience during the trip. I would also like to extend my thanks to Heinke Buhtz and Conny Burghardt for their help in offering me the lab facilities during my lab work.

I would also like to extend my sincere thanks to the Alfred Wegener Institute (AWI) on Heligoland, who generously provided me with a lab bench to work on site. My gratitude goes to Dr. Jochen Girschke from the institute for avian research on Heligoland for his support in mouse collection on the upper land. Special thanks to all the people on Heligoland who participated and/or provided agreement for trapping, in particular, the Kläranlage, Backfabrik am Hafen, Jugendherberge, Firma Hagemann, Leuchttrum, Schwimmhalle, Flugplatz, EDEKA Oberland, Polizeibüro, Krankenhaus, Sportplatz and all the families who welcomed our traps in their gardens and in the vicinity of their houses, Karen Wilshire und Maaten Boersma, Wichmann family, and Hackmeyer Haus.

I would like to thank all the members of the Max Planck Institute in Plön for the friendly working atmosphere. Special thanks should be given to Dr. Maria Abu Chakra who provided helpful comments for the improvement of this thesis.

Lastly, I would like to thank my beloved extended family in Khartoum, even if far away, always close to me, most important to my mother, my brother and sisters who have continued to be on my side, motivating, and asking; I will always be indebted to you. I would also like to acknowledge the support provided by my little family during my PhD, I am indebted to you for your love and patience which all put me into the challenge to pursuing my studies.

Zusammenfassung

Seit die Evolutionsbiologie mit der Evolutionstheorie selber ihren Anfang fand, steht das Erforschen von Inseln in ihrem Mittelpunkt, um die molekularen Mechanismen, die Evolution, Adaption und Artbildung unterliegen, zu verstehen. Invasive Arten sind besonders interessant hinsichtlich der Erforschung von Anpassung, da sie Hinweise darüber geben, wann und wie die Kolonisation der Inseln stattgefunden hat.

Im Fokus dieses Projekts liegt die Untersuchung evolutionärer Prozesse, welche zusammen mit geographischer Isolation die bereits phänotypisch beschriebene, auf Helgoland vorkommende Hausmaus *M. m. helgolandicus* geformt hat. Helgoland ist eine kleine Insel im Süd-Osten der Nordsee und wurde im frühen 15. Jahrhundert vom Menschen besiedelt. *M. m. helgolandicus* wurde das erste Mal von Zimmermann 1953 beschrieben. Seit diesem Zeitpunkt wurde angenommen, dass *M. m. helgolandicus* eine eigene Unterart von *M. musculus* darstellt, jedoch mit deutlichen morphologischen Unterschieden zu *M. m. domesticus*, welche auf dem westeuropäischen Festland vorkommt.

In dieser Studie wurden vier, nucleäre diagnostische Marker (*Abpa*, *D11 cenB2*, *Btk* und *Zfy2*), sowie die charakteristische, relative Schwanzlänge (TBLR) verwendet, um diese Mäuse von Mäusen der Unterarten *M. m. musculus* und *M. m. domesticus* zu unterscheiden. Zusätzlich wurden mithilfe der mitochondrialen Kontrollregion (D-loop) und 21 Mikrosatelliten die Populationsstruktur bestimmt und somit mögliche Kolonisationswege untersucht. Darüber hinaus sollte die Sequenzierung der gesamten mitochondrialen DNA von 11 Individuen über den Zeitpunkt der Kolonisation Auskunft geben. Hierbei wurde die Mutationsrate von Mäusen des Kerguelen Archipelagos zu Grunde gelegt, da von ihnen die Kolonisationsgeschichte bereits sehr gut erforscht ist. Zur Beschreibung von möglichen von *M. m. domesti-*

cus oder *M. m. musculus* introgressierten Haplotypen wurden sowohl ganze Genome von drei weiteren *M. m. helgolandicus* Individuen hinsichtlich Einzel-Nukleotid-Unterschieden (SNPs) analysiert, als auch die Daten von zwei möglichen Quell-Populationen. Diese Studie betrachtet aber auch morphologische Merkmale von *M. m. heloglandicus*. Im Speziellen wurden die Mandibeln zwischen Individuen von Helgoland und vom Festland verglichen, um erste Hinweise auf morphologische Anpassung an das Leben und die Ernährung auf der Insel, festzustellen.

Basierend auf den Ergebnissen der diagnostischen Marker, relativer Schwanzlänge, Mikrosatelliten und mitochondrialer DNA kann *M. m. helgolandicus* hauptsächlich *M. m. domesticus* zugordnet werden. Die helogländer Mauspopulation weist eine, verglichen mit Mauspopulationen vom Festland, geringe genetische Diversität auf. Die mitochondriale DNA zeigt hauptsächlich einen, spezifisch auf Helgoland vorkommenden Haplotypen und einen selten vorkommenden Haplotypen, der nur von einem Individuum getragen wird und wahrscheinlich introgressiert ist. Demzufolge sieht es so aus, als habe eine einzige Kolonisation der Insel vor ein paar Hundert Jahren stattgefunden. Trotz der sehr isolierten Lage der Insel sind einzelne seltene Fälle von Migration vom Festland zu beobachten, wobei allem Anschein nach ist die Population auf Helgoland beständig gegenüber Einwanderern und behält ihren "eigenen" Genpool. Zudem weist *M. m. helgolandicus* verlängerte Mandibeln auf, ein Merkmal, das nur auf Helgoland zu finden ist. Sehr wahrscheinlich ist dies ein Zeichen für Anpassung an veränderte Nahrungsquellen in einer neuen Umgebung, hin zu einer vermehrt carnivoren Ernährung. Das Genom der helgoländer Mäuse ist sehr durchmischt mit *M. m. musculus* Haplotypen, dies könnte auf eine mögliche Hybrid-Speziation während der Kolonisation hinweisen.

Abstract

Islands are a center of interest in evolutionary biology since the emergence of evolutionary theory itself. They are studied to understand the molecular mechanisms of evolution, adaptation and speciation. Invasive species are of particular interest since they may garner clues and evidence about the processes that took place during the onset of colonization and for understanding mechanisms of adaptation.

The aim of this project is to study the evolutionary processes that altogether with isolation shaped the phenotypically known house mouse *Mus musculus helgolandicus* inhabiting the island of Heligoland. Heligoland is a small island located in the South-East corner of the North Sea and was first colonized by humans in the dawn of the fifteenth century. *M. m. helgolandicus* were first described by Zimmermann in 1953. Since then they have been thought to form a separate subspecies, which is morphologically different from its continental counterpart *M. m. domesticus* inhabiting the Western European region.

Here, four nuclear diagnostic markers (*Abpa*, *D11 cenB2*, *Btk* and *Zfy2*) and the discriminatory relative tail length (TBLR) were used to differentiate these mice from the other two subspecies *M. m. musculus* and *M. m. domesticus*. In addition, the possible routes of colonization and population structure for the invasive mice were investigated using mitochondrial (mt)D-loop DNA sequence and (21) microsatellite loci respectively. Furthermore, whole mtDNA genome was sequenced for 11 individual mice to estimate the onset of colonization on the island from the calculation of mutation frequency in comparison to that of house mouse from Kerguelen archipelago, which has a documented colonization history. Moreover, the whole genome sequence of three individuals was generated and analysed for single nucleotide polymorphisms (SNPs) which were then used along with data for two po-

tential source populations from the two subspecies inhabiting Europe to assign the possible patterns of introgression of haplotypes in *M. m. helgolandicus*. This study also revisits the morphology of *M. m. helgolandicus*, in particular, the mandible to assign morphological differences among Heligoland mice on one side and among Heligoland and mainland populations on the other side.

Based on the results from diagnostic markers, relative tail length, microsatellites and mtDNA analyses, *M. m. helgolandicus* are predominantly of *M. m. domesticus* origin. *M. m. helgolandicus* population on Heligoland exhibited low genetic diversity compared with other populations from the mainland. The mtDNA data shows that there is a major mtDNA haplotype specific to Heligoland and a minor haplotype represented by a single individual presumably introgressed. Hence, there was a single primary colonization into the island a few hundred years ago and more interestingly, the isolated island shows a case of recent migration from the mainland revealing a signal of refractory to immigration. *M. m. helgolandicus* displays an elongated mandible which is distinctive for Heligoland. Most likely it was acquired by adaptive forces due to diet changes from a novel environment, with particular a shift to carnivory. The genome is highly intermixed with *M. m. musculus* haplotypes, pointing to a possible hybrid speciation scenario during the colonization phase.

Declaration

The project was initiated by my supervisor Prof. Dr. Diethard Tautz. And we both designed the project layout through rounds of discussion.

Chapter 2

A total of 8 mice samples were provided by Dr. Jochen Dierschke from the Institute for Avian Research in Heligoland.

Chapter 3

Dr. Anja C. Schunke provided the body measurement data for *M. m. musculus* and *M. m. domesticus* subspecies, she also provided the specimens from old mice collections through loans. She introduced me to the morphometrics practical work and landmarking approach analysis, however, all the work was completely achieved by me. The mandible raidographs for different *Mus* populations used for comparison were obtained from Dr. Louis Boell, however, the landmarking and further steps were analysed by me.

Chapter 4

The whole genome sequencing was initiated by Prof. Dr. Diethard Tautz. The reference data set for *M. m. musculus* and *M. m. domesticus* subspecies used for the introgression analysis was provided by Dr. Fabian Staubach. Analysis of of the HapMix results was conducted at some stage with an R script also provided by Dr. Fabian Staubach.

List of Figures

List of Tables

1 General introduction

1.1 Evolution on islands

Since the dawn of Darwin's evolutionary theory in the nineteenth century, biologists have directed their research and efforts to decipher how species evolve. Studying evolution on islands has played an important role in the development of our current knowledge of how and why evolution occurs. Islands and their inhabitants show distinct forms from their mainland counterparts, mostly because they have been isolated on the small islands (relict of previously widespread forms) and subsequently acquired random genetic changes (drift) and/or adaptation to the new environment. Islands as discrete pieces of the environment, which are isolated from the continental large scale processes such as gene flow, may present precise understanding of microevolution, speciation, and adaptive radiation (Berry, 1998).

The development of evolutionary theory through natural selection was based on the study of the Galapagos by Darwin and the study of the Malay archipelago by Wallace, who were attracted by the special fauna found on such islands. The extraordinary importance of islands for scientists and why they are perfect pieces for research emerged from the mere fact that each island on its own can be a more homogeneous area than the mainland. Moreover, that islands are the exclusive home of certain species, the so-called endemic species. The endemic species, which in most cases are not so much different from their counterparts or their closest relatives on the mainland or another nearby island provided the evidence for the theory of geographic speciation. When a population becomes isolated on an island or in an insular location on a mainland it acquires during this isolation, specific genetic changes to fulfill the adaptive requirements to novel environmental conditions. As

a result of such changes a new species can be found. Such an example is Madagascar, which became isolated from the Godwanan land mass during the Cretaceous period. Although Madagascar is not far from the African continent, many groups of birds and mammals that arose thereafter on the mainland failed to migrate. Hence, Madagascar followed an independent evolutionary path, where endemic species diversified into ecological niches that were occupied by other species groups on the mainland (Losos and Ricklefs, 2009; Mayr, 1967).

If an island is inhabited through colonization rather than survival as relict, new environmental and ecological conditions must be considered as important selective agents stressing individuals (Berry, 1996). Colonization is usually initiated by a small number of individuals, constituting partial information of the total genetic variability of their parental species. The colonization event results in a demographic bottleneck (enormous decrease in numbers) associated with potential major effects on genetic variation as a consequence to intermittent genetic drift. As a result of genetic depletion, post-colonization populations may be less capable to adapt to sudden environmental changes leading to extinction in some cases. The reaction of the colonizing population depends largely on the alleles and their frequencies in the founders or more specifically on its gene pool size, and when they increase in number and expand, they will be more resilient to immigration events from the mainland (Berry, 1996). The exposure to a new environment on an island with a different selection pressure could lead to large evolutionary changes that can further increase, if correlated with an adaptive shift. This could result in major genetic changes and the formation of a new species (Mayr, 1967). The evolution and diversification of species on islands could occur either under an allopatric or a sympatric model (Figure 1.1).

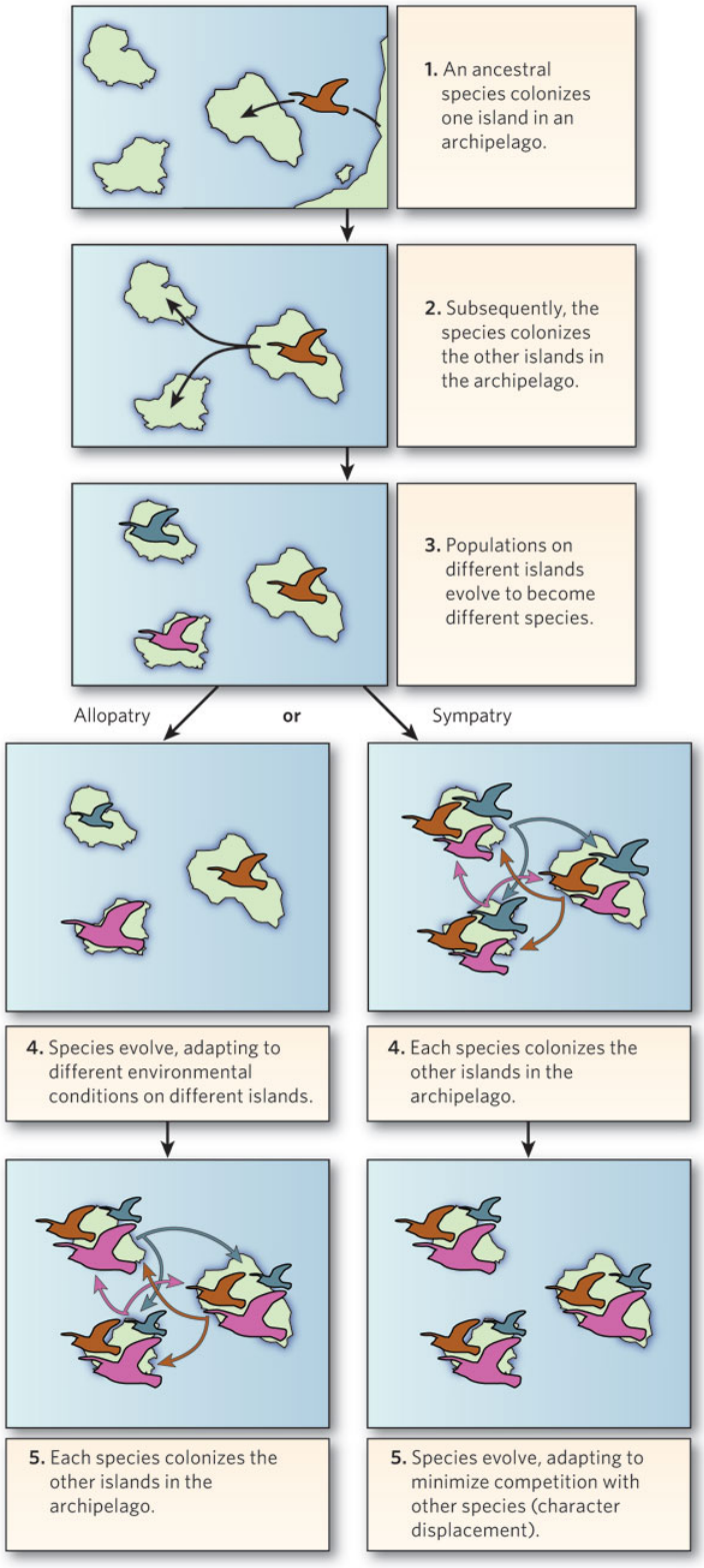


Figure 1.1: Cartoon illustrating the colonization steps of an archipelago by an ancestral species (upper panel). The lower panel shows the allopatric speciation model (left) and the sympatric speciation model (right). Figure was taken from Losos and Ricklefs (2009).

The best-known example for allopatry (geographic isolation) is the evolutionary radiation of Darwin's finches, observed by the production of 13 species in the Galapagos archipelago. On the other hand the sympatry model is more difficult to prove, since one would have to show that a preceding phase of allopatric speciation is highly unlikely (Losos and Ricklefs, 2009). However, there is increasing evidence for sympatric speciation on islands as well, e.g. palm trees on Lord Howe island (Savolainen et al., 2006) and *Anolis* lizards on the islands of the Greater Antilles in the Caribbean Sea (Losos et al., 1998). There have been some experimental studies to address different cases where populations failed to establish colonies or became extinct. A study on populations of house mice (*M. m. domesticus*) was conducted by (Berry et al., 1982) on two Shetland islands where populations were unsuccessful as a result to lack of food in one case and to competitors in another. Detailed studies focusing on persistence (the colonizing ability and extinction rate) found that the species differed in their migration capability with larger species having an advantage due to larger body sizes. However the smallest species have proved success when invading an empty island mainly due to a decrease in food requirement (Peltonen and Hanski, 1991).

Notwithstanding, the patterns of genetic variation on islands are lower than on the mainland for the same given subspecies, it is worth focusing on island populations which had been under the influence of various evolutionary forces and environmental changes. The growing interest in expanding research on islands has focused on determining the relative importance of factors related to the percentage of endemic species on islands such as the island size, the richness of it's fauna and the degree of isolation from the nearest mainland coast or the nearest island (Mayr, 1967). Hence, more empirical studies will unravel the evolutionary importance of islands.

1.2 House mice

The house mouse has mainly been popularized as a successful model organism for biomedical research. However, this organism is also ideal for evolutionary studies, since it has expanded across the world in several waves, thus adapting to many new niches. The house mouse *Mus musculus* originated on the Indian subcontinent a million years ago (Boursot et al., 1993) and was transported by people and spread throughout the globe in different times (Cucchi et al., 2005). These mice are intimate commensals that have been carried on ships and expanded their natural range far beyond Eurasia and diversified into three subspecies, *M. m. castaneus*, *M. m. domestiucs* and *M. m. musculus* (Din et al., 1996). *M. m. musculus* colonized central and eastern Europe and northern Asia, *M. m. castaneus* in southern Asia, and *M. m. domestiucs* has been introduced to Africa, Americas and Australia by Western European ships (Figure 1.2); (Boursot et al., 1993; Frazer et al., 2007; Guénet and Bonhomme, 2003; Prager et al., 1996). Hybrid zones in regions of subspecies contact were also established (Boursot et al., 1993).

The onset of the house mouse westward expansion mainly in the Fertile Crescent of the Middle East is dated back to 10,000 years BP (Rajabi-Maham et al., 2008). This region is considered the cradle for the commensalism with humans and the onset of expansion, which is in concordance with zooarchaeological findings (Cucchi et al., 2005). Along with human expansion and settlements, the house mouse subspecies were able to invade new regions such as Madeira and Faroe archipelagos in the Atlantic ocean (Boursot et al., 1993; Gündüz et al., 2001), Japan in the North Pacific (Moriwaki et al., 1986), and Kerguelen archipelagos in the South Indian ocean (Berry et al., 1978; Hardouin et al., 2010).

Fossil records showed that mice arrived in Spain around 3,000 years ago and colonized other European regions with human migrations (Cucchi et al., 2005). The

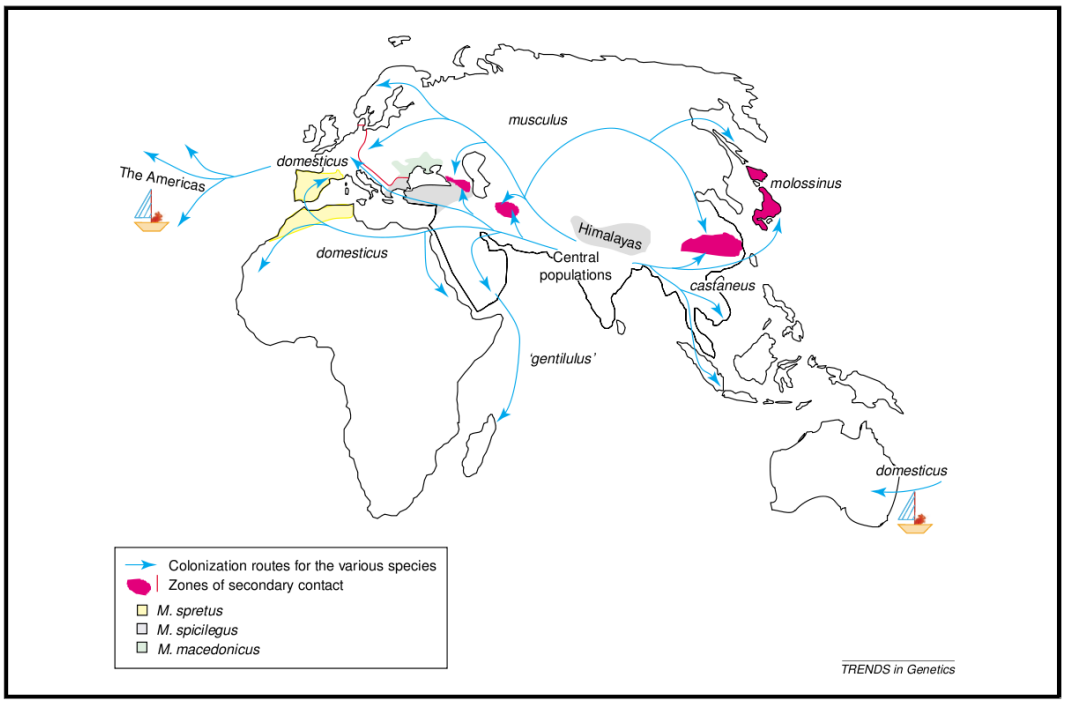


Figure 1.2: Colonization routes of the house mouse and closely related species (taken from Guénet and Bonhomme (2003)).

distribution of the house mice in Europe (Figure 1.2) is shaped by the two genetically differentiated subspecies *M. m. musculus*, which inhabits central and Eastern Europe and *M. m. domestiucs*, inhabiting the Mediterranean basin and Western Europe (Auffray et al., 1990; Boursot et al., 1993). The house mouse *Mus musculus* is well known as commensal to humans and this special feature was the basis for studies linking the routes of distribution and colonization history of these mice with documented human movements (Pocock et al., 2005; Searle et al., 2009a). For example studies on house mice from the Faroe island and from the North Atlantic region documented supportive findings that the genetics of house mice is likely to mirror the population genetics of humans (Jones et al., 2011, 2012).

In addition, classical inbred laboratory mice have been widely used for both genetic and medical research. However, they lack the variety of genetic polymorphisms due to the fact that they are derived from a handful of founders and heavily inbreed. Accordingly, recent research is looking for new strains produced from crosses of dif-

ferent wild *Mus* species, which will offer advantageous information for evolutionary analysis (Guénet and Bonhomme, 2003).

The mouse laboratory strain C57/BL6J was fully sequenced and is used as the reference genome to represent the house mouse. The genome of C57/BL6J strain constitutes genomic composition of the three subspecies *M. m. musculus*, *M. m. castaneus*, and *M. m. domestiucs*, see Figure 1.3 (Frazer et al., 2007; Wade et al., 2002; Yang et al., 2007). However, it is unclear how much of this composite nature is due to crossing between subspecies at the time of the establishment of the strains, or whether it reflects natural introgression (Staubach et al., 2012).

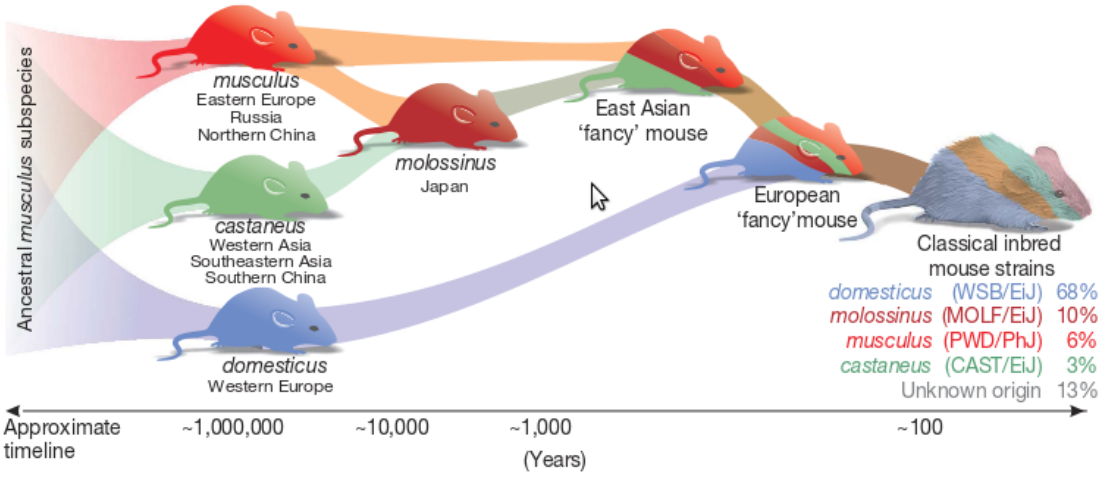


Figure 1.3: Cartoon showing the ancestral *M. Musculus* subspecies and the relative contributions of the *Mus musculus* subspecies to produce varieties of mice with different coat colours and haplotypes in classical strains. Figure adapted from Frazer et al. (2007)

1.3 The hybrid zone

The concept of species or speciation is the evolutionary process that is being delineated by a degree of reproductive isolation. Reproductive isolation prevents the emerging species from freely exchanging genes. The house mouse *M. musculus* provides an excellent model system for speciation research; its subspecies show intermediate levels of reproductive isolation which ease their crossing in the lab and also

allow some to hybridize in nature (Teeter et al., 2007). Besides that, the recent development in genome sequencing and the availability of house mouse data, all are being put forward to expand this area of research.

The taxonomy of the genus *Mus* provided extensive studies of the house mouse in the laboratory where hybrids are produced relatively easy (Berry and Scriven, 2005). Although those hybrids are rare in the wild (Forejt, 1996), there are hybrid zones along the regions of contacts where these mice have not been completely reproductively isolated such as *M. m. musculus* and *M. m. castaneus* in the far east where their hybrids are known as *M. m. molossinus* in Japan (Yonekawa et al., 1988). On the other hand in Central Europe *M. m. domestiucs* and *M. m. musculus* form a narrow hybrid zone illustrated in Figure 1.4 that extends from Scandinavia to Bavaria and runs through variable environments, without obvious geographical barriers (Boursot et al., 1993; Sage et al., 1993; Ďureje et al., 2012).

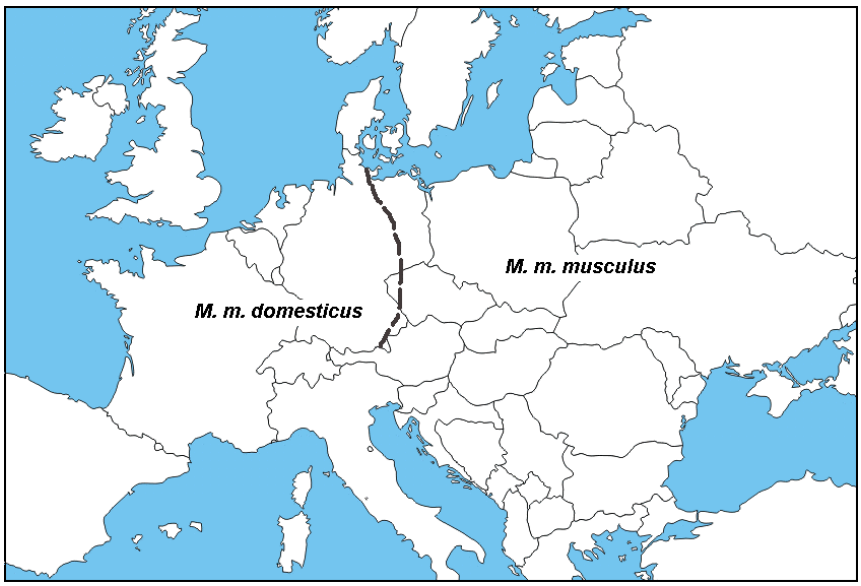


Figure 1.4: A map showing the hybrid zone (zone of contact) of the *M. m. musculus* and *M. m. domesticus* (the map was obtained from (http://d-maps.com).

The European hybrid zone has been the focus of genetic studies using different markers ranging from autosomal to maternal and paternal based genomic regions (Payseur and Nachman, 2005; Raufaste et al., 2005; Sage et al., 1993). The hybrid

zone has been studied along the extended zone e.g. in Denmark, Germany, and Czech Republic, however there are still cryptic relationships of mice in under-represented geographical regions of Europe that should be the aim of future research (Dod et al., 1993, 2005; Macholán et al., 2007; Payseur et al., 2004; Prager et al., 1993; Raufaste et al., 2005; Sage et al., 1993, 1986; Teeter et al., 2007).

1.4 The island of Heligoland

The island of Heligoland (54° Í1 N, 07° 5́3 E), is a small island in the North sea in North Western Germany and consists of two small islands (Figure 1.5 & 1.6). The main island which is known as Heligoland is a Triassic red sandstone rock, 1 km^2 long, 61 m high and 46 km away from the German coast (Spaeth, 1990). The smaller island, Dune Island, which was formerly connected to Heligoland is a sandy island with low sand dunes and lies about 1 kilometer to the east of Heligoland. The island of Heligoland is inhabited by >1000 people and has two major distinctive land parts. The upper land is mainly surrounded by the sandstone cliffs and the lower land is completely covered by the island village (Dierschke et al., 2010; Ritsema, 2007).

The birth of Heligoland is the result of different geological stages: About 230 million years ago the sea covering what is known nowadays as the North sea evaporated during a hot climate leaving large deposits of salt in the South-eastern area. Then some 220 and 210 million years ago the shell lime deposits were formed mainly from the shells of organisms. In the last 50 million years Heligoland came to the surface as a result of salt drifting and lifting up of the smaller particles from the Zechstein period. Heligoland became an island about 3,000 years ago when the sea level reached stable level and the different erosion waves resulted in a reduced island area and height (Ritsema, 2007).

The geographical location of Heligoland gave it a turbulent political and military

Figure 1.5: Map showing the location of Heligoland island on the North Sea (https://maps.google.de/maps

history and hence was a pawn of great powers. To illustrate, from c.1400-1807 Heligoland was under the Hanseatic cities and Schleswig-Gottorp and the Kingdom of Denmark. The two powers ruled the island in consequent rounds either through war and planned attacks or by taking over islanders who worked mainly on fishing, hunting and wrecking. The ruling power of the island, when had no choice, either surrendered or ceded the island to the attacking power. The oldest document on Heligoland dated the occurrence of the first church back to 1435 which is in line with the first colonization of people from the Kingdom of Denmark and from Schleswig-Gottorp (Ritsema, 2007).

In 1807 Heligoland was taken by the British as a location for military activity and during this time the island was at some point isolated from the German and Danish mainlands. In 1890 Heligoland was ceded to Germany by Britain in the treaty of Heligoland-Zanzibar which was also known as the Anglo-German Heligoland Treaty. During the German ruling period the island was also used as a military harbour mainly during the world wars where civilian population were in exile on the mainland

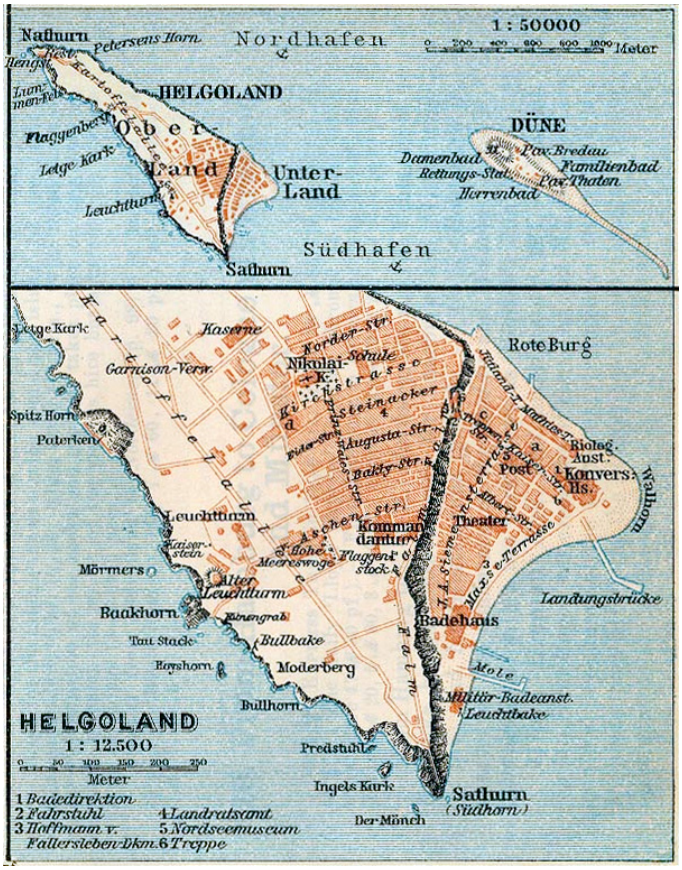


Figure 1.6: Map of Heligoland from 1910 showing the structure of the island. (http://www.lib.utexas.edu/maps/historical/baedeker_n_germany_1910/)

and the island was under extensive military activities. In the Second World War (1944-1947) the island was evacuated and was recolonized in the fifties (Drower, 2002; Olson and Shadle, 1991).

1.5 Possible colonization routes of house mouse into Heligoland

The main island of Heligoland is colonized by the so called house mouse *M. m. heligolandicus* described to be phenotypically different from the two subspecies inhabiting Europe *M. m musculus* and *M. m. domesticus* (Zimmermann, 1953). Hence it is considered as a separate subspecies and their classification was based on mor-

phological features e.g. skull shape and measurements (Reichstein and Vauk, 1968; Zimmermann, 1953).

Even though there might have been subsequent migration to Heligoland along with humans, the colonization waves are more likely to be from the same subspecies inhabiting Western Europe *M. m. domesticus* and less likely from *M. m. musculus* subspecies. Another scenario could be that the island was colonized from the hybrid zone where *M. m. musculus* and *M. m. domesticus* are in contact. However, this scenario is unlikely, given that both Denmark and Scandinavia were colonized by house mouse from Northern Germany *M. m. domesticus* and hence colonized Heligoland afterwards (Prager et al., 1993).

Revisiting the history of Heligoland can ease the interpretation of possible colonization routes and different demographic events that may have taken place on the island. Here I present possible migration routes for house mouse colonization of Heligoland deduced from the history of the island which was ruled by Danish, British, and German powers. The map shown in Figure 1.7 provides insights into three different possible routes of house mouse colonization. The Roman numbers (I, II, III) on the figure represents the periods in a chronological order and the arrows show the directions of the suggested routes.

The elucidation of population genetic structure of the *M. m. heligolandicus* in relation to other known subspecies will resolve the local controversy of the origin of these mice. The study of the mouse population from Heligoland can shed light not only on the genetic structure of these mice, but also on their history as well as the mechanisms and processes of island colonization and adaptation.

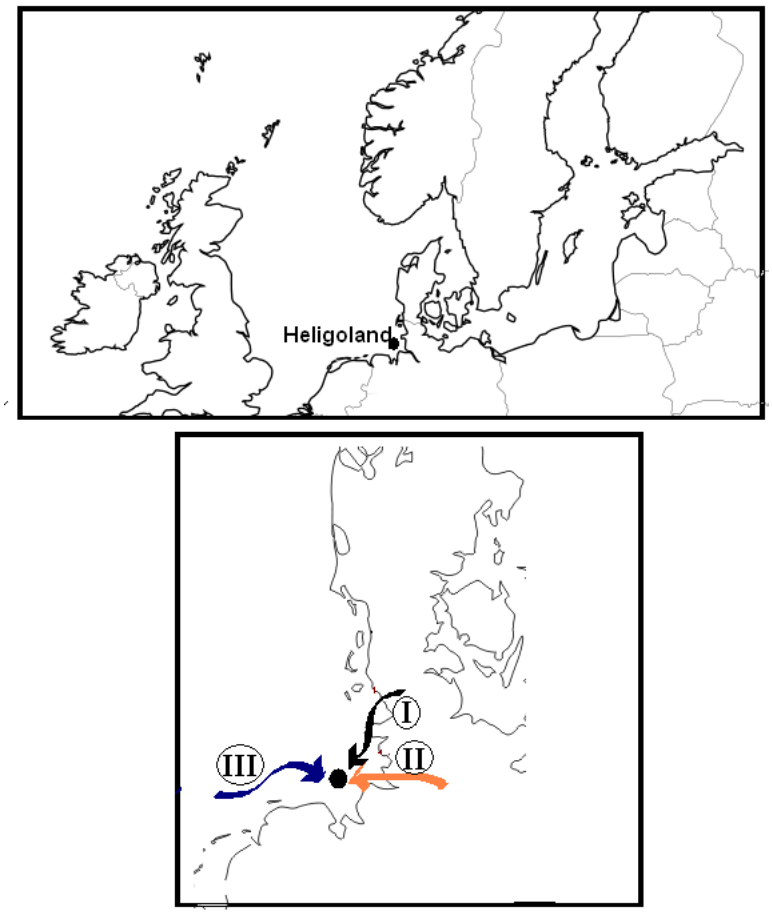


Figure 1.7: Possible migration routes for house mouse colonization of Heligoland, the arrows outline the origin of founder individuals emanating from Denmark, Holstein, and Great Britain. (http://www.freeworldmaps.net/printable/europe/)

1.6 Genetic studies on *M. m. heligolandicus*

M. m. heligolandicus is under-represented in previous genetic studies and the studies documented included few samples and were limited to a single mode of analysis. Figueroa et al. (1987) conducted a large survey on populations from the two subspecies *M. m. musculus* and *M. m. domesticus* along with 6 individual mice from Heligoland and other species of the genus *Mus*. The study was based on the analysis of two regions on chromosome 17, D17Tu1 and D17Tu2, the former was mapped to the centromeric region of chromosome 17 and the latter to the S region of the ma-

jor histocompatibility complex (MHC)H2. The most striking observation from their study was that the distribution of D17Tu1 probe was subspecies specific among populations from the two studied subspecies. *M. m. domesticus* populations tended to show extensive polymorphism in contrast to the monomorphism observed among *M. m. musculus* populations. More interestingly, the samples from Heligoland shared two patterns for the D17Tu1 probe D17Tu1a and D17Tu1d similar to that observed among populations from *M. m. domesticus* with one exception for D17Tu1g pattern which was found only in Heligoland, Australia and in inbred mice but not among the other studied wild mice. On the other hand, the distribution of the D17Tu2 was found to be widely distributed among the two subspecies sharing five patterns and the samples from Heligoland showed only one pattern D17Tu2b (Figueroa et al., 1987).

Further studies were based on the analysis of mitochondrial DNA and included a single mouse from Heligoland. The first study by Sage et al. (1990) based on the distribution of the fragment patterns of mtDNA using restriction enzymes in Europe, North Africa, Middle East, and the Americas along with one mouse from Heligoland. Their results assigned the sequence from Heligoland to a common clade distributed in Western European region. Prager et al. (1993) used mtDNA control region diversity and information of the colonization of Scandinavia by the house mice from Northern Germany (East Holstein). Their survey included the same mtDNA sequence from Heligoland that was used by Sage et al. (1990). They found that the house mouse from Schleswig Holstein in Northern Germany colonized different Scandinavian regions. Also the sequence analysed from Heligoland had a distinct haplotype not shared with any of the analysed sequences except that it was characterized by the addition of an 11bp direct repeat. This repeat was shared by sequences from Northern Germany, Denmark and Scandinavia. The repeat was also observed in sequences from Holstein and Jutland hybrid zones, where most of these

mice were assigned to *M. m. domesticus* based on mtDNA, and assigned to *M. m. musculus* based on their nuclear genome (Prager et al., 1993).

1.7 Aims of the study

The whole study aims at deep insights into the origin of *M. m. helgolandicus* and was designed to cover different aspects of recent biological research schemes. This study aims to explore the evolutionary processes that shaped the origin and genetic structure of *M. m. helgolandicus*. Furthermore, understands in details which subspecies have been part of their genetic composition and if their original haplotype structure had evolved as a result of isolation.

The study consists of three parts. The first part is a detailed analysis of the genetic composition of the Heligoland mice using different molecular markers, such as diagnostic molecular markers, microsatellite typing and the distribution of mitochondrial DNA control region haplotypes. The mtDNA is considered in the context of previously published sequences to inspect the demographic history and if frequent colonization invaded the island and shaped the recent genome composition of the mice. The second part, implements geometric morphometrics on the mandible. It is a landmarking approach on a contemporary mice collection along with two older collections collected in different time periods to asses the status of morphological adaptation. The third part is mainly concerned with the whole genome sequence of the house mouse from Heligoland. In particular, to assign signals of introgression and possible adaptation in their genome as a composite of the two well known subspecies

2 Genetic analysis and insights into the origin of *M. m. helgolandicus*

2.1 Introduction

Recent studies have focused on the origin of house mouse populations inhabiting various regions in an attempt to reconstruct their colonization history and to find molecular signatures resulting from recent expansions of the subspecies (Bonhomme et al., 2010; Rajabi-Maham et al., 2008). In particular, some studies have been concerned with patterns of colonization on islands e.g. Madeiran mice (Förster et al., 2009), British Isles in the north Atlantic ocean (Searle et al., 2009b), Kerguelen archipelagos in southern Indian ocean (Hardouin et al., 2010), and New Zealand in the west southern pacific ocean (Searle et al., 2009a). The revealed patterns of the house mouse colonizations based on molecular markers provided evidence for the suggested expansion routes in Europe, namely, the Mediterranean route and the Bosphorus/Black Sea route (Rajabi-Maham et al., 2008).

Recent phylogeographic studies on house mouse mtDNA sequences, supports previous findings linking patterns of house mouse phylogeography and human activities during human historical movements, e.g. Iron Age and Viking Age (Jones et al., 2012; Searle et al., 2009a,b). Besides that, the house mouse genetic diversity on islands within an archipelago is positively correlated to human population size, e.g. the Mykines island within Faroe archipelago (Jones et al., 2011).

Nuclear markers with fixed differences between *M. m. domesticus* and *M. m. musculus* subspecies have been used widely, to differentiate between subspecies and also to determine any forms of locus specific introgression among populations (Dod

et al., 2005; Lanneluc et al., 2004; Munclinger et al., 2002). These markers are *Abp*, *D11 cenB2*, *Btk*, and *Zfy2*. Androgen-binding protein *Abp* is a member of the secretoglobin family which is encoded on mouse chromosome 7. It is a major component of saliva that has been proposed to be part of mate recognition in mice (Dod et al., 2005; Laukaitis et al., 2008). *D11 cenB2* is a sequence closely linked to the centromere on mouse chromosome 11 and is considered to play a role in chromosomal segregation (Lanneluc et al., 2004). *Bruton agammaglobulinemia tyrosane kinase* gene *Btk* is an X-linked gene close to a B1 insertion in the mouse X-chromosome. The insertion is known to be fixed in *M. m. domesticus* populations but is absent in *M. m. musculus* (Munclinger et al., 2003, 2002). *Zfy2* is a Y-linked gene, located within the last exon of the Zinc finger protein 2. This marker is typed for the presence or absence of an 18 bp deletion which is known to be fixed in *M. m. musculus* and absent in *M. m. domesticus* (Munclinger et al., 2002; Orth et al., 1996).

Microsatellites are short repetitive sequences scattered throughout the genome and have proved to be versatile markers with the potential to document recent events. This is due to their high mutation rate. Their polymorphisms originate from variability in length rather than in the primary sequence. Microsatellites rapidly became the markers of choice in genome mapping, linkage studies, and subsequently in population genetics (Ellegren, 2004; Tautz, 1989). Teschke et al. (2008) designed a large set of microsatellites using a pooling approach described in (Thomas et al., 2007). The assessed microsatellites (915 loci) were chosen to cover the whole genome of the house mouse and they were analyzed in population pairs, one from *M. m. musculus* and the other from *M. m. domesticus*. Combined sets of these microsatellites have been used recently for population structure analysis and determination of house mouse genetic diversity within and among populations (Hardouin and Tautz, 2013; Linnenbrink et al., 2013).

Phylogeographic studies started to draw attention in the late 1980s and early

1990s and since then, they have become one of the central pillars of population genetics. Their application involves the estimation of a tree or network using phylogenetic reconstruction (Nielsen and Beaumont, 2009). The mitochondrial DNA (mtDNA) is a part of the genome that is inherited in a non-Mendelian manner only maternally (from the mother) and is known as a fast-evolving region with a considerable high mutation rate (Ballard and Whitlock, 2004). It has been extensively used in the last four decades as a phylogenetic tool, mainly to document matrilineal footprints and reconstruct demographic history of both populations and species. In particular, the mtDNA control region has been of great importance to phylogeographic studies of the house mouse *M. musculus*, to assign recent versus old colonization events and also to provide evidence for single versus multiple waves of colonization (Bonhomme et al., 2010; Rajabi-Maham et al., 2008; Searle et al., 2009a,b).

In this study, I use the four diagnostic nuclear markers that are well known to differentiate between *M. m. domesticus* and *M. m. musculus*. I type microsatellite loci to identify the genetic composition and population structure of *M. m. helgolandicus* within different mainland populations representing *M. m. domestiucs* from Western Europe and *M. m. musculus* from Asia. Additionally, I use mtDNA control sequences for these mice with published sequences representing different populations of *Mus musculus* species. More in detail, to infer their initial pattern of colonization, to pin point if it was led by single/multiple colonizing events and see if there is evidence for successful secondary migration waves. Further more, I determine aspects of colonization history of *M. m. helgolandicus* based on matrilineal movements. Specifically I estimate the colonization age of these mice on Heligoland from the mutation frequency assigned from the sequenced mtDNA genomes compared to that of the Kerguelen mice studied by (Hardouin and Tautz, 2013).

2.2 Materials and methods

2.2.1 Sample collection

A total of 17 mouse samples from Heligoland island were collected on the mainland (Heligoland) in the periods 2004-2012 by researchers at the Institute for Avian Research and by ourselves in the summer of 2012. The collection was done in a single field trip from two localities defined as upper and lower lands (see Figure 2.1). The collection process did not apply a specific sampling scheme because the island is too small to allow to follow such a scheme.

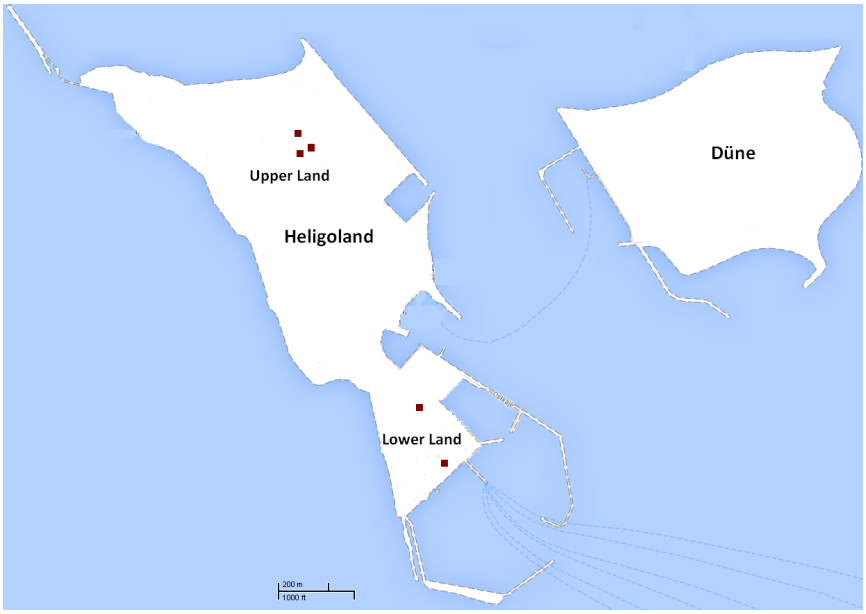


Figure 2.1: Map of the Heligoland (Heligoland to the left and Düne to the right) with illustration of sampling sites on the mainland. (https://maps.google.de/)

Mice were trapped live and for each mouse; body weight, body measurements and photos for skin coat color were taken. The mice were sacrificed by CO_2 inhalation and were dissected on site and organ tissues from each mouse were prepared and later each mouse was preserved in absolute Ethanol for morphological analysis and future use.

2.2.2 DNA extraction

Total genomic DNA was extracted from mice tissue samples (mostly liver) using a salt extraction protocol. The tissue was incubated in a lysis buffer with Proteinase K [0.20mg/mL] in concentration at 55° C overnight on a shaking platform. 500μL Sodium Chloride [4.5M] was added to break down fat and proteins of the cell membrane and to strengthen phosphate bonds of the DNA molecule. Then 300μL chloroform was added to separate the DNA from the protein phase. DNA was precipitated using [0.7 of the total volume] pure Isopropanol and the DNA pellet was washed with 500μL [70%] Ethanol and dissolved in 30μL Tris-EDTA (TE) buffer prepared as follows [1M Tris: Tris (hydroxymethyl) aminomethane (MWT 121.4g/mol) 60.57g) in 0.5L deionized water, pH adjusted to 8.0 using HCl and 0.5M EDTA: Diaminoethane tetraacetic acid (MWT 372.2 g/mol) 18.6g in 100mL deionized water, pH adjusted to 8.0 using NaOH]. The concentration of DNA was measured using Nanodrop 1000 version 3.0.

2.2.3 Diagnostic nuclear markers

Extracted genomic DNA was used to analyze four nuclear genetic markers that are known to differentiate between *M. m. domesticus* and *M. m. musculus*. Specific primers for each genetic marker were used and each sample was scored by PCR and gel electrophoresis. The *Abp* marker was tested for PCR subspecies-specific alleles as in (Dod et al., 2005; Laukaitis et al., 2008). *D11 cenB2* was typed for PCR subspecies-specific alleles as in (Lanneluc et al., 2004). *Btk* marker was typed and scored as in (Munclinger et al., 2002) for the presence or absence of the B1 inserstion in the *Btk* gene found on chromosome X. The *Zfy2* Marker was tested for the absence or presence of an 18 bp deletion found on the Y chromosome following the protocol of (Munclinger et al., 2002). The details of these markers and the

primers used are all listed in supplementary Table 1.

2.2.4 Microsatellite typing

I chose 21 microsatellites (provided in supplementary Table 2) from (Teschke et al., 2008) to genotype the population from Heligoland with populations from France and Germany collected by Linnenbrink et al. (2013) and a population from Northern Germany (district of Ploen) collected by our colleagues at the Institute in 2007.

Of each primer set the forward primer was labeled with FAM or HEX dye on the 5' end. The PCR reactions were carried out in 5μL final volumes using 5ng DNA template and the standard protocols of QIAGEN Multiplex PCR kit. The PCR was programmed as follows: initial incubation step at 95°C for 15 min followed by 28 cycles at 95°C for 30s, 60°C for 1.30 min, 72°C for 1.30 min with a final extension at 72°C for 10 min. PCR products were diluted 1:20 in water. 1μL of the diluted PCR product was added to a previously prepared mixture of (10μL HiDi formamide) and (0.1μL of 500 ROX) size standard (Applied Biosystems, USA). A denaturation step was then performed with the following incubation times: 90°C for 2 min and 20°C for 5 min. The samples were analyzed using GeneMapper version 4.0 for Windows (Applied Biosystems, USA).

2.2.5 Microsatellite data analysis

The genotyped data from this study was combined with data for 3 populations from Kazakhstan, Germany and France, genotyped previously for the same microsatellite loci (Teschke et al., 2008). The total number of individuals analysed were 221 from 9 different populations. The number of individuals per population and their geographical locations are detailed in Table 2.1. The average number of microsatellite alleles per locus and the observed and expected heterozygosities were calculated per population using POPGENE program version 1.32 (Yeh et al., 1997).

Genetic distances among individuals were calculated using the proportion of shared alleles implemented in MSA version 3.15 (Dieringer and Schlötterer, 2003). The distance matrix obtained was converted into a tree using the Neighbor-Joining algorithm provided with the PHYLIP software package version 3.69 (Felsenstein, 1991) and graphically displayed in MEGA version 5.0 (Tamura et al., 2011). In addition, a tree based on the genetic divergence between populations was created using Cavalli-Sforza and Edwards (1967) chord distance D_c, which is implemented in POPULATIONS version 1.2.32 and available at (http://bioinformatics.org/project/groupid=84).

To decipher the population structure among the populations, I used the computer software STRUCTURE version 2.3.3 (Hubisz et al., 2009; Pritchard et al., 2000). STRUCTURE implements a multi-locus model-based clustering method that is used extensively to infer population structure and assign individuals to a predefined number of clusters. It assumes Hardy-Weinberg equilibrium within populations and linkage equilibrium among loci. Of each independent run I employed the admixture model for individual ancestry and the F model for allele frequency correlation and without prior information on localities of samples. I used 1,000,000 MCMC (Markov Chain Monte Carlo) repetitions and a burnin of 100,000 iterations with a number of clusters K from 1 to 12, each simulated ten times.

2.2.6 mtDNA control region

The mtDNA D-loop was amplified using the primers 5'CATTACTCTGGTCTTGTAAACC and 5'GCCAGGACCAAACCTTTGTGT from (Hardouin et al., 2010). The reactions were carried out in 10μL final volume with the following cycling parameters: 95°C for 15 min followed by 35 cycles of 95°C for 30s, 60°C for 1.30 min, 72°C for 1 min and an elongation step at 70°C for 15 min. Samples were then purified with Exonuclease/Shrimp Alkaline Phosphate (Exo/SAP) (USB Corp.) with

the following incubation conditions: 37°C for 20 min and 80°C for 20 min. Then each of the amplified sequences was subjected to cycle sequencing reaction using the following conditions: 96°C for 1 min followed by 29 cycles of 96°C for 10 sec, 55°C for 15 sec and 60°C for 4 min. The sequences were edited and visualized using CodonCode Aligner version 4.1.1 (CodonCode Corp.) and were adjusted to 852 bp according to (Bibb et al., 1981) and were aligned with previously published data obtained from (Ihle et al., 2006; Prager et al., 1993; Searle et al., 2009b) using MEGA version 5.0 (Tamura et al., 2011). The haplotype data was calculated using DnaSP version 4.50.3 (Rozas et al., 2003). The network was calculated using the Median Joining method and drawn with Network version 4.5.1.0 (Fluxus Technology Ltd.) (Joly et al., 2007).

2.2.7 Complete mtDNA sequencing

Mitochonrial genomes were sequenced for 9 mice using a set of primers described in (Hardouin and Tautz, 2013; Stewart et al., 2008) and provided in supplementary Table 3. The sequences were edited and visualized using CodonCode Aligner version 4.1.1 (CodonCode Corp.). Three mitochondrial genome sequences were additionally obtained from the whole genome sequenced data (details in section 4.2.1). A total of 12 mtDNA genome sequences were aligned using MEGA version 5.0 (Tamura et al., 2011). I determined the number of mutations in these sequences in comparison to the consensus sequence and I estimated the mutation frequencies from the total number of nucleotides sequenced using the procedure applied by (Hardouin and Tautz, 2013; Stewart et al., 2008). I also used the total number of mutations and nucleotides sequenced to estimate the age of these mice on Heligoland since their probable first colonization event.

2.3 Results

2.3.1 Subspecies-diagnostic nuclear markers

The four nuclear markers were typed for all the samples from Heligoland and control samples from the other three subspecies *M. m. musculus*, *M. m. domesticus*, and *M. m. casteneus* were also included. The amplified markers were checked for their fragment sizes by gel electrophoresis and used to mark the genetic background of Heligoland mice at these markers.

Table 2.1: Expected product sizes for the typed nuclear markers.

Nuclear marker	**Expected product sizes (bp) and specific alleles**		
Gene/marker	***Domesticus***	***Musculus***	***Casteneus***
Abpa	192	-	-
Abpb	-	290	290
D11 cenB2	Allele specific to *domesticus*	Allele specific to *musculus*	Allele specific to *domesticus*
Btk	342	206	206
Zfy2	202	184, 202	184, 202

The amplified fragments of *Abp* and *D11 cenB2* were scored for their allele specific PCR fragments. *Abp* marker fragment size for all the mice samples from Heligoland showed the same size of fragment as *M. m. domesticus* for *Abpa* but not for *Abpb*. The *D11 cenB2* showed the fragment size of alleles specific to *M. m. domesticus*. On the other hand, the amplified fragments of *Btk* and *Zfy2* were scored for the presence or absence of an insertion and deletion at each marker respectively. *Btk* marker for all the mice samples from Heligoland has the fragment size of 342 bp which means that these mice have the insertion at the *Bruton agammaglobulinemia tyrosine kinase* gene. The presence of the insertion is indicative of *M. m. domesticus*. The typed *Zfy2* marker for all the samples from Heligoland exhibited the presence

of only one fragment of size 202 bp, hence the absence of the 18 bp deletion in the Zinc finger protien (*Zfy2*) which is also proofed as a diagnostic marker for *M. m. domesticus*. The results of the four nuclear markers are detailed in Table 2.1.

2.3.2 Population genetic diversity

A total of 17 mice from the island of Heligoland were genotyped using 21 microsatellite markers in addition to 4 populations from France, 3 populations from Germany and one population from Kazakhstan. These populations represent the two species *M. m. musculus* and *M. m. domesticus*. The genotyped data for all individuals where population genetic parameters such as average number of alleles per locus, observed and expected heterozygosity values were calculated for each population (Table 2.2).

Table 2.2: Population genetic parameters for microsatellite loci typed in this study.

location	Population	N	H_{obs}	H_{exp}	A_{av}
Germany	Cologne-Bonn	45	0.533	0.799	11.19
	Ploen-District	18	0.381	0.770	7.86
	Schoemberg	12	0.435	0.697	6.52
Heligoland	Heligoland	17	0.328	0.479	3.33
France	Massif Central	46	0.597	0.765	11
	Louan	12	0.549	0.732	5.81
	Divonne les Bains	12	0.591	0.786	7.76
	Nancy	12	0.603	0.798	7.24
Kazakhstan		47	0.614	0.759	13.24

N, number of individuals scored; A_{av}, mean number of alleles per locus; H_{ob}, observed heterozygousity; H_{exp}, expected heterozygousity.

The population from Heligoland shows reduced heterozygosity (0.479) and lower average number of alleles (3.33) when compared to other populations from the mainland Cologne-Bonn (0.799/11.19) and Massif Central (0.765/11). Such a reduction in genetic diversity reflects the possibility of different scenarios e.g. population inbreeding and bottleneck. The influence of local inbreeding on the island of Heligoland is in

line with the fact that the sampling scheme in this study didn't follow the standard sampling protocol (Ihle et al., 2006) and that the small size of the island has the potential to promote inbreeding. It has been known that natural populations of the house mouse exhibit local inbreeding and communal nesting (Berry and Bronson, 1992), which can result in local reduction of genetic diversity (Ihle et al., 2006).

The constructed tree of individuals using the calculated distances of the proportion of shared alleles is shown in Figure 2.2. The clustering of *M. m. domesticus* populations from *M. m. musculus* is clearly observed. The population from Heligoland (green color) forms a single cluster with only one case of possible recent migration. Although this measurement does not account for microsatellite mutations, it has been shown to be informative for phylogenetic reconstruction (Harr et al., 1998).

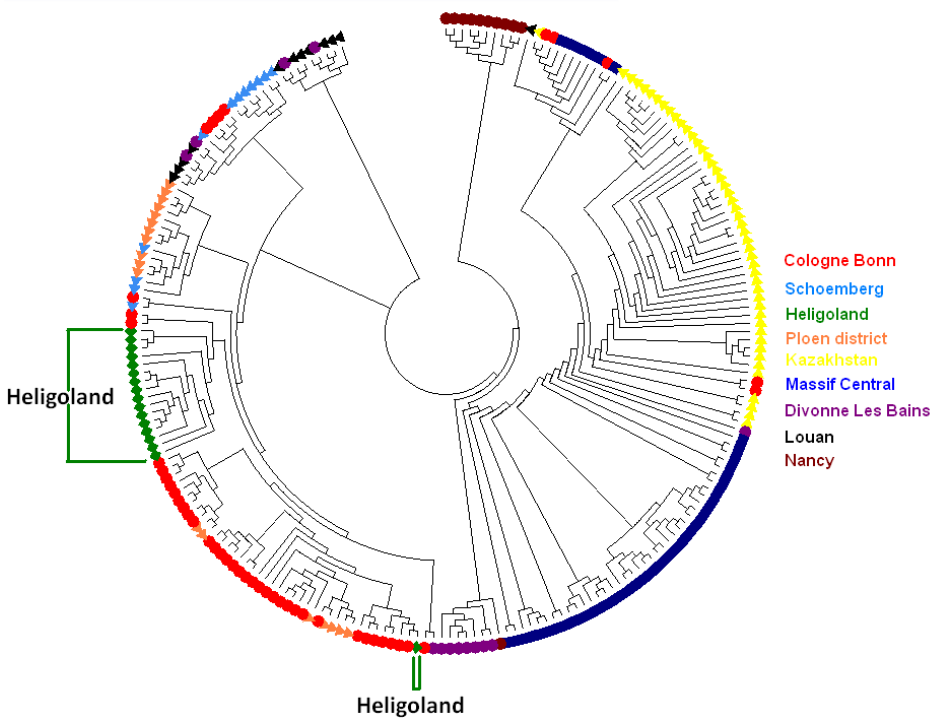


Figure 2.2: Neighbour-joining tree based on the calculated proportion of shared alleles for individuals from Heligoland and mainland.

The neighbor joining tree based on Cavalli-Sforza and Edwards (1967) chord

distance D_c, is shown in Figure 2.3. The tree reveals the basic population differentiation between *M. m. musculus* represented here by the population from Kazakhstan and *M. m. domesticus* by populations from France and Germany. Consistent with the findings of (Ihle et al., 2006).

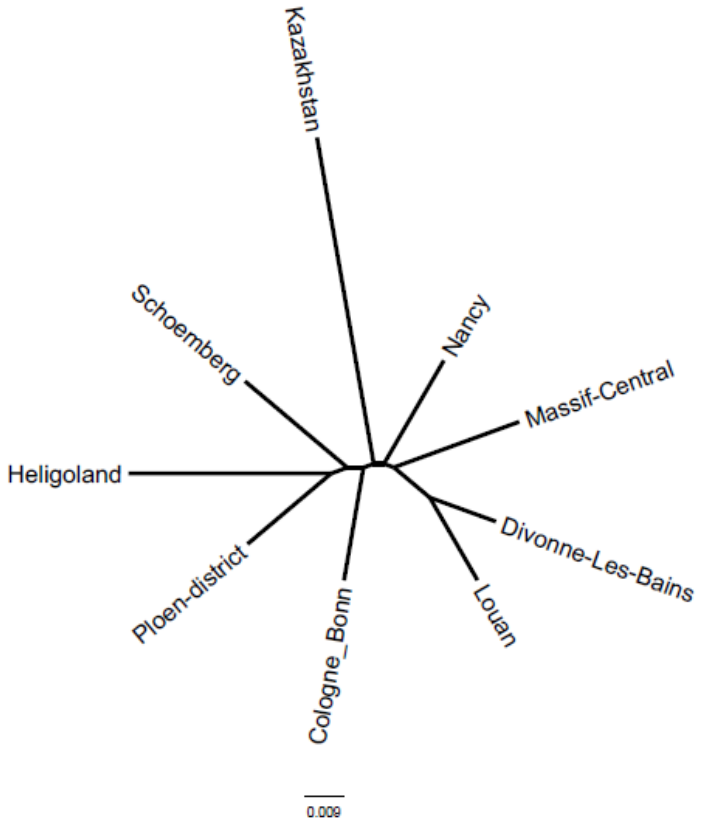


Figure 2.3: Neighbor joining tree based on Cavalli-Sforza and Edwards (1967) chord distance

The assignment of individuals to genetic populations was assessed using STRUCTURE and the output was generated using the programs CLUMPP version 1.1.2 (Jakobsson and Rosenberg, 2007) and Distruct (Rosenberg, 2004) respectively. The estimation of the realistic K value was analysed according to (Evanno et al., 2005). The basic *M. m. musculus* and *M. m. domesticus* structure can be observed from Figure 2.4 where the population from Kazakhstan was assigned to one cluster and all other populations were assigned to the other cluster at $K = 2$. Even though

the optimum value of K according to the method is $K = 2$, it is noteworthy to interpret the structure at $K \geq 4$ were individual clustering show more pronounced patterns and consistent assignments. Accordingly, the structure output for $K = 2$ - 7, each for 10 replicates is illustrated in Figure 2.4. Hence, the pattern of clustering indicates that the populations anlaysed are consistently assigned to their clusters at larger values rather than at lower values of K.

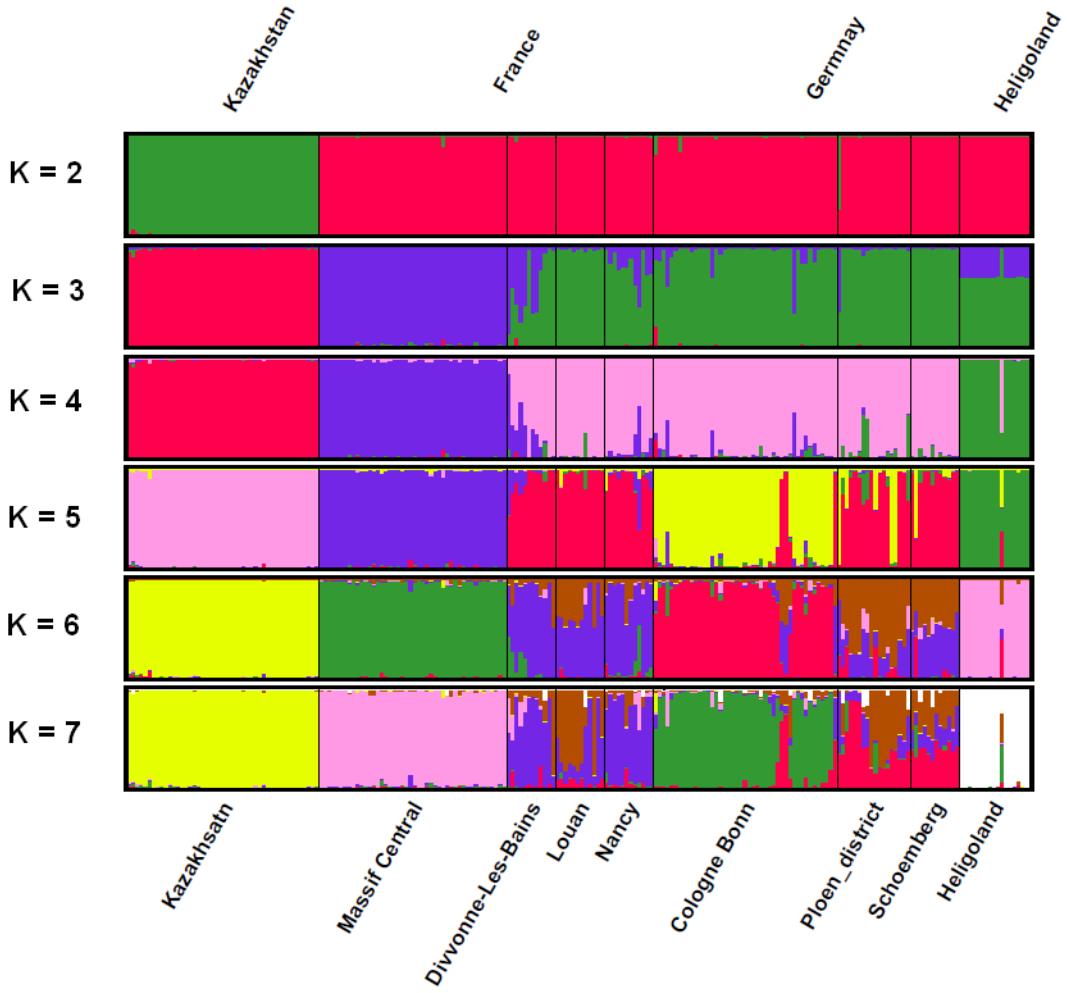


Figure 2.4: Clustering of 221 individuals from 9 *M. musculus* populations, assuming 2-7 clusters K. The optimal number of clusters is two and the mean (across replicate runs) log likelihood for $K = 2$ was -18374.19. Each individual is represented by a column divided into K colors with each color representing a cluster. Different populations are separated by a black line and are labeled below the figure by sample locations and above the figure by geographic region.

These observed clusters are consistent with the subspecies level of differentiation and the geographical distribution of the populations. Moreover, a closer resolution on the population structure showed that the population from Heligoland was assigned to a single cluster at $K >= 4$ in contrast to the results at lower K values. The structure analysis results supports the pattern seen on the allele sharing tree, which confirms the notion of the distinctness of Heligoland population from the observed microsatellite data except for one individual from Heligoland which is more likely a recent immigrant to the island.

2.3.3 mtDNA analysis

2.3.3.1 mtDNA D-loop

I obtained 852 bp of the mtDNA control region (D-loop) from all the 17 mice samples and I found that there are only two haplotypes belonging to Heligoland, grouped within *M. m. domesticus* haplotypes (Figure 2.5). One major haplotype represented by a total of 16 sequences with a single mutation on position 15995, this mutation divides the major haplotype into two sub-haplotypes each of 8 sequences which most likely arose on Heligoland (Table 2.3). I found an insertion (TAACTCTCTTT) of 11 bp in the D-loop sequence between positions 16091-16101 in the 16 mice representing the major haplotype, but not in the sequence of the minor haplotype. This insertion was previously found in one sequence from Heligoland and was determined as a distinct haplotype known as haplotype DEU_U47469.41 Holstein according to (Prager et al., 1993). The sequence representing the DEU_U47469.41 Holstein haplotype from Heligoland (light green) in Figure 2.5 groups within Heligoland sequences (green color) from this study. Moreover, I found a deletion of 2 bp (TA) at position 15 559-15 560. Hence, the haplotype grouped within the 11 bp clade found previously in 49 mice from Holstein in Germany, in 82 mice from Swedish, Finish,

and northern Danish populations. In addition to 13 mice from two localities within the hybrid zone (Prager et al., 1993). The other haplotype is a minor haplotype only represented by one sequence and grouped within haplotypes from Germany, British Isles and Great Britain. This haplotype harbours many point mutations from the haplotype of Heligoland and possibly belongs to a recent immigrant (HG09).

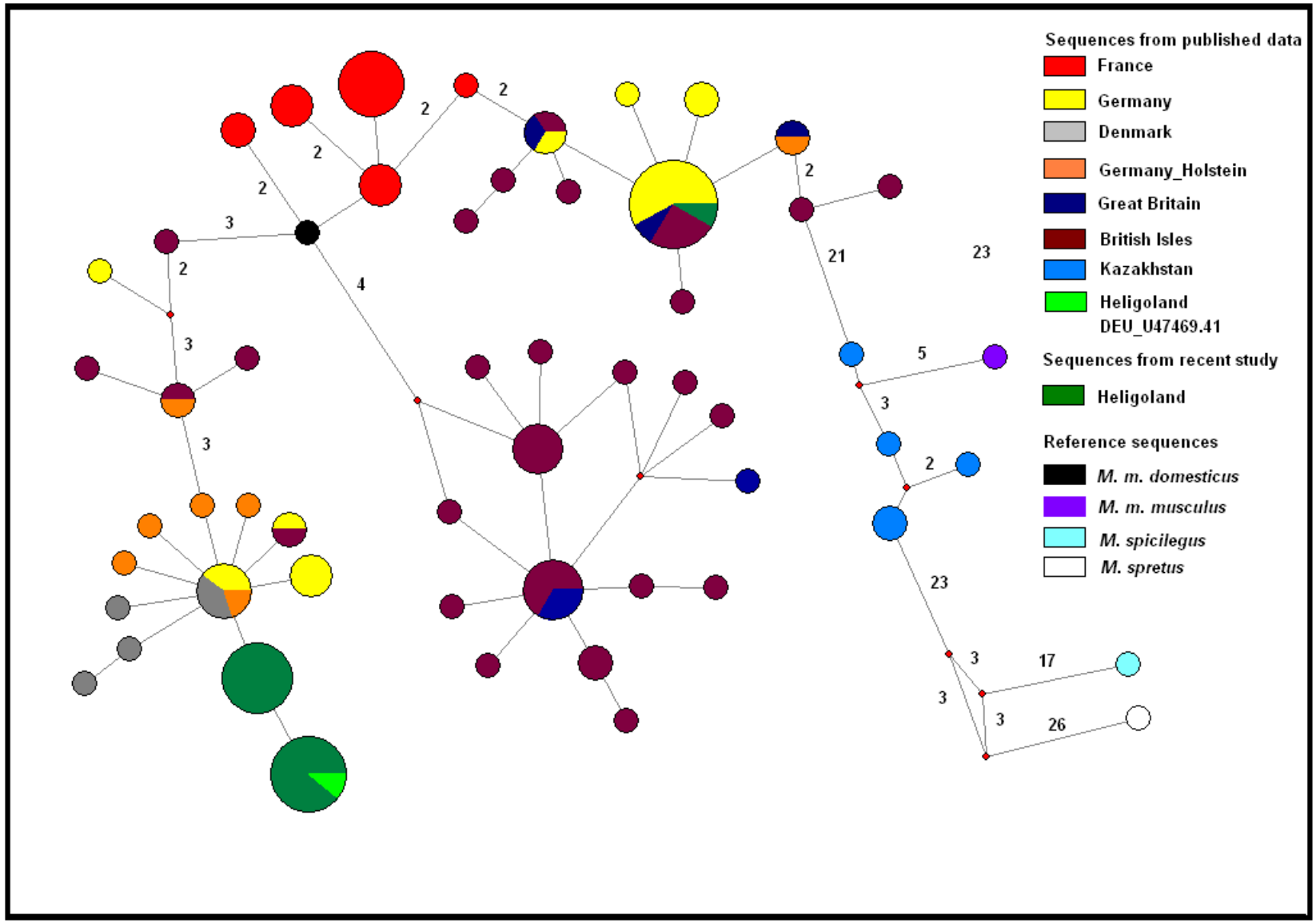


Figure 2.5: Network based on mtDNA D-loop haplotypes calculated using Median Joining for the house mouse from Heligoland and previously published sequences along with *M. m. domesticus*, *M. m. musculus* and two *Mus* species. The area of each circle is proportional to haplotype frequency. Each node is one mutational step away from the next one, numbers indicate the cases of more than one step. Small red circles indicate branch splits.

2.3.3.2 mtDNA genome

From the sequences of 12 mitochondrial DNA genomes I excluded the sequence of the recent immigrant. In total I found 11 point mutations from a total of 11 sequences, seven substitutions are in coding regions resulting in amino acid changes. One in $tRNA^{val}$, ND2, $tRNA^{Asn}$, the replication origin, two in ND4, one in ND5, one in ND6, two in CYTB and one in the D-loop region (Table 2.3). I used the number of mutations assigned in the mtDNA excluding the D-loop (10 substitutions) to calculate the mutation frequency per nucleotide sequenced. For this I applied the sequence divergence measurement which is given by the following formula (frequency = No. mutations / No. nucleotides sequenced), given that the mtDNA sequence wit the D-loop exempted is 15448 nt and the number of point mutations is 10. Applying the formula to the data I find: 10 mutations among 154 480 nucleotides sequenced (10 x 15448 nt), results in a mutation frequency = 6.4 x 10^{-5}. Without a documented date of colonization or onset of house mouse invasion to Heligoland, I assume here that these mice are comparable to the house mice from Kerguelen which colonized the archipelagos in a single primary wave dated back to 200 years ago (Hardouin and Tautz, 2013). I applied the mutation frequency calculated for Kerguelen mice 3 x 10^{-5} from a total of 7 mutations in 16 mice sequenced. Dividing the mutation frequency for Heligoland to that of Kerguelen (6.4 x 10^{-5}/3 x 10^{-5}), gives a ratio of 2.13. Assuming the same factors on both islands I estimate the first colonization wave introduced to Heligoland (2.13 x 200 years) = 400 years ago.

Table 2.3: Observed nucleotide substitutions in the complete mitochondrial DNA sequences of Helgoland mice, the data was analysed based on the UCSC genome browser annotations (mm9).

position	$tRNA^{val}$ 1080	ND2 4771	$tRNA^{Asn}$ 5157	rep_origin 5163	ND4 10688	ND4 10689	ND5 12009	ND6 13681	CYTB 14698	CYTB 15163	CR 15595
consensus sequence	A	C	C	C	T	A	T	C	T	G	T
HG_01		T									
HG_02								A			
HG_05							C				
HG_06										A	C
HG_08	G										
HG_10											C
HG_11										A/G	C
HG_12									C		
HG_13									C/T		
HG_14					C	T					C
HG_1450_2			T	T							C
outgroups											
Rat	A	C	C	C	C	A	C	A	T	G	T
Human	A	C	C	C	A	A	A	C	T	G	T
Orangutan	A	C	C	C	T	A	C	C	T	G	T
Dog	T	T	C	C	A	A	G	G	T	G	T
Horse	A	A	C	C	A	A	G	A	T	G	C
Opossum	T	A	C	C	A	C	A	C	T	G	T

2.4 Discussion

Revisiting classification

Based on the evidence from the analysis of the subspecies diagnostic molecular markers, microsatellite markers, and mtDNA the mice from Heligoland clearly resemble the characteristics of the *M. m. domesticus* subspecies. My proposed scenario for the origin of Heligoland mice is in accordance with the fact that the mice that invaded the island were from the Western European region, in particular, from Denmark and northern Germany. These are predominated by *M. m. domesticus* subspecies. In addition, this is also line in with the given scenario of Heligoland being isolated and is situated on the *M. m. domesticus* side of the hybrid zone.

Genetic diversity

The genetic diversity in the context of microsatellite variation reflected by mean number of alleles and heterozygousity measures is low (Table 3.2). The low genetic diversity reflects the influence of local inbreeding of this population and the assumption that these mice experienced a colonization bottle neck (Berry and Bronson, 1992). It is noteworthy to point out that during the history of the island additional bottlenecks could have occurred during wars as a result of the burning of houses and the destruction of infrastructure.

Genetic drift as an evolutionary force that has strong influence on small populations, could have played another role in reducing genetic diversity of these mice. Additionally, taking into account the low diversity from microsatellites and that from the mtDNA haplotype data analyzed here, it is clear that the colonization of Heligoland by a single matrilineal haplotype is correlated with the low level of polymorphism since colonization. Despite the low genetic diversity, the genetic structure

of Heligoland mice is homogeneous with a distinctive structure as revealed at larger K values rather than the optimum value at K=2. This could be related to the behaviour of the model implemented in Structure and the importance of including large number of markers or a number of ancestry informative markers. The population structure of the house mice from Heligoland shows mostly no admixture of genotypes from the geographically close populations from Germany and France for the addressed markers except for one individual, but this is most likely a recent immigrant. This result was also reflected by means of the proportion of shared alleles and genetic distances between populations which are all in line with mtDNA haplotype diversity on Heligoland.

Colonization history of Heligoland house mice

Heligoland mice show a low haplotype diversity $H_D = 0.588$ and nucleotide diversity $\pi = 0.0020$ compared to house mice from other archipelagos such as the Madeiran were $H_D = 0.90$ and $\pi = 0.0014$ (Gündüz et al., 2001), New Zealand were $H_D = 0.66$ and $\pi = 0.0042$ for *M. m. domesticus* (Searle et al., 2009a), and $H_D = 0.955$ and $\pi = 0.0082$ for Great Britain and Ireland (Searle et al., 2009b). Continental *M. m. domesticus* populations tend to have higher values for both haplotype and nucleotide diversity, $H_D = 0.896$ $\pi = 0.0082$ for Norway (Jones et al., 2010) and $H_D = 0.98$ $\pi = 0.0084$ for Turkey (Gündüz et al., 2005). The low haplotype diversity on Heligoland is the result of only one major haplotype. Some other studies found a link between house mouse low haplotype diversity and human population (Förster et al., 2009; Jones et al., 2011). While Heligoland is inhabited by $>$ 1000 people (Dierschke et al., 2010), the lower haplotype and nucleotide diversity can not be directly linked to human population. However, a close look at human population genetic diversity on Heligoland might support what I have suggested here, a follow study could conduct such a comparative analysis. It is worth noting that the island

is 46 km away from the closest mainland coast and maritime activities are high which increase probability of migrating and immigrating mice. On a similar perspective, the analysis of the mtDNA D-loop does not present any disparate haplotype diveristy on Heligoland as shown by the calculated haplotype diversity, the presence of a single major haplotype and a minor haplotype, represented by one mouse. These results lead to a major finding of a single primary colonization on Heligoland and no further genetic immigration. This finding is compatible to that from Kerguelen archipelago and sub-Antarctic islands; the study suggested that when a population is established and settled on an island, further introductions are not powerful enough to interfere with the genetic composition of the resident population (Hardouin et al., 2010).

It is obvious that the mice on Heligoland succeeded in establishing colonies when conditions were good enough. The mice population on Heligoland shows a level of local adaptation and a long term commensalism with humans despite the recent history of the island during the World Wars. Despite evacuation and the consequent bombing followed by recolonization the island genetic integrity was conserved.

Most interestingly, a single individual was found that does not fit into the general genetic pattern and that I interpret as a recent immigrant that must have come from one of many ships that continuously land in Heligoland. It testifies that new mice arrive on the island, but they failed to get established, i.e. neither form new colonies, nor are they genetically integrated to an appreciable degree, at least with respect to the mitochondrial haplotype. This suggests that the resident population is so highly adapted to the local conditions that new invaders have little chance for survival.

In summary, it appears more likely that the population of house mouse on Heligoland is a well established population and that it remained isolated with a relatively small population size, resulting in lowered genetic diversity, as measured by microsatellites and mtDNA data. Moreover, it is obvious that the first mice to arrive on Heligoland were possibly from Denmark or Northern Germany South and West

of the *M. m. musculus*/*M. m. domesticus* hybrid zone respectively. That is to say, *M. m. domesticus* bearing the DEU_U47469.41 Holstein mtDNA haplotype were the ancestors of Heligoland house mouse.

3 Aspects of Insular evolution and adaptation in the mandible of *M. m. helgolandicus*

3.1 Introduction

Extensive studies on different mammals showed that mammals isolated on islands often exhibited significant body size changes and in some cases morphological evolutionary changes for particular body parts (Foster, 1964; Lomolino, 1985, 2005; Sondaar, 1991; VanValen, 1973). Foster (1964) was the first to describe what was later known as the Island Rule (Lomolino, 1985, 2005; VanValen, 1973). The rule was supported by empirical data, mainly based on island-mainland comparisons (Foster, 1964; Lomolino, 1985, 2005). It stems from observations that small mammals such as rodents tend to be bigger in size than their counterparts from the mainland. Foster (1964) and VanValen (1973) generalized the island rule that dwarfism is observed in large species and gigantism is observed in small species when colonizing an island.

In addition, Heany (1978) studied patterns of evolution on islands, he suggested that the major force driving evolution there is the relationship between island area and body size. Where the limitation of resources turns to be a factor leading to body size decrease in large mammals and the lower rates of predation and competition to be factors leading to body size increases in small mammals. On a similar perspective, Lomolino (1985) showed that competition and immigration selection lead to gigantism, whereas the availability of food resources and predation lead to dwarfism.

Both Heany (1978) and Lomolino (2005) suggested that the effects of these factors are more apparent on smaller islands. This assumption was supported by a

recent study where evolutionary bursts on small islands were evident. In addition, the absolute amount of change in island mammals was negatively related to the island area; smaller islands show greater amount of change (Millien, 2011). This links the fast evolution of mammals to the accelerated rates of adaptation on small islands as a result of new environmental responses, in addition to the large ecological contrast between the island and mainland. This includes species diversity, isolation, abiotic factors, and demographic factors such as smaller population size (Losos and Ricklefs, 2009; Millien, 2011).

Although the island rule has been supported with empirical data, few studies showed cases where the rule was violated. For example, a study of *Apodemus* on the small islands of the Japanese archipelago. Only the large species of *A. speciosus* showed a trend towards larger size among the small island populations. In contrast to the smallest species *A. argenteus* which was affected by the environmental gradient. The observed morphological differentiation of these island populations was attributed to the genetic background which is interpreted as the combined effect of the genetic basis of the founding population and the subsequent genetic drift (Renaud and Millien, 2001).

Understanding the aspects of evolution in the context of morphological changes is much more challenging when considering all factors influencing evolution on islands, the strength of isolation and different selection pressures. It is generally assumed that the genetic background and the body size of the animal are the factors that lead to morphological evolution, however some extensive studies found that the size of the island and the degree of its isolation from the mainland are strongly correlated with factors leading the morphological evolution (Berry, 1996; Heany, 1978; Lomolino, 1985).

Phenotypic evolution in house mice

Phenotypes are complex structures that are determined by the combined effects of several genes. Hence, the study of phenotypes is important for assessing evolutionary patterns of biological shape and it has been of considerable relevance to documenting historical patterns and different aspects of phylogeography (Renaud et al., 2007).

The mandible is a morphological character mainly involved in mastication and food processing, it has gained attention in the morphometric field because it can unravel patterns of adaptation processes associated with possible ecological shift (Renaud et al., 2009). The mandible of the house mouse is divided into two anatomical regions defined as the alveolar and ramus regions. The alveolar region is divided into the incisor and molar zone, while the ramus is divided into three regions known as coronoid, condyle, and angular processes, see Figure 3.2 & Table 3.1 (Boell et al., 2013). The evolution of such a character might have been influenced by different factors. The mandible is well known to be under selective pressure for food processing mechanism and is also controlled by bone plasticity which is influenced by diet (Renaud and Auffray, 2010; Renaud et al., 2010).

The mouse mandible is of major interest for studying evolution, in particular with the recent advances in geometric morphomotrics, such as the landmarking approach and related quantitative statistical analysis (Klingenberg, 2010). The evolutionary history of the house mouse and its recent patterns of colonization have been connected to humans through commensalism (Sage et al., 1993). The commensal life style has been proposed to have evolved independently in the different subspecies of the house mouse, and can be explored by morphological analysis.

Functional characters related to commensalism such as lower aggressive behavior (Corti and Rohlf, 2001) and diet shift (Renaud and Auffray, 2010) have proven to affect the mandible shape changes. Recent studies have shown that *M. musculus*

species can be distinguished on the basis of their mandibular morphology and that there is a considerable variation among populations within subspecies, which is much higher for both shape and size in peripheral regions than in central regions where the house mouse is distributed (Siahsarvie et al., 2012).

Evolutionary studies of house mouse and other rodents have illuminated many examples of morphological evolution that resulted from different mechanisms. These include non-genetic or environmental factors, gene flow from morphologically different source populations, genetic drift and responses to natural selection. The meta-analysis on morphological data over the last century in four widely-separated island rodent populations concluded that the observed phenotypic changes are best explained by natural selection, and that the rates of evolution are higher on smaller and remote islands. Furthermore, the analysis also addressed the unlikeliness of gene flow as a source of explanation referring to the fact that the studied islands are remote and that if new introductions took place, they needed to bring the change. In addition, the analysis showed that genetic variation within a population is a prerequisite for adaptive responses and multiple colonization waves or rapid expansion following invasion may serve to maintain eliminated genetic variation caused by bottlenecks and founder effects sufficient enough to allow island population to respond to selection. Moreover, the rapid adaptive response more readily takes place on isolated populations (Pergams and Ashley, 2001).

The (Renaud et al., 2010) study on epigenetic effects on mouse mandible found that there are two major sources of plastic shape variation that can affect mandible morphology: muscular dystrophy and the efficiency of food processing. Moreover, that these factors do not alter the bone of the mandible directly during the mouse development, but rather they both modify the muscular attachment force on the mandible during late postnatal growth. When the mice were exposed to food of different consistencies after weaning that resulted in a shape change of the mandible.

Even though the studied factors were related to bone remodeling of the mandible, their morphological characteristics observed were different, muscular dystrophy was observed to cause a shape change distributed all over the mandible whereas the response to food consistency was more localized around the molar zone and the insertion of the masseter muscles. Those findings concluded that differences related to food processing caused more targeted changes related to a given function of the mandible reflected by the type of food (Renaud et al., 2010).

In contrast, another recent study by Boell et al. (2013) suggested pleiotropic effects where multiple parts of the mandible were affected at the same time. The major finding of that study was that the gene dosage has lower effects on the mandible shape changes, but more pronounced than the average additive effects revealed by quantitative trait loci (QTL). Moreover, deciphering the gene dosage effects on the mandible morphology will expand our understanding of the different aspects of morphological changes and more into details about the evolutionary mechanism of such a complex trait (Boell et al., 2013).

Landmark geometric morphometrics

Morphometrics is mainly concerned with the quantitative measurement of the biological shape. The landmark-based geometric morphometrics summarizes the morphology in terms of landmark configurations in 2 or 3-dimensional Cartesian coordinates (Webster and Sheets, 2001).

The application of landmark geometric morphometrics has proven to be powerful and has been widely used with the potential to provide insights into how a given shape differs. Additionally, it is gaining more attention with the increasing ease of digitally acquiring landmark data and the advancement and availability of applicable software (Webster and Sheets, 2001).

The advancement in the morphometric field helped dissecting the change in

mandible morphology from wild house mouse populations as well as inbred strains of the genus *Mus*. Recent morphometric approaches have become applicable in a wide range of studies substituting for previous established methods which were based on cranial or dental structure (Auffray et al., 1996). In addition, the application of geometric morphometrics and different related statistics has allowed the establishment of new approaches useful for the differentiation among the subspecies (Boell and Tautz, 2011; Siahsarvie et al., 2012).

Mus musculus helgolandicus

Mus musculus helgolandicus was first studied by Zimmermann (1953). The morphological analysis conducted by Zimmermann (1953) on these specimens from Heligoland and the two populations representing the two well known subspecies *M. m. domesticus* and *M. m. musculus* inhabiting western Europe showed that the mice from Heligoland are morphologically different and hence were defined *M. m. helgolandicus* according to that study. The whole analysis was based on body measurements and skull analysis. The major analysis of the skull focused on the condylobasal length (CBL) which is the length of the skull, measured from the anterior points of the premaxilla to the posterior surfaces of occipital condyles.

Later in (1968) Reichstein and Vauk showed that the house mouse of Heligoland was not only characterized by the combination of features of the two closest mainland forms *M. m. musculus*, and *M. m. domesticus*, but they exhibited differences in other forms. They surveyed morphological variation among *M. m. helgolandicus*, *M. m. musculus*, and *M. m. domesticus* focusing on comparative analysis between condylobasal length and zygomatic arch width (Figure 3.1). Their findings supported that of Zimmermann (1953) and led to the confirmation that these house mice are a separate subspecies, at least on a local basis.

I investigate, here; the mandible morphology of *M. m. helgolandicus* for dif-

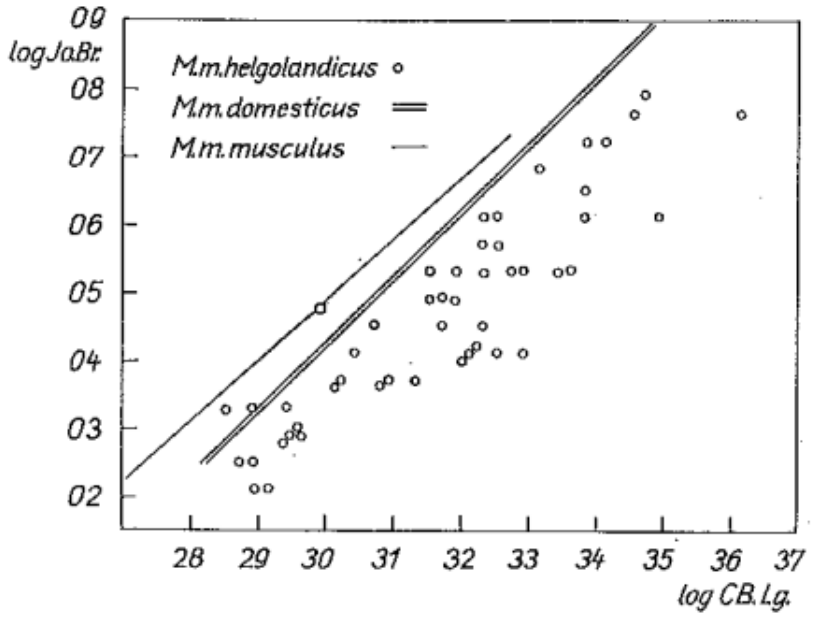


Figure 3.1: Correlation diagram of the log of condylobasal length against the log of zygomatic arch width between *M. m. musculus*, *M. m. domesticus* and *M. m. helgolandicus*, Figure adapted from (Reichstein and Vauk, 1968)

ferent specimens representing a contemporary collection and two older collections dating back to 1930s by Zimmermann (1953) and 1960s-1970s (private collection). I am interested in the mandible shape variation of these mice as compared to other mandibles found among wild mice from the mainland.

This study might improve our understanding of the previously assigned differences in these mice. Additionally, it might confirm the findings from the molecular analysis or potentially provide a new perspective of the evolution of morphological aspects. In both cases this will improve our understanding and outline the demographic history of these mice on Heligoland.

The selected landmarks for mandible analysis were gleaned from a previous study that included both wild and inbred mice specimens, which all belong to the genus *Mus*. Hence the landmarks are assumed to be homologous and applicable for the mandible statistical comparisons considered here (Boell and Tautz, 2011).

3.2 Materials and methods

3.2.1 Tail measurements and coat coloration

All mice of the contemporary collection were scored for dorsal and ventral coat color using Turner's standard color chart (chart used locally at our mouse facility). For each specimen (excluding juveniles and those with damaged body parts resulting from snap trapping), head-body length and whole body length were measured and tail length was determined by subtracting head-body length from total body length. Then these measurements were used to determine the relative tail length which is expressed as tail:body length ratio (TBLR). It is known as a reliable discriminator between *M. musculus* subspecies; *M. m. musculus* tends to have a smaller (TBLR) than the other two subspecies (Boursot et al., 1993; Kraft, 1985; Marshall and Sage, 1981; Searle et al., 2009a). Here, the (TBLR) was calculated for 6 specimens from Heligoland and 6 specimens from each of the following populations, Cologne Bonn, Massif Central and Kazakhstan representing the two subspecies *M. m. domesticus* and *M. m. musculus*.

3.2.2 Geometric morphometrics

3.2.2.1 Animal specimens

I analyzed a total of 65 skull specimens for the house mice from the island of Heligoland collected at different time periods (details in section 2.2.1). The oldest was collected by Zimmermann (1953) early in the thirties and was obtained from the Zoological Museum in Berlin (ZMB) as a loan. The second was collected by amateur collectors during the fifties and seventies and was obtained through a loan from the Institute für Haustierkunde (IFH) in Kiel. The contemporary collection was collected during our trip in 2012 and by the researchers at the Institute for Avian

research in Heligoland during 2004 to 2012.

3.2.2.2 Specimens preparation

All mouse specimens were subjected to preparation prior to the scanning process following the same protocol. The skulls of trapped mice from this study were prepared from whole body specimen (preserved in Ethanol) by first decapitating the head by a process that the whole skull with the mandible attached were complete. The old material borrowed from the museum and the Institute für Haustierkunde (IFH) in Kiel were prepared taking care that the mandible is intact and attached to the skull, for these specimens I used the provided information for sex and labeling from the containers of the borrowed material. In some cases mandibles were only available without skulls or only one intact hemimandible. The 65 skull specimens were scanned with a micro-computertomograph (microCT- VivaCT 40, Scanco, Bruettisellen, Switzerland). The left hemimandible of each of the specimens scanned was outlined using the software options provided by the microCT. Details of the specimens are supplied in supplementary Tables 4, 5 & 6.

3.2.2.3 Mandible landmarking

Two dimensional coordinates of 14 mandibular landmarks were digitized on each hemimandible of the scanned and outlined specimens. In addition, incomplete mandibles due to damage resulted from snap trapping of mice or the impact of museum storage processes, were digitized by either using the intact hemimandible (left/right) or the best available landmarks.

The digitization was performed in two independent rounds to avoid technical errors and to get hands on digitization steps. The digitization was performed using two software utilities from Morphometrics tpsUtil (Rohlf, 2004) and tpsDig (Rohlf, 2005) respectively. The positions of the landmarks analyzed in this study are illus-

trated in Figure 3.2 and their definitions are detailed in Table 3.1.

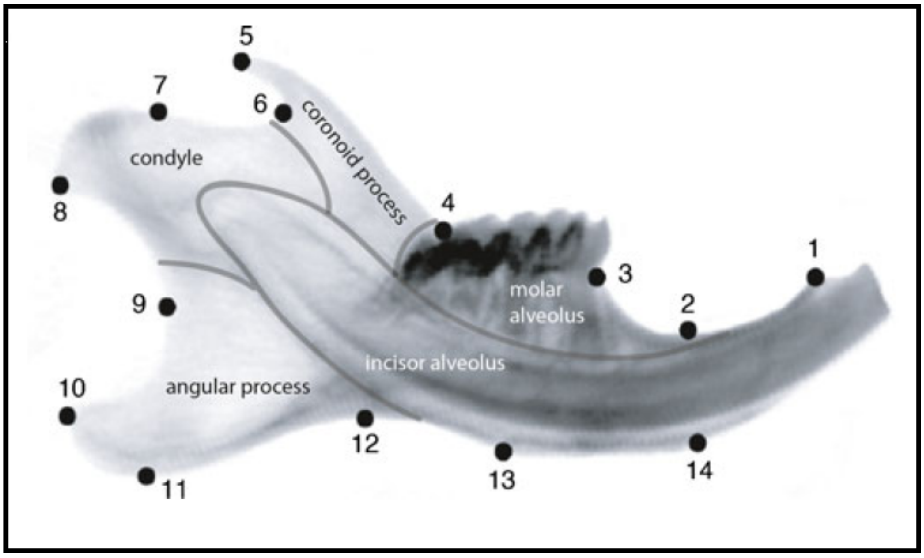


Figure 3.2: Lateral view of a mouse hemimandible showing the 2 dimensional 14 landmarks used in this study (figure adapted from (Boell et al., 2013)).

Table 3.1: Definitions of landmarks used in geometric morphometric analyses (Adapted from (Boell et al., 2013; Boell and Tautz, 2011)).

LM No.	Landmark
1	Anterior terminus of bone dorsal of the incisor
2	Minimum of depression on dorsal side of incisor ramus
3	Bone/teeth transition anterior of M1
4	Intersection of ascending ramus with tooth row
5	Tip of processus coronoideu
6	Minimum of depression posterior to processus coronoideus
7	Anterior margin of condylar articular surface
8	Posteroventral tip of condyle
9	Minimum of depression formed by condyle and processus angularis
10	Posterodorsal tip of processus angularis
11	Posteroventral tip of processus angularis
12	Minimum of depression formed by processus angularis and incisor ramus
13	Posterior margin of muscle insertion area on ventral side of incisor ramus
14	Anterior margin of muscle insertion area on ventral site of incisor ramus

*Landmarks numbers as in Figure 2.2.

To avoid any technical errors that could result from the measurements (observer factor) I selected randomly 14-16 specimens (hemimandible) from each of the populations studied by Boell and Tautz (2011) and digitized them all for a combined

analysis with Heligoland collections. This can provide insights on possible shape variations among *M. m. helgolandicus* and various *Mus* species from the mainland.

I included six different populations published in Boell and Tautz (2011). The subspecies *M. m. domesticus* is represented by three different populations from Germany (Frankfurt), Iran (Teheran) and Kerguelen (Gouillou). The subspecies *M. m. musculus* is represented by a population from Hungary. A population from Johnston Atoll in Taiwan was included to represent *M. m. castaneus*. And a population from Madrid was also included as a representative for the species *M. spretus* (Table 3.2). To avoid distortion of statistical analysis few samples of each data set were excluded, either for the suspected young age or for the suspected old age as well as mandibles with malformation diagnosis.

Table 3.2: Populations used for geometric morphometrics in this study.

Region	Population/Locality	Species	Source/Reference	N
Germany	Frankfurt	*M. m. domesticus*	SMF/(Boell and Tautz, 2011)	15
Iran	Teheran	*M. m. domesticus*	SMF/(Boell and Tautz, 2011)	16
Kerguelen	Gouillou island	*M. m. domesticus*	J-L.C/(Boell and Tautz, 2011)	15
Hungary	Hungary	*M. m. domesticus*	SMF/(Boell and Tautz, 2011)	15
Taiwan	Johnston Atoll	*M. m. domesticus*	NMNH/(Boell and Tautz, 2011)	14
Spain	Madrid	*M. spretus*	R.R/(Boell and Tautz, 2011)	15
Germany	Heligoland	*M. m. helgolandicus*	ZMB/(Zimmermann, 1953)	23
Germany	Heligoland	*M. m. helgolandicus*	IFH	27
Germany	Heligoland	*M. m. helgolandicus*	MPI	17

SMF=Senckenberg Museum Frankfurt; J-L.C= Jean-Louis Chapuis; NMNH= National Museum of Natural History, Smithsonian Insitution, Washington; R.R= Ruth Rottscheidt; ZMB= Zoological Museum Berlin; IFH= Institute Fuer Haustierkunde, Kiel; MPI=Max-Planck Institute for Evolutionary Biology, Ploen.

3.2.2.4 Geometric morphometrics analysis

The landmark coordinates for the different data subsets were processed with the Procrustes fit implemented in MorphoJ (Klingenberg, 2011). MorphoJ implements a full Procrustes superimposition method which is not very much different from

other Procrustes fitting (partial Procrustes fits), but what is important is that when analyzing data sets with unusually large variation, the full Procrustes fit will put less weight on observations that are far from the average shape and will therefore be more effective against the influence of outliers (Klingenberg, 2011).

The procrustes superimposition in MorphoJ is performed to produce new variables for the analyzed mandible shapes which corresponds to the raw coordinates. The superimposition translates the configurations of the raw coordinates to a point where only the shape between landmarks is the major differentiating factor (Klingenberg and McIntyre, 1998). The landmark coordinates derived from application of Procrustes fit in MorphoJ were then used to generate one covariance matrix for the dataset from Heligoland and another for the whole data set. The Procrustes distances calculated among the specimens from Heligoland and specimens from different continental populations were used to obtain and draw a neighbor joining tree using phylip version 3.69.

3.2.2.5 Centroid size

The size of the mandible for each specimen was estimated from its calculated centroid size in MorphoJ. The centroid size of the mandible is calculated as the mean values of 3 coordinates (x, y, z) for all the 14 landmarks assigned. Statistically it is the square root of the sum of the squared distances between each landmark and the centroid of the mandible and it is proportional to the square root of the mean of all squared landmarks distances. It is not a direct measure of the size, simply because it is calculated for different configurations of landmarks used to summarize the shape (Bookstein, 1991). Centroid size was calculated mainly to test for differences in size among populations and they were visualized using box plots.

3.2.2.6 Statistical analysis

The Covariance matrices obtained from the datasets were used to inspect mandible shape differentiation among and within populations from Heligoland and the mainland. The differentiation was first assigned using the multivariate analysis implemented in principal component analysis (PCA). PCA is a widely used method for exploratory multivariate analysis and one of its uses was applied here as an ordination method to inspect the principal features of shape variation in the dataset. Secondly canonical variate analysis (CVA) was used to assign features of shape variation with prior assumptions of known group membership. CVA was also used to calculate the Procrustes distances and their probabilities between all samples in each dataset using the built-in options in MorphoJ.

The CVA implemented in MorphoJ was used to determine the shape features that best distinguish among multiple groups of specimens assuming group membership to be known a priori. The matrices of pairwise Procrustes distances among all possible pairs of groups were calculated. The Procrustes distance was calculated as the square root of the sum of squared point distances between two shapes in superimposed configurations at centroid size (Klingenberg and McIntyre, 1998).

The CVA and discriminant function analyses were applied to deeply look into the observed pattern of mandible morphology differentiation among Heligoland and continental populations. The discriminant function applies a multivariate t-test that tests the equality of the means of the two given groups with variables being normally distributed and where the number of cases is at least two more than the number of variables. The discriminant function analysis is implemented in MorphoJ and the analysis automatically conduct a parametric T-square test for the difference between group means (Klingenberg, 2011).

3.3 Results

The coat coloration was scored in a total of 17 mice from Heligoland. Dorsally, the color ranges between brown (35.29%), dark brown (41.18%), and black (23.53%). For ventral color scoring almost all mice were creamy brown (70.59%) and a few were white (23.53%) and one was brown (5.88%). The most frequent combination among the mice from Heligoland was a dark brown back and a creamy brown belly (35.29%) with only one mouse with a white tail tip (1mm). This tip color is rarely seen in house mouse from the mainland.

All adult mice from Heligoland with tail length of ($>$ 74mm), hence were identified as *M. m. domsticus* and the Tail Body Length Ratio (TBLR), which is a reliable discriminator between subspecies was calculated. *M. m. musculus* tends to have a smaller TBLR than the other two subspecies (Boursot et al., 1993; Marshall and Sage, 1981). The TBLR mean for the mice from Heligoland is 1.015 (range: 0.90-1.113), compared with 0.664 (range: 0.611-0.756) for *M. m. musculus* from Kazakhstan and 0.940 (range: 0.873-1.045) and 0.915 (range: 0.865-1.011)for *M. m. domesticus* from Cologne Bonn and Massif Central respectively.

The difference between the mean Heligoland and the two subspecies values was highly significant for *M. m. musculus* ($P = 0.0001$) and not significant for the two *M. m. domesticus* populations ($P = 0.108$, two tailed t-test. Figure 3.3 shows the average number calculated for house mouse from Heligoland and other poulations from the mainland with error bars indicating standard deviation around the mean.

3.3.1 Mandible size among populations

The size differences among populations estimated from the centroid size were visualized using box plots and are illustrated in Figure 3.4. The centroid size for Heligoland is large compared to that of other populations from the mainland. More

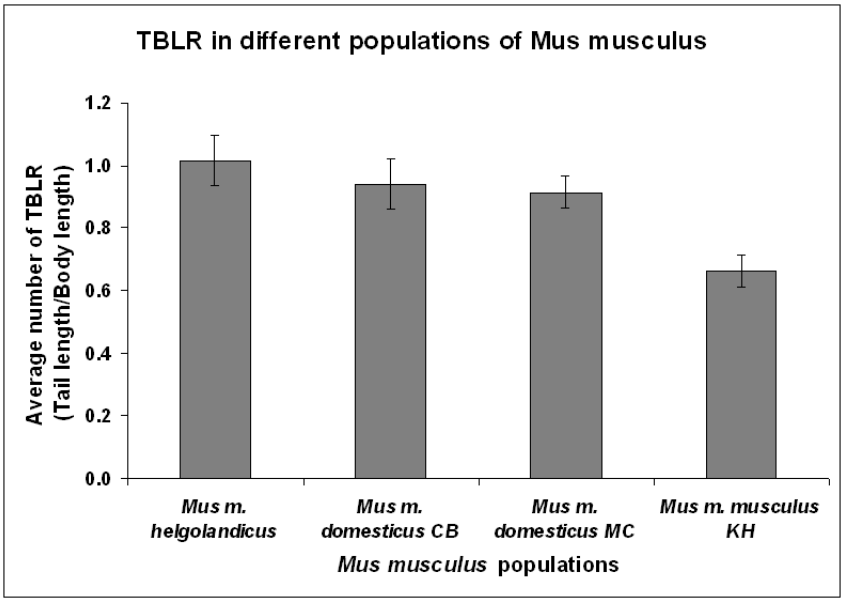


Figure 3.3: Chart of TBLR among *M. m. helgolandicus* and mainland populations

in details Heligoland (IFH) collection shows the largest size followed by Heligoland (ZMB) and (MPI) respectively. The difference between the means showed significant values between Heligoland and the other populations ($P<0.0001$).

3.3.2 Mandible shape differentiation

3.3.2.1 Mandible shape of the house mouse from Heligoland

The mandible shape differentiation among the specimens from Heligoland is observed from the results of variance in mandible shape summarized on the first two principal component axes (Figure 3.5). The PCA was based on a covariance matrix generated from the landmark coordinates after the Procrustes superimposition explained in section 3.2.2.4.

The PCA was analyzed for mouse mandible specimens from Heligoland collected at different time periods and will be referred to as follows, (ZMB, IFH, and MPI) in a chronological order based on collection time period (more details are explained under the *M. m. heligolandicus* section of this chapter). Notwithstanding, the first

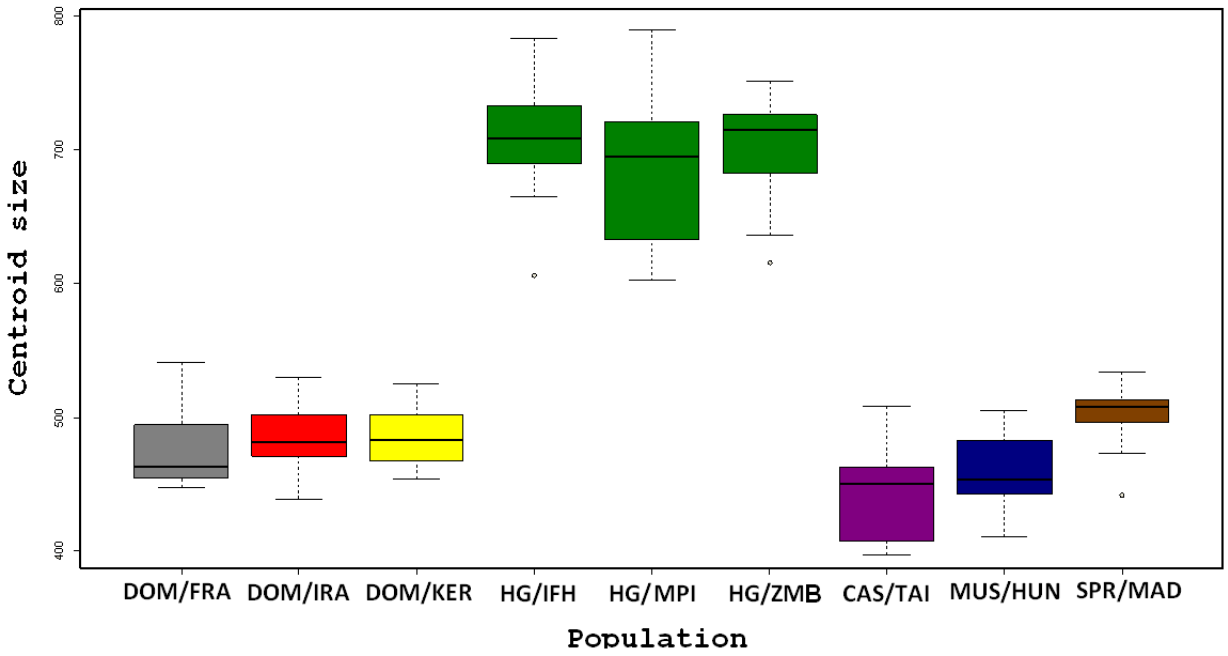


Figure 3.4: Box plot of centroid size in populations of the house mouse from Heligoland and the mainland. The house mice from Heligoland represented by the three different collections are shown in dark green color. Populations of *M. m. domesticus* from Frankfurt is shown in grey, from Iran is shown in red and from Kerguelen is shown in yellow. Population of *M. m. castaneus* origin is in violet, *M. m. musculus* is in blue and *M. spretus* is in brown. The abbreviations for the populations are detailed in Table 3.2.

axis (PC1) explaining 26.67% and the second axis (PC2) explaining 17.68% of the total variance, no clear differentiation between the specimens of the three collections from Heligoland is observed. The specimens are scattered along the first and the second axes (Figure 3.5).

To confirm the strength of the signal and to determine the possibility of sequence changes on the mandible morphology on Heligoland I also conducted a discriminant function analysis between pairs of populations using MorphoJ (Klingenberg, 2011). The analysis discriminated between each of the analysed pairs and the cross validation among population pairs showed overlapping among specimens from each of the analyzed pairs and was supported by the non significant parametric P-values between the pairs. The shape differences between population pairs from discriminant

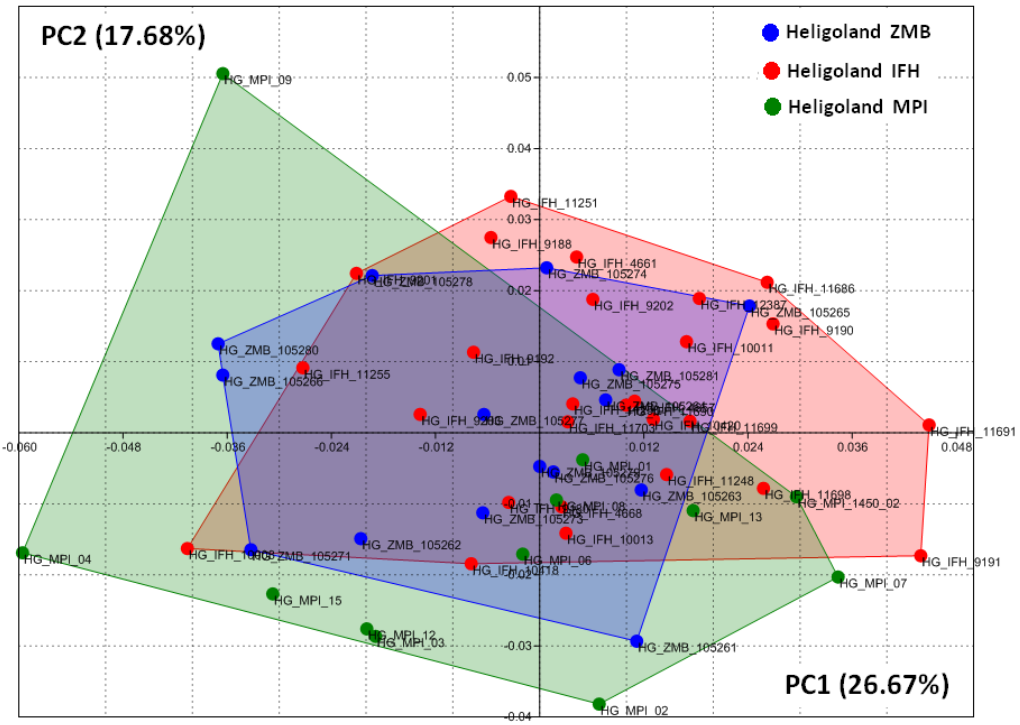


Figure 3.5: PCA scatter plot for the first two axes among populations from Heligoland.

function are illustrated in Figure 3.6 **A**, **B** & **C**.

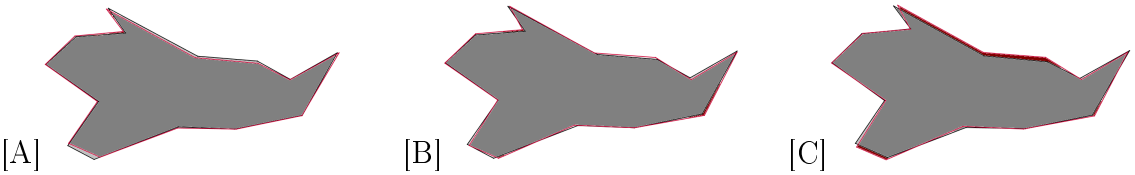


Figure 3.6: Wireframe graphs from discriminant function analysis among population pairs from Heligoland for the first axis (PC1). Each graph illustrates the shape change from black to red. **A)** Wireframe graph between Heligoland IFH (black) and MPI (red). **B)** Wireframe graph between Heligoland IFH (black) and ZMB (red). **C)** Wireframe graph between Heligoland MPI (black) and ZMB (red).

3.3.2.2 Mandible shape of house mouse between Heligoland and mainland populations

The pairwise Procrustes distances and their *P*-values shown in Table 3.3 & 3.4 respectively show that the three populations collected from Heligoland are significantly

distant from the other continental populations. In addition significant differences are also observed among continental populations and coin with patterns of mandible morphology found previously for the same populations by Boell & Tautz (2011) who pointed to the ground of natural variations among wild populations.

Table 3.3: Pairwise Procrustes distances among Heligoland and continental populations.

	Dom-Frankfurt	Dom-Iran	Dom-Kerguelen	Helgo-IFH	Helgo-MPI	Mus-Castaneus	Mus-Hungary	Spre-Madrid
Dom-Iran	0.038							
Dom-Kerguelen	0.0326	0.0391						
Helgo-IFH	0.0423	0.058	0.0525					
Helgo-MPI	0.0376	0.0490	0.0493	0.0241				
Mus-Castaneus	0.0247	0.0283	0.0294	0.0458	0.0385			
Mus-Hungary	0.0264	0.0245	0.0376	0.0508	0.0464	0.0251		
Spre-Madrid	0.051	0.0311	0.0524	0.0616	0.0559	0.0477	0.0377	
Helgo-ZMB	0.0362	0.0505	0.0477	0.0232	0.0260	0.0374	0.0439	0.0579

Table 3.4: *P*-values from permutation tests (10000 permutation rounds) for pairwise Procrustes distances among Heligoland and continental populations. Significant probabilities after Bonferroni correction are in bold

	Dom-Frankfurt	Dom-Iran	Dom-Kerguelen	Helgo-IFH	Helgo-MPI	Mus-Castaneus	Mus-Hungary	Spre-Madrid
Dom-Iran	**<0.0001**							
Dom-Kerguelen	**0.0001**	**<0.0001**						
Helgo-IFH	**<0.0001**	**<0.0001**	**<0.0001**					
Helgo-MPI	**0.0001**	**<0.0001**	**<0.0001**	**0.0089**				
Mus-Castaneus	**0.0263**	**0.0052**	**0.0026**	**<0.0001**	**<0.0001**			
Mus-Hungary	**0.0035**	**0.013**	**<0.0001**	**<0.0001**	**<0.0001**	**0.022**		
Spre-Madrid	**<0.0001**	**0.0004**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**0.0001**	
Helgo-ZMB	**<0.0001**	**<0.0001**	**<0.0001**	**0.0006**	**0.0104**	**<0.0001**	**<0.0001**	**<0.0001**

The constructed neighbor joining tree based on the Procrustes distances among populations is shown in Figure 3.7 with branch length data. There is a clear division between the house mice from Heligoland and the other subspecies included in the analysis.

Patterns of differentiation between mainland and island populations was visualized using PC analysis and are illustrated in Figure 3.8. The first axis (PC1 = 22.9% of variance and PC2 = 21.3%) are mainly determined by the differences between Heligoland (red) and the other species and subspecies (different colors).

The populations from Heligoland differed largely from their counterparts from the mainland and the shape difference on the first axis implies major shape changes

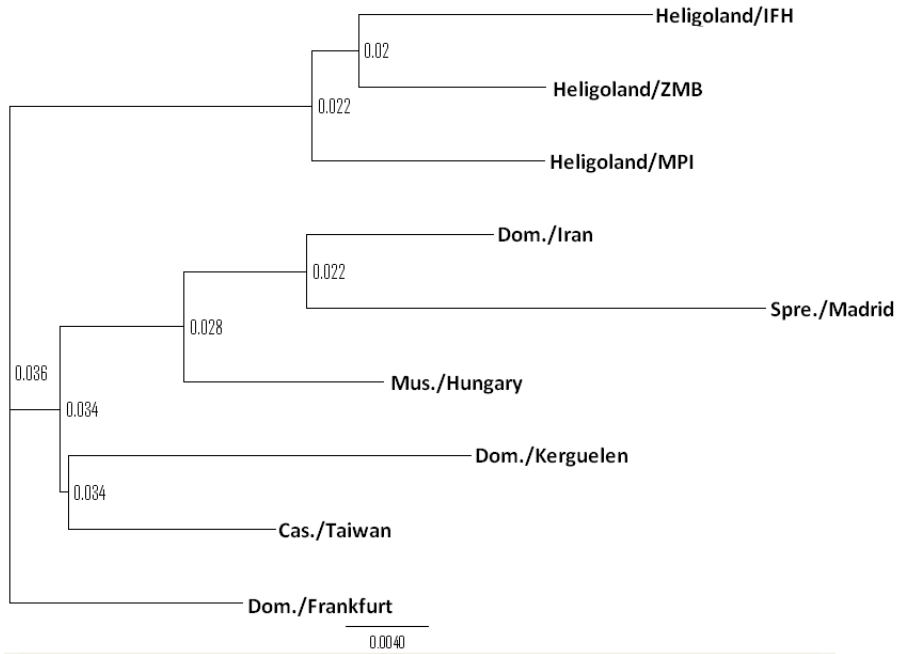


Figure 3.7: Neighbor joining tree based on pairwise Procrustes distances among populations

led by the elongated incisor zone, the narrower and shorter coronoid and condylar processes and the narrower angular process. In contrast, the second axis mainly concerns the shape changes led by elongated incisor zone, wider molar zone, sharper condylar and shorter angular and coronoid processes.

The shape changes along the first two PCs is not shown for the whole data set comparisons, but more specifically detailed comparisons between Heligoland and a population representing *M. m. domesticus* from Frankfurt and Kerguelen island and *M. m. musculus* from Hungary will be explained in the following text.

To assign any patterns of shape variation among different collections from Heligoland and the closest mainland population, I analyzed a smaller data set which included the three collections from Heligoland along with the population of *M. m. domesticus* origin from Frankfurt in Germany. I conducted a principal component analysis based on the mandible Procrustes coordinates and the results were visualized using PCA and are shown in Figures 3.9 & 3.10. The first axis (PC1) explains

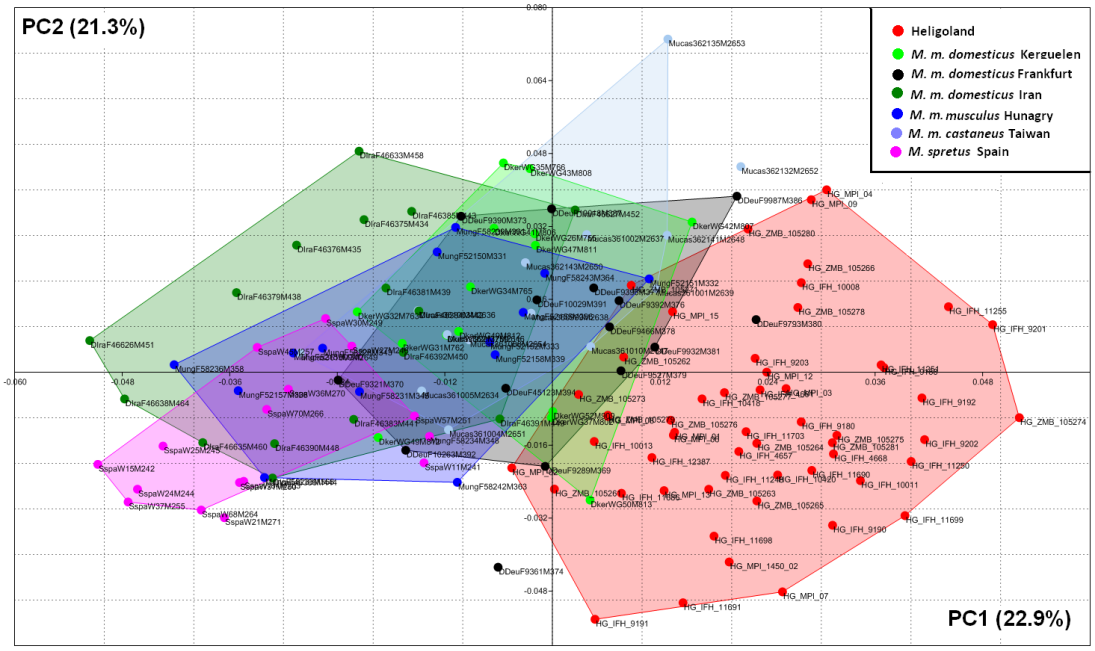


Figure 3.8: Mandible shape variation among *M. m. helgolandicus* and continental populations represented on the first two axes of a PCA scatter plot.

(26.34%) of the total variance and shows some differentiation between the specimens of the three collections from Heligoland and the population of *M. m. domesticus*. These concern shape changes in the ramus region of the mandible in particular the sharper condylar process and the shorter coronoid process. The variance is also featured by the deeper incisor alveolus and the elongated incisor zone. The second axes (PC2 = 15.16%) of the total variance is driven by a differentiation between Heligoland and the population of *M. m. domesticus* and is based mainly on a narrower incisor, molar, and ramus zones.

The shape differences among populations from Heligoland and *M. m. domesticus* population from Frankfurt are represented for the first two PCs in a wireframe graph in Figure 3.10 **A** & **B**. The observed differences between the house mouse of Heligoland and continental population were mostly pronounced through the elongation of the mandible on the first axis and the narrowing on the second axes of the principal components.

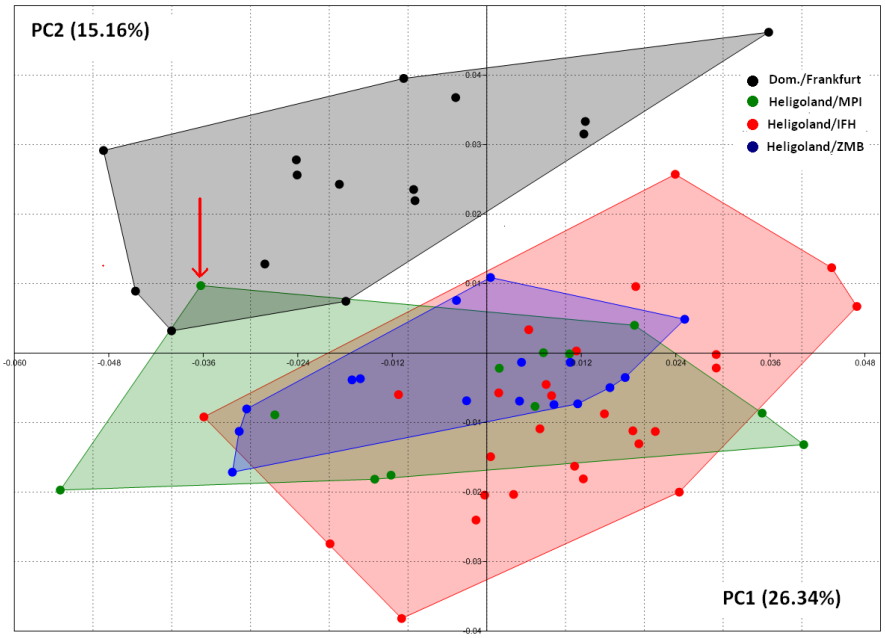


Figure 3.9: Scatter plot for the first two PCs for *M. m. helgolandicus* and *M. m. domesticus* population from Frankfurt.

Figure 3.10: Shape changes along the first two PCs between *M. m. helgolandicus* and *M. m. domesticus* populations in a wireframe graph. Shape changes are from grey *(domesticus)* to red *(helgolandicus)*.

The observed scattering along the first axis caused by one specimen from Heligoland recent collection (specimen is pointed by red arrow) is supporting the mtDNA haplotype data analysed in section 2. It showed that this mouse has a haplotype not specific to Heligoland. Similar to the haplotype data, the observation from the PC analysis confirms that this single specimen shows a mandibular morphology not specific to Heligoland and probably represents a recent invader from the mainland.

The canonical variate analysis which is more powerful in disentangling hidden signals was analyzed between Heligoland and the population of *M. m. domesticus*. It showed a concerted signal of distinct mandible morphology observed through shape changes from the mainland population to Heligoland. Results are illustrated in Figure 3.11.

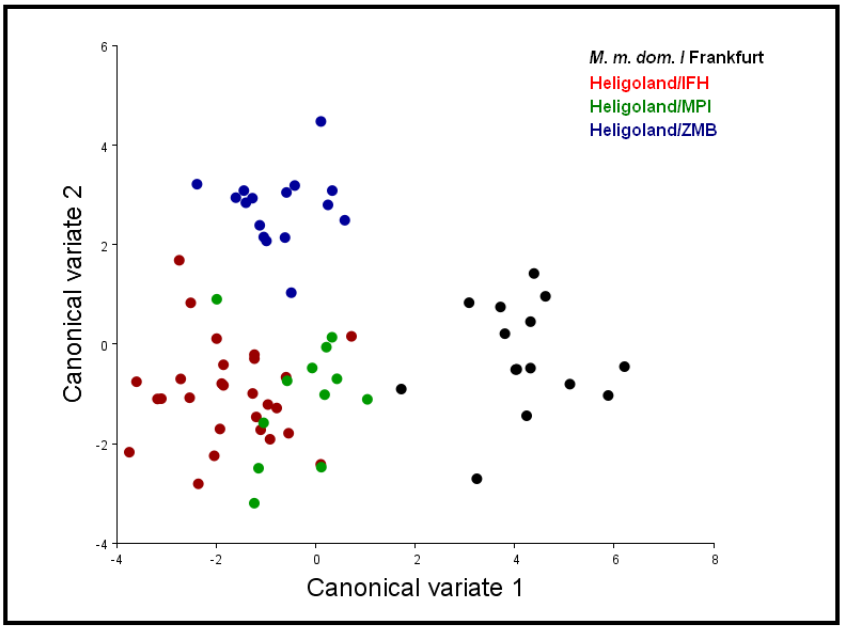


Figure 3.11: Canonical variate analysis (CVA) scatter plot of CV1 and CV2 among populations from Heligoland and a population of *M. m. domesticus* from Frankfurt.

In the discriminant function analysis, all the specimens from Heligoland and Frankfurt were correctly identified to their respective group. Three specimens from Heligoland were misclassified as from the mainland, however when the discriminant frequencies were plotted there were no obvious overlap between the groups. The discriminant analysis confirms the observed pattern of insular evolution revealed here and reflected by the distinct mandible morphology of the house mouse from Heligoland. The P-value from the t-test is significant among Heligoland and the continental population and the details of the test are summarized in Table 3.5). The output from the discriminant function analysis showing the discriminant histograms and their frequencies are illustrated in Figure 3.12.

Table 3.5: Discriminant function analysis between Heligoland and a population of *M. m. domesticus* origin from Frankfurt/Germany

Discriminant Function Analysis			
Difference between means			
Procrustes distance	0.03726244		
Mahalanobis distance	5.1131		
T-square: 308.1291	P-value < 0.0001		
From discriminant function			
True	Allocated to		
Group	*M. m. domesticus*/Frankfurt	*M. m. helgolandicus*/Heligoland	Total
M. m. domesticus/Frankfurt	15	0	15
M. m. helgolandicus/Heligoland	0	55	55
From cross-validation			
True	Allocated to		
Group	*M. m. domesticus*/Frankfurt	*M. m. helgolandicus*/Heligoland	Total
M. m. domesticus/Frankfurt	12	3	15
M. m. helgolandicus/Heligoland	4	51	55

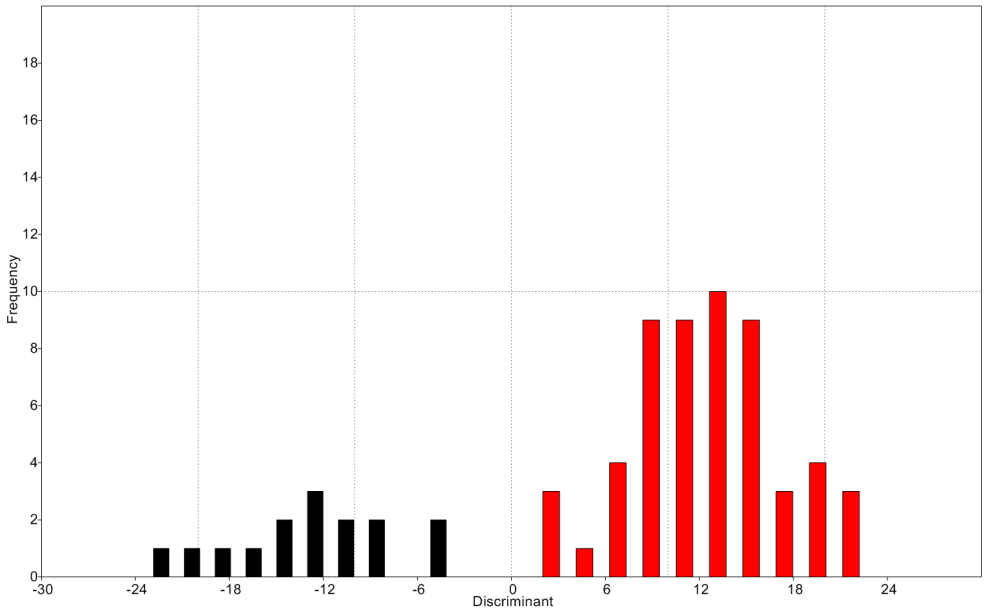


Figure 3.12: Histogram showing the discriminant function analysis between *M. m. helgolandicus* (red columns) and mainland population from Frankfurt in Germany representing *M. m. domesticus* subspecies (blck columns) plotted on the x-axis and their frequencies on the Y- axis

The PCA scatter plot between Heligoland and *M. m. musculus* population from Hungary is illustrated in Figure 3.13, the first axis explains 22.81% of the total variance and second axis explains 20.25% of the variance. The shape differences among the populations are represented for the first two PCs in a wireframe graph in Figure 3.14 **A** and **B**. The observed differences between the house mouse of

Heligoland and the continental population were mostly pronounced on the first axis through the narrower angular process, shorter condylar process and wider coronoid process. On the other hand the second axis reflects the variation mainly through the elongated mandible and shorter coronoid and angular processes.

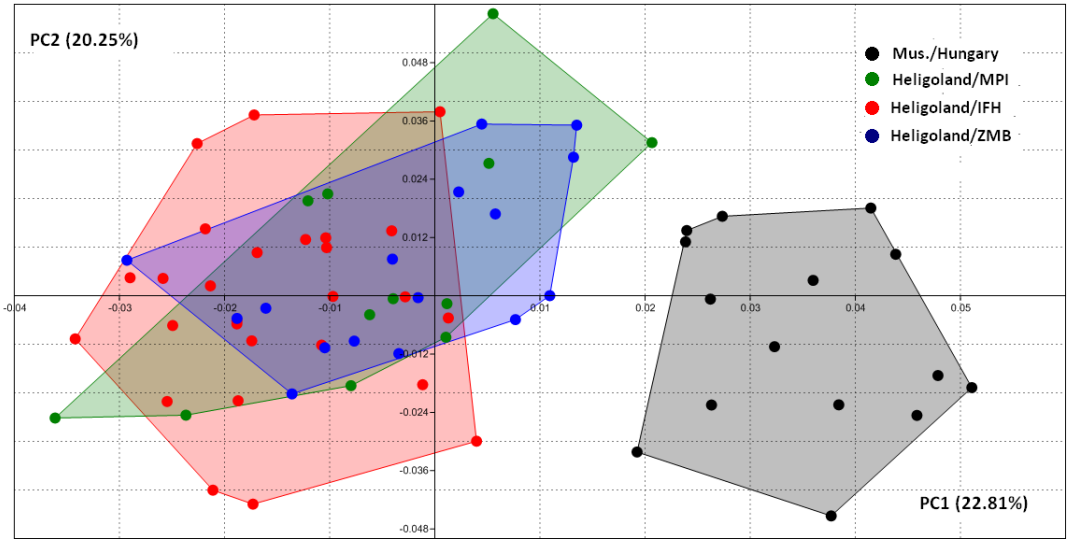


Figure 3.13: Scatter plot for the first two PCs for *M. m. helgolandicus* and *M. m. musculus* population from Hungary.

Figure 3.14: Shape changes along the first two PCs between *M. m. helgolandicus* and *M. m. musculus* population in a wireframe graph. Shape changes are from grey *(musculus)* to red *(helgolandicus)*.

To better understand the novelty of the mandible morphology on Heligoland. I compared the mandible of the different collections from Heligoland to that of *M. m. domesticus* population from Kerguelen island. The generated PCA scatter plot between Heligoland and *M. m. domesticus* population from Kerguelen island is illustrated in Figure 3.15, the first axis explains 28.13% of the total variance

and second axis explains 16.36%. The shape differences among populations from Heligoland and that of Kerguelen population are represented for the first two PCs in a wireframe graph in Figure 3.16 **A** and **B**.

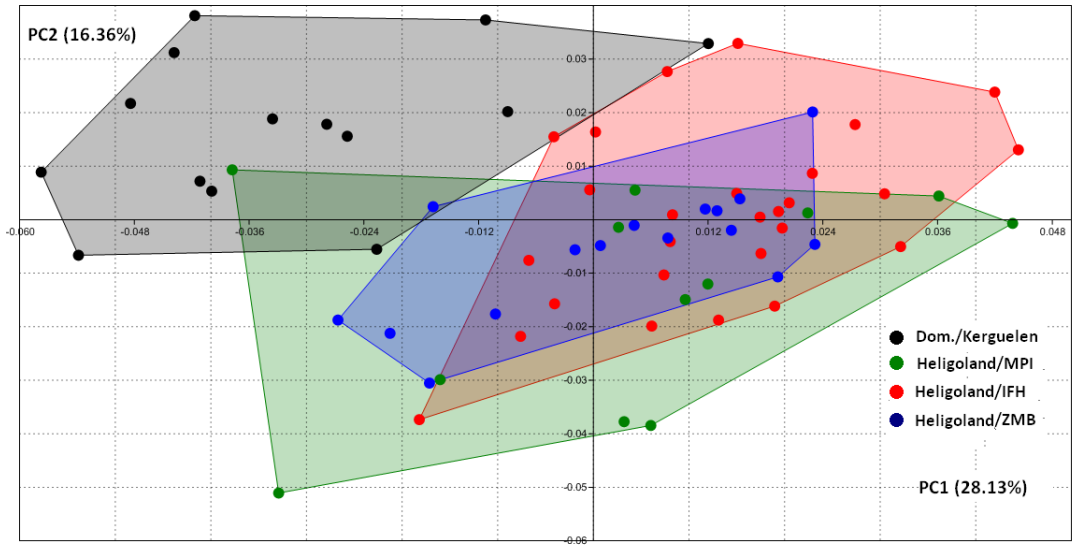


Figure 3.15: Scatter plot for the first two PCs for *M. m. helgolandicus* and *M. m. domesticus* population from Kerguelen.

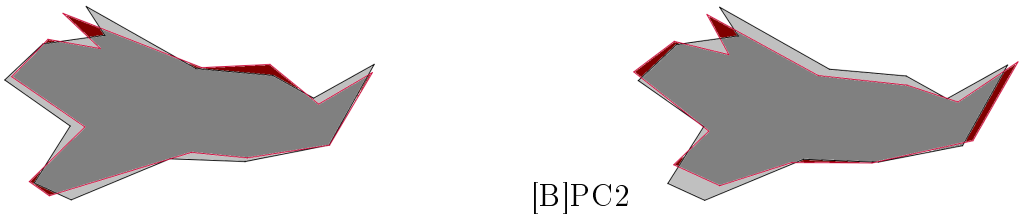


Figure 3.16: Shape changes along the first two PCs between *M. m. helgolandicus* and *M. m. domesticus* population from Kerguelen Island in a wireframe graph. Shape changes are from grey *(domesticus)* to red *(helgolandicus)*.

The observed differences between the house mouse of Heligoland and the island population were mostly pronounced through the narrower incisor zone, coronoid, angular, and condylar processes. In contrast, mandible elongation and shortening of the angular process were the major variants on the second axis.

Table 3.6: Discriminant function analysis between *M. m. helgolandicus* and *M. m. domesticus* from Kerguelen

Discriminant Function Analysis			
Difference between means			
Procrustes distance	0.04856136		
Mahalanobis distance	8.1114		
T-square: 734.2244	P-value < 0.0001		
From discriminant function			
True	Allocated to		
Group	*M. m. domesticus*/Kerguelen	*M. m. helgolandicus*/Heligoland	Total
M. m. domesticus/Kerguelen	14	0	15
M. m. helgolandicus/Heligoland	0	55	55
From cross-validation			
True	Allocated to		
Group	*M. m. domesticus*/Kerguelen	*M. m. helgolandicus*/Heligoland	Total
M. m. domesticus/Kerguelen	14	0	14
M. m. helgolandicus/Heligoland	0	55	55

3.4 Discussion

This study applies geometric morphometrics, in particular, the mandible landmarking approach. The mandible is a well characterized phenotypic structure that is related to food processing and hence to the fitness of the organism. In addition, it has widely been studied in the house mouse for both lab strains and wild populations. This chapter reveals the aspects of insular evolution of the house mouse inhabiting the small island of Heligoland and clarifies previous findings on general skull shape and body measurements which all led to local classification of these mice defined as a separate subspecies *M.m. helgolandicus* from the other two subspecies inhabiting Europe *M. m. domesticus* and *M. m. musculus*.

The coat coloration observed in the mice of Heligoland was not indicative of their origin given that coloration is variable in all three subspecies of the genus *Mus*. This character is not diagnostic and I cannot consider it here as a taxonomic feature (Boursot et al., 1993; Marshall and Sage, 1981). On the other hand, the relative tail length of the mice from Heligoland showed significant mean differences from *M. m. musculus*, which in this case is indicative of the *M. m. domesticus* like mice.

Analysis of the centroid size showed that the mice of Heligoland have a significantly larger mandible than other subspecies and species included in this analysis. Both findings from the analysis of TBLR and centriod size of house mice from Heligoland are likely linked to morphological changes due to adaptation to the new environment on the island. This supports the notion of an ongoing insular evolution on the island and more specifically the "island rule" that small mammals grow bigger on small islands (Berry, 1996; Foster, 1964; Heany, 1978; Lomolino, 1985, 2005; Sondaar, 1991; VanValen, 1973).

Patterns of mandible morphology of house mouse from Heligoland

In the present study, the house mice of Heligoland collected at different time periods present distinct mandible morphology than their counterparts from the mainland and an archipelago (Kerguelen house mice). These findings support previous studies (Reichstein and Vauk, 1968; Zimmermann, 1953) in a sense that the house mouse of Heligoland is distinct from the mainland subspecies. Even though the previous studies focused on the total skull shape and used conventional methodology, I would like to claim here that those conclusions were supported in this study through advanced geometric morphometrics methodology.

The results from the principal component and discriminant analyses on the mandible morphology showed no clear evidence for an ongoing morphological sequence of changes on Heligoland among the three collections. Hence, this is indicative of settlement of mouse population and stability on the island which is line in with findings from molecular analysis in chapter 2.

Moreover, when the mandible morphology among Heligoland put into a wider context by including continental populations, the insular divergence is obviously observed. The main factors that could have played a role in shaping the mandible

morphology of house mouse from Heligoland are isolation, environmental factors and the suggested colonization history of the island; where colonization occurred from a single mitochondrial haplotype 400 years ago which was revealed mainly from results of molecular analysis.

Results from the PC analysis showed that the mandible of *M. m. helgolandicus* is distinct from other continental populations and that was not led by the general mandible shape, but more specifically with some parts of the mandible such as the angular process, coronoid and condylar processes. These findings might reflect the levels of plasticity on the evolving mandible on Heligoland and they are influenced by the mastication process because muscles are jointly part of these structures (Renaud and Auffray, 2010).

The distinct mandible of *M. m. helgolandicus* from the two major subspecies inhabiting Europe and from other species e.g. *M. spretus* is indicative of an island specific morphology and hence no ancestry can be revealed from the morphological data for these mice.

Hence, these mice reflect a different morphology of the mandible than the so called ancestral morphology. In this analysis Iranian specimens are considered representatives of the ancestral mandible morphology which was pointed out by (Siahsarvie et al., 2012) in their large survey which included a wide range of populations. That study found out that populations from Iran are closer in their morphology to the ancestral morphology which had suggested a possible Iranian origin of the house mouse along with the findings that this region is the origin of the preserved ancestral morphology. These results also provide evidence for the previously revealed patterns of the morphological evolution in the mandible of different house mouse subspecies in that the more geographically closer subspecies are to the suggested origin of the species the less variable their morphologies. In contrast, the study found that the peripheral subspecies which are geographically distant from the suggested origin are

morphologically variable and that the variation is 10 fold greater than that of the center (Siahsarvie et al., 2012). These previous results are in line with my results thus confirming that the house mouse from Heligoland is a peripheral population and that its invasive characteristic gave the mandible its observed morphology. This is mostly led by the evolutionary plasticity, selection and the nature of the changing environment.

On the other hand, the distinct morphology of these mice from the invasive house mice inhabiting Kerguelen archipelago shows an evidence for the influence of the variable environments and the nature and consistency of diet on each island. This finding with previous findings by Boell and Tautz (2011) and Renaud et al. (2013) on the invasive subspecies on Kerguelen archipelago show that if the insular environment on Heligoland is similar to that of the Sub-antarctic archipelago, we would expect the mandible of *M. m. helgolandicus* to resemble that of Kerguelen. Furthermore, the distinctive mandible of *M. m. helgolandicus* from that of Kerguelen, suggests not only a contrasting environment, but also a different adaptive responses to diet shifts.

In this study, the major PCs are associated with variation of the angular and coronoid processes, which were found previously to be part of a single functional complex that serves for attachment of the masticatory musculature (Atchley and Hall, 1991; Klingenberg et al., 2001). Studying the complex genetic architecture of these mandibular zones in which several genes are correlated with mandible changes might also be highly considered in such an analysis to point these effects in different species and at different timescales (Klingenberg et al., 2001). The angular process was also found to be more prone to bone remodeling during late postnatal growth (Renaud et al., 2010).

Insular adaptation

The mild climate on Heligoland, which is almost an offshore climate (oceanic climate), is generally similar to Western European climate conditions and cannot explain the divergence of species. However, variable resources of food available for mice on the island are distributed with respect to the geological structure of the island (upper and lower land) which could have an impact on the mice colonies establishment. Additionally, the high level of bird migrations with different bird species coming to the island all year long might have had an influence especially when rodent predators are considered such as eagles and sea gulls.

The most striking finding from the geometric morphometrics analysis is that the mandible of *M. m. helgolandicus* has a distinctive shape from the other mainland mandibles. The morphological analysis of the mandible of *M. m. helgolandicus* compared to other populations from the mainland showed that the mandible is characterized by elongation and sharp angular and condylar processes. These characteristics are pointing toward a carnivorous/insectivorous diet which is in contrast to the known diet for the genus *Mus* being omnivorous. On a similar perspective, diet shift from plant seed to macroinvertebrates had been documented for mice on sub-Antarctic islands (Smith et al., 2002) which was indicative of local adaptations to environmental changes. Moreover, these findings might better explain the resilience of new waves of colonization on islands, where new arrivals need not only to be introduced to the established population, but also high competition capabilities (Hardouin et al., 2010).

Although the distinct mandible of *M. m. helgolandicus* could have been influenced to some extent by the different collections preparation protocols, it has been shown previously that these factors have minimal effects and don't hinder the observation of an ongoing morphological change and its underlying genetic patterns

(Boell and Tautz, 2011). Moreover, the hypothesized origin of the ancestors of *M. m. helgolandicus* suggested by this study is from Western Europe and most likely of *M. m. domesticus* origin, that is to say if the genetic background is the major factor shaping the mandible of Heligoland house mouse I could have observed some patterns similar or close to patterns from that region.

Hence, it is more obvious that the morphological changes of the mandible from Heligoland arose most likely as a result to positive selection and adaptation to the new environment on the island. This is due to the fact that adult traits such as the mandible are highly integrated with the fitness of the individual and hence influenced by selection pressure. These findings supports previous findings by Boell and Tautz (2011) where their study included the same continental and island populations for the same specified landmarks along with inbred strains derived from wild populations under identical conditions. The authors concluded that adaptive evolution may contribute to the mandible shape changes between populations, mostly pronounced in newly colonized niches.

Overall, the findings of this study shed light on the importance of islands fauna and their inhabitants being powerful resources for understanding the processes of phenotypic divergence for such a morphological character as the mandible. Moreover, the results from this study point towards the ongoing phenotypic divergence that is strongly related to the onset of speciation and biological diversity. Noteworthy, these findings are supporting the idea of adaptive changes facing *M. m. helgolandicus* which are mainly the new environment of Heligoland in contrast to that of the mainland. Along with the diet type there which is also suggested to be different from that on the mainland.

Extended sampling from Heligoland and the involvement of fossil materials might provide evidence for the direction of shape changes from the first possible ancestors invaded the island. Additionally, in depth molecular analysis for the museum ma-

terial might reveal the cryptic patterns of colonization history on Heligoland which were only reflected here by samples from the contemporary mouse collection.

4 Patterns of introgression in *M. m. helgolandicus*

4.1 Introduction

From DNA molecules to Whole genome sequencing

Genome re-sequencing is the process that consists of three basic steps comprising sample preparation, physical sequencing and re-assembly. The sample preparation step includes the breaking down of the DNA template into smaller fragments, which will be amplified into multiple copies using a variety of molecular methods such as the polymerase chain recation (PCR). The physical sequencing step is the process of determining the sequence of the nucleotide bases (A, G, C, and T) within each fragment of DNA and the identified number of bases in each fragment is known as the read length. In the re-assembly step a software is used to align the overlapping reads and hence allows the original genome to be assembled into contiguous sequences (Schatz et al., 2010).

New methods referred to as next-generation sequencing have been developed with the application of different methodologies that are all based on a template preparation, sequencing and imaging, and following steps of genome alignment and assembly (Metzker, 2010). The sequencing of many millions of DNA fragments in parallel is the bases of the next generation sequencing technologies which are widely being used today such as the pyrosequencing method, the 454 Genome Sequencer (http://www.454.com), the reversible dye-terminator-based Illumina Genome Analyser, HiSeq and MiSeq (http://www.illumina.com), the ligation-based SOLiD Genome Sequencer (http://www.lifetechnologies.com) and a semiconductor-based Ion Per-

sonal Genome Machine (PGM) and Ion Proton (http://www.lifetechnologies.com/us/en/home/brands/iontorrent.html) (Yalcin et al., 2012). Moreover, the application of next generation sequencing technologies is largely growing. It includes different aspects of molecular research such as de novo genome sequencing, re-sequencing, detection of coding and non-coding transcripts, identification of sequence variants, epigenetic profiling, and interaction mapping (Minoche et al., 2011).

The increasing demand for large survey studies on individual's genetic variation and population diversity has resulted in a great shift from conventional sequencing methods and microarray based methods to whole genome sequencing, which has turned out into high-throughput data with a growing demand for analytical tools and statistical methods for whole genome comparisons that all with human analysis will answer questions of concern in evolutionary biology and therapeutic approaches (Kirkness and C., 2010). It is noteworthy that the third generation sequencing technologies are under development with optimistically lower running costs (Kim et al., 2014).

Adaptive evolution in the house mouse *Mus musculus*

One of the major goals in evolutionary biology is to reveal the forces leading to population divergence both genetically and phenotypically. The evolution of genomes is shaped by different processes and it has been apparent that it depends on the balance between mutation, neutral evolution, selection and adaptation which are still only partly understood.

Positive selection, which is known as the tendency of beneficial traits to increase in prevalence (frequency) in a population, has an important role in the evolution of the house mouse, as it is the driving force behind evolutionary adaptation. Simply, for a trait to undergo positive selection, it must have two characteristics. First, the trait must be beneficial; in other words, it must increase the organism's fitness for

surviving and reproducing. Second, the trait must be heritable so that it can be passed to the next generation. Beneficial traits are extremely varied and may include anything from protective coloration, to the ability to utilize a new food source, to a change in size or shape that might be useful in a particular environment. If a trait results in more offspring who share the trait, then that trait is more likely to become common in the population than a trait that arises randomly.

The influence of selection on populations might mainly depend on the amount of gene flow between populations. Selection will drive phenotypic divergence when there is a limited gene flow among populations, whereas the time since population divergence will drive neutral genetic divergence (Ogden and Thorpe, 2002). However, if gene flow is evident between populations, there might be a positive correlation between phenotypic and genetic divergence because local adaptation can act to reduce gene flow among populations (e.g. selection against migrants or hybrids) (Egan et al., 2008). Indeed, in many species, there is a correlation between ecological and genetic divergence (Nosil et al., 2009). Thus, measuring gene flow among populations is key to deciphering the evolutionary forces leading to phenotypic differentiation (Domingues et al., 2012).

Studies of genome scans for selective sweeps have shown that loci under positive selection can be identified in natural populations and that sweep signatures might also result from effects of drift on populations under demographic factors such as population bottlenecks (Akey, 2009; Oleksyk et al., 2010). Hence, statistical methods have been developed to allow distinguishing sweep signatures from effects of drift from those of positive selection with an increasing input from recent advances in high-throughput genome data (Sabeti et al., 2006; Tang et al., 2007). Selection might target phenotypes related to morphology, physiology, immune response or reproductivity of the organism. Hence, detection of signatures of selective sweeps has been of great importance for the detection of adaptive trait loci in natural popula-

tions of the house mouse, mainly by typing neutral markers and statistically assign regions of reduced polymorphism in different populations (Ihle et al., 2006).

Of interest are invasive species which colonize new habitats and hence are considered model system for such studies mainly to decipher the role of different evolutionary forces in population differentiation. Islands which often referred to as "natural laboratories" (Mayr, 1942) are influenced by genetic drift which plays a larger role on small founding populations and natural selection which is more powerful when invasive population colonize new habitats with novel environmental conditions to which the invasive population must adapt to (Mullen et al., 2009).

Gene flow and introgression

Hybridization and introgression are well known to play a critical role in the evolution of species. Allelic introgression from closely related species or other species is a major factor that has shown important relevance for understanding the genetic composition of wild populations. The importance of such a factor resulted from the fact that wild populations can hybridize in nature, and this has been found in many plant species and has been proposed recently to be ongoing in animal populations (Mallet, 2007). Hybridization has been highly proven to taking more advantageous roles in plant species and that on average around 25% of plant species hybridize with at least one other species. In contrast, hybridization in animals is more controversial, with around 10% of animal species that are known to hybridize (Mallet, 2005). Recent advances in genome sequencing and the increasing capacity and power of statistical tests have contributed much to the understanding of hybridization and introgression mechanisms.

Hybridization can only takes place when reproductive isolation is weak, in such a case genetically divergent individuals (representing different subspecies, species) crosses and produce genotypes with less fertility or less viability than the crosses

between genetically similar individuals (Arnold, 1992; Arnold et al., 1999).

Arnold et al. (1999) reviewed the controversial assumptions based on natural hybridization, mainly that the evolutionary history of hybridizing forms will not be influenced by natural hybridization as the probability of producing novel genotypes with higher relative fitness is low, and that all hybrid genotypes will be less fit. In addition to that, hybridization sometimes is likely to be influenced by environmental factors such as the intervention of human and climate changes.

In recent years extensive studies on different biological systems have been conducted pointing toward the creativity of hybridization leading to the emergence of new species/subspecies from two different species, however bearing species-like characteristics. The newly formed species "hybrid species" can colonize a new niche where none of its parental species could be found (Mallet, 2007; Nolte and Tautz, 2010). Hence, contrasting previous assumptions which mostly looked at hybridization as destructive. Nowadays the feasibility of the available genotyping techniques and the massive amount of data are very useful for studying cryptic population structure. For example, sequences of mtDNA and Y-chromosome genes have been useful for studying population hybridization, because they have no recombination events and thus retain the genetic information of parental populations (Avice, 2000).

Genome patterns of introgression in the house mouse *M. musculus*

During the few past decades research studies have focused on the genetics of reproductive isolation and identified some genetic components of reproductive isolation mainly through laboratory crosses between the three major subspecies *M. m. musculus*, *M. m. domesticus*, and *M. m. castaneus* and also between strains derived from these subspecies. Among the house mouse, there are cases of hybridization between subspecies that have been reported in the wild (Nunome et al., 2010; Teeter

et al., 2007; Ďureje et al., 2012; Yonekawa et al., 1988). The study of hybrid mice started long ago and was mostly concerned with regions of secondary contact and regions spanning the hybrid zones of house mouse with more attention given to the European hybrid zone (Ďureje et al., 2012). On there analysis, (Teeter et al., 2007) surveyed large number of autosomes across the mouse genome and used the patterns of introgression to map genomic regions that contribute to the maintenance of genetic isolation between recently diverged species. In contrast, other studies focused on isolated populations e.g. in New Zealand which revealed evidence for complex patterns of introgression that reflects the recent and still going hybridization (Searle et al., 2009a). Hence, island as a natural laboratory for evolutionary studies, serves as a great potential to deciphering patterns of introgression and positive selection.

Populations with admixed genomes arise when mating occurs between individuals from reproductively isolated populations where geographical barriers are not powerful enough to hinder gene flow. The genomes of admixed individuals consist of chromosomal fragments of distinct ancestry from each of the ancestral populations. For example, the genome of hybrid mouse from the European hybrid zone contain segments of both *M. m. musculus* and *M. m. domesticus* ancestry, to say that at a specific chromosomal location in the genome of that individual is expected to inherit 0,1 or 2 copies of *M. m. musculus* ancestry and vice verse. The inference of chromosomal fragments of distinct ancestry and their frequency across the genome, have important role in re-constructing the population history of the studied species/subspecies.

Staubach et al. (2012) studied the patterns of allelic introgression in natural populations of the house mouse *M. musculus* each from *M. m. domesticus* and *M. m. musculus* subspecies. The *M. m. domesticus* populations were from Southern France and Western Germany. On the other hand, the *M. m. musculus* populations were from the Czech Republic and Kazakhstan. The analysis on patterns of introgression

was based on the Affymetrix Mouse Diversity Genotyping Array. The array was used to genotype 11 wild caught individuals per populations of each subspecies. It was used given that it was designed to cover the variation of *M. m. domesticus* and *M. m. musculus*, where the two subspecies constituting high percentages of the genome of laboratory inbred strains. The microarray was designed with more than half million SNPs with 1 SNP/5 kb. The genotyped data has in total 471,271 SNPs, less than the stated number of SNPs defined by the mouse genome diversity microarray, as the authors corrected for the false positive SNPs applying a filtering step with the RLMP algorithm (Staubach et al., 2012).

The study revealed that populations of the *M. m. musculus* subspecies have on average a larger introgressed regions than *M. m. domesticus* subspecies and at least 10% of the mouse genome is subject to introgression. The patterns of introgression were revealed noticing that some of the haplotypes in a given population were more similar to haplotypes found in the other subspecies. This pattern was known as the pattern of population specific introgression and was observed across the subspecies boundaries.

Simulations as powerful tools for the assessment of the best model of fit were applied in the previous study to better explain the observed patterns of introgression with different possible scenarios. The observed patterns of introgression of haplotypes were assessed for the natural populations of the house mouse with a series of simulations including varying population size and different rates of migration.

The study pointed toward a major finding that the genetic make-up of the mouse genome is primarily shaped by selective sweeps and adaptive introgression (Staubach et al., 2012). Furthermore, the study found that the frequency and size of introgressed regions assigned and their distribution can not be explained by a neutral introgression model. These results are illustrated in Figure 4.1.

Although the molecular analysis to characterize the origin of the house mouse

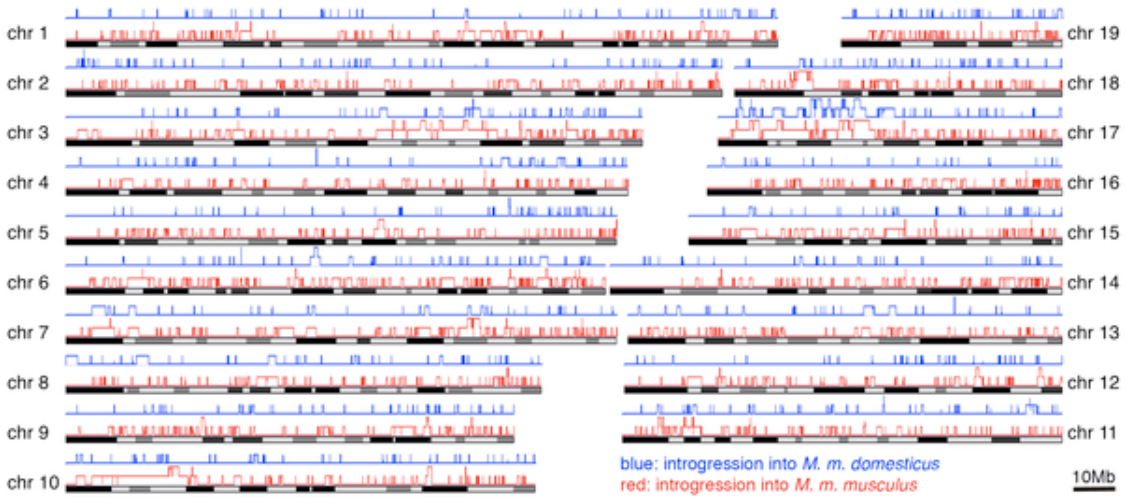


Figure 4.1: Distribution of introgressed regions across the mouse genome. The introgressed regions into *M. m. domesticus* are shown in blue and into *M. m. musculus* are in red. Elevated blocks indicate regions found in both populations of the respective subspecies. Figure was taken from (Staubach et al., 2012)

of Heligoland in the context of *M. m. musculus* and *M. m. domesticus* from this study has showed that the mice of Heligoland are featuring *M. m. domesticus* subspecies, the morphometrics analysis from this study confirmed the findings of the two previous studies that the house mouse of Heligoland is assigned to a distinct subspecies (Reichstein and Vauk, 1968; Zimmermann, 1953). These findings were motivating to sequence the whole genome of house mice from Heligoland to look into insights of their origin from possible signatures of introgression and to asses if adaptive potential was a leading force on the morphological distinctness of these mice.

Despite the expected importance of introgression in the evolution of invasive species, there have been no comprehensive studies of patterns of introgression across their genomes. This chapter aims at assessing the signatures of introgression in the genome of the invasive house mouse *M. m. helgolandicus* using the high throughput genome sequencing data and the ancestral segments detection software implemented in Hapmix. Here, I report on the patterns of introgression across the genome of the

house mouse of Heligoland based on the whole genome sequences of three individual mice. My goals were to identify the size and location of regions of high levels of introgression from the subspecies inhabiting Western Europe. Furthermore, to use functional annotations of genes spanning these regions to get insights into the biological processes associated with patterns of introgression.

4.2 Materials and methods

4.2.1 Whole genome sequencing

4.2.1.1 DNA Extraction

The genomic DNA of three individual mice from Heligoland was extracted from liver tissue using the salt extraction protocol described in section 2.2.2. The quality of the DNA was checked with agarose gel and the concentration of DNA was measured with NanoDrop spectrophotometer (Thermo Inc. USA). The DNA samples were sent to the Cologne Center for Genomics at the university of Cologne for whole genome sequencing using Illumina HiSeq2000 technology.

4.2.1.2 DNA Library construction and genome sequencing

DNA library preparation was carried out by the sequencing center according to the standard Illumina TruSeq protocol for sequencing on HiSeq 2000 (Illumina Inc., San Diego, USA). Consequently, two paired-end libraries with insert size of $\tilde{2}30$ bp were generated for deep sequencing of each genome using HiSeq 2000 (Illumina Inc.). The constructed DNA libraries for the 3 samples were tagged and then pooled and sequenced with a paired end cluster generation kit on 6 Illumina HiSeq2000 (2x100bp) lanes, resulted in 70-80Gb of filtered data for each sample.

4.2.2 Sequence analysis

4.2.2.1 Trimming of the reads

The paired-end reads obtained from the sequencing step in FASTQ format were subjected to trimming step using Trimmomatic version 0.30. The trimming step consists of trimming low quality bases and removal of adapters and other illumina-specific sequences and dropping of reads below 60 bases long (Bolger et al., 2014). The command used for this step is provided in the Appendix.

4.2.2.2 Indexing of the reference genome

Prior to mapping of the reads, the reference genome (NCBI build 37/mm9) was downloaded in fasta format from the UCSC genome browser (Downloads) utility for the mouse genome. The downloaded genome was indexed using the Burrows Wheeler (bwa) version 0.6.2-r126 (Li and Durbin, 2010).

4.2.2.3 Sequence mapping and alignment to the reference genome

Paired end reads were mapped to the indexed mouse reference genome by sequence alignment (aln) using the Burrows Wheeler Alginer (bwa) version 0.6.2-r126 (Li and Durbin, 2010). The mapped reads produced were in Sequence Alignment/Map format (SAM), which were then subjected to Samtools utility functions view, sort and index respectively to produce the Binary Sequence/Map format (BAM). PCR duplicates were removed using the rmdup function provided by Samtools utility. The summary statistics for each of the genomes sequenced here were obtained using Qualimap (García-Alcalde et al., 2012). Here the mapped and aligned reads to the reference sequence were implemented under the command line interface in a BAM file format to summarize the information on the number of mapped reads, number of paired end reads, genome coverage, insert size, AGCT% content, and mapping

quality. The Integrative Genomics Viewer (IGV) version 2.3 was used to visualize the aligned reads for each of the sequenced genomes (Thorvaldsdóttir et al., 2013)

4.2.2.4 SNP calling and detection

The mpileup function of samtools version 0.1.18 was used to detect single nucleotide polymorphisms(SNPs) in relevance to the reference genome (NCBI build 37/mm9) (Li, 2011; Li et al., 2009) along with the bcftools view function version 0.1.17-dev (Li, 2011). The vcftools version 0.1.9.0 were used to generate the variant call format file which is a representation of the respective sequence variations of the analyzed sequences (Danecek et al., 2011). The details of SNP calling steps are summarized in Table 4.1.

Table 4.1: Software used for SNPs calling and detection.

Step No.	Software	Function	Output	References
1	Mpileup	Samtools	Binary Alignment Format(BAM)	(Li, 2011; Li et al., 2009)
2	View	Bcftools	Binary Call Format (BCF)	(Li, 2011)
3	VarFilter	Vcftools	Variant Call Format (VCF)	(Danecek et al., 2011)

4.2.2.5 Identification of SNPs and analysis of variants

The annotation software snpEff was used to annotate the variants and their effects on the genome (such as amino acid changes) (Cingolani et al., 2012). The software takes the variant call format as input and analyses the variants, annotates them and calculates the effects they produce on known genes. For the annotation analysis the software requires genome database and for that purpose, the mouse reference genome NCBIM37.64 was used in an attempt to annotate the variants across the genomes from Heligoland.

4.2.3 Introgression analysis

4.2.3.1 Hapmix-Inference of local ancestry in admixed populations

To characterize patterns of introgression across the genomes of the three house mice from Heligoland, the hidden Markov model approach implemented in Hapmix software was used. Hapmix (Price et al., 2009) is used mainly to infer the ancestral state of a given admixed individual for all possible chromosomal segments in respect to two hypothetical potential source populations. Hapmix treats the two hypothesized source populations as totally phased and combines a phasing algorithm that allows the calculation of the average inferences about ancestry over all the possible phased haplotypes. Hence, it compares the unphased data from putatively admixed individuals to the phased data from the reference ancestral populations (Price et al., 2009).

4.2.3.2 Reference data for introgression analysis

Given that, the mouse genome diversity array was annotated according to the mouse dbSNP128, I used the functional annotation of genetic variants implemented in ANNOVAR (Wang et al., 2010) combined with the dbSNP128. The annotated variants were used to detect overlapping regions with the mouse genome array and hence used for introgression analysis.

The data for the two reference populations was obtained from (Staubach et al., 2012). The reference data of *M. m. domesticus* origin was represented by a German population and of *M. m. musculus* origin by a population from Kazakhstan. Each population was represented by 22 autosomal chromosome samples from 11 unrelated wild caught individuals (Staubach et al., 2012). The phased data from these populations was used as a potential source populations for the putatively admixed individuals from Heligoland.

4.2.3.3 Patterns of introgression

Hapmix uses the Hidden Markov Model (HMM) to model linkage disequilibrium within populations and it allows for miscopying of ancestry fragments from the non-ancestral population and the use of unphased data. In addition, it is capable of depicting older ancestral fragments which are much shorter than the recent events. Here, the patterns of introgression were depicted using Hapmix HAPLOID mode. The parameters used were 100 generations since admixture and miscopying value of 0.0005. These values have been found to detect smaller introgressed haplotypes with reasonable power. The minimum per SNP certainty threshold to call a SNP introgressed was 0.9 and the recombination parameters used as described in Staubach et al. (2012).

The haploid mode estimates the likelihood that a haplotypic region in an admixed individual from Heligoland is statistically correlated to Kazakhstan population or to the German. Introgression was explained by the inferred probabilities of an individual to have 1 or 0 copies from the first population (Kazakhstan), or 9 copy for unknown ancestry. Hence, if the ancestry of a chromosomal region was assigned to the *M. m. musculus* subspecies (Kazakhstan population), this region was considered introgressed. The inferred probabilities of introgression at each locus was merged with the SNP input file used for running Hapmix. The new merged file was subjected to an R script to detect the boundaries of introgressed haplotypes, their length and frequency from the number of introgressed haplotypes within a given region.

4.2.3.4 Data visualization and GO of introgressed regions

The patterns of introgression file resulted from the previous analysis was loaded as custom track on the UCSC genome browser. The Genome Graphs utility of the browser was used to visualize the genomic regions affected by introgression and to retrieve gene lists overlapping with the respective regions across chromosomes (Kent

et al., 2002). In addition the Tables function (Karolchik et al., 2004) was used to calculate fractions of genome affected. Gene lists were then analyzed with the online tool GOrilla available on (http://cbl-gorilla.cs.technion.ac.il/). The tool was used to detect enrichment terms of genes that appear densely at the top of the ranked list of genes using *Mus musculus* reference genome. Here, I focused on ontology associated with "Biological process" with a significance threshold at P-value < 0.001 (Eden et al., 2007, 2009).

4.3 Results

4.3.1 Whole genome sequence analysis

The depth of coverage for each genome was calculated using Qualimap and the mean coverage for the 3 genomes is 10X. The parameters used for sequence alignment and mapping and the calculations of whole genome statistic are provided in the Appendix. The depth of coverage for each of the 3 genomes is illustrated in supplementary Figure 1 (**I**, **II**, & **III**) respectively. In addition supplementary Figure 2, shows a snapshot of the mapped reads to the reference genome.

4.3.2 Detection of SNPs

The calling of SNPs using the mpileup function of Samtools resulted in a total of 7,974,665 million SNPs from the three whole genome sequenced individuals. Some of these variants are shown in supplementary Figure 3. To get use of the reference population data, I used the SNPs assigned by Staubach et al. (2012) to find overlapping SNPs with those detected from the whole genome data in this study. To do that, I first annotated the 7,974,665 million variants with ANNOVAR using the filter based annotation of genetic variants and mouse snp128 database (Wang et al.,

2010). That resulted in a total of 4,078,197 annotated variants written to a new file.

The annotated variants were then subjected to an awk script along with SNP data for the reference data and a total of 121,819 overlapping variants was produced and used with Hapmix software to depict possible ancestry for the chromosomal fragments. The total number of variants used by Hapmix is dependent on the number of variants in the phased reference population data, hence Hapmix depicts the ancestral fragments in the admixed individuals in relevance to the reference populations. To say that, the total number of variants assigned here covered 471,271 variants.

4.3.3 Genome annotations

The annotation results obtained from the analysis of variant call format (VCF) file for the the 3 sequenced genomes from Heligoland show large number of variants across the genomes. The highest percentage of variants with effects on the genome is obviously found in introns and secondly in intergenic regions. The variants were assigned to different genomic types and regions and their details are provided in Table 4.2 & Figure 4.2

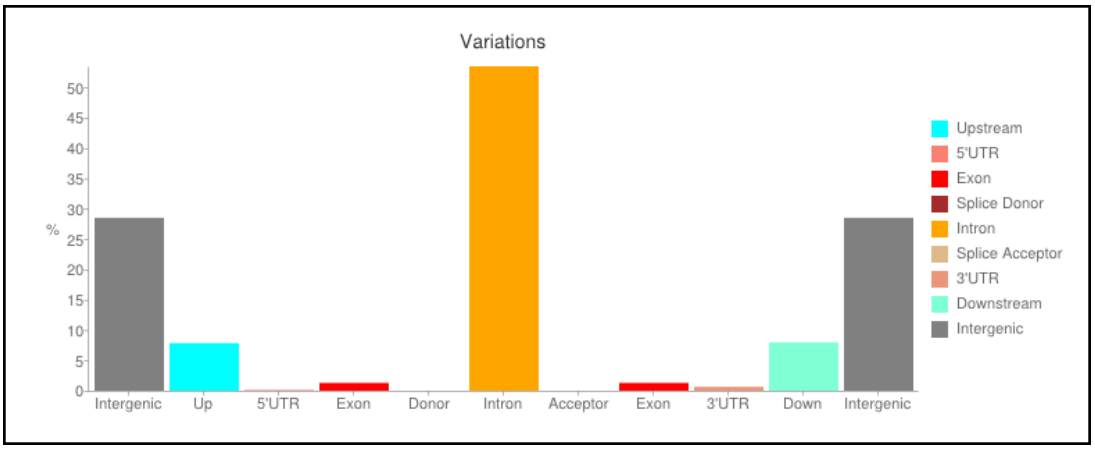


Figure 4.2: Genome annotation chart obtained from the analysis of the VCF file for the 3 genomes from Heligoland using snpEff software and NCBIM37.64 database.

Table 4.2: Genome annotation table obtained from the analysis of the variant call format (VCF) file for the 3 genomes from Heligoland using snpEff software and NCBIM37.64 database.

Number of effects by functional class

Type	Count	Percent
MISSENSE	42648	34.809%
NONSENSE	473	0.386%
SILENT	79399	64.805%

Number of effects by type

Type	Count	Percent
DOWNSTREAM	1348803	7.966%
EXON	94963	0.561%
INTERGENIC	4824422	28.493%
INTRAGENIC	631	0.004%
INTRON	9054940	53.478%
NONE	27836	0.164%
NON_SYNONYMOUS_CODING	42488	0.251%
NON_SYNONYMOUS_START	10	0%
SPLICE_SITE_ACCEPTOR	241	0.001%
SPLICE_SITE_DONOR	320	0.002%
START_GAINED	3113	0.018%
START_LOST	80	0%
STOP_GAINED	473	0.003%
STOP_LOST	70	0%
SYNONYMOUS_CODING	79344	0.469%
SYNONYMOUS_STOP	55	0%
UPSTREAM	1328116	7.844%
UTR_3_PRIME	107682	0.636%
UTR_5_PRIME	18497	0.109%

Number of effects by region

Type	Count	Percent
DOWNSTREAM	1348803	7.966%
EXON	217483	1.284%
INTERGENIC	4824422	28.493%
INTRON	9054940	53.478%
NONE	28467	0.168%
SPLICE_SITE_ACCEPTOR	241	0.001%
SPLICE_SITE_DONOR	320	0.002%
UPSTREAM	1328116	7.844%
UTR_3_PRIME	107682	0.636%
UTR_5_PRIME	21610	0.128%

4.3.4 Patterns of introgression

Results from Hapmix on chromosomal ancestry fragments were mainly obtained as a likelihood of 1 or 0 copy from the *M. m. musculus* from Kazakhstan or 9 copy of an unknown ancestry. The boundaries of introgressed haplotypes, length and frequency were detected using an R script. The SNP data information used as input was merged with the ancestry file output and so the sum of introgressed haplotypes

and size of regions that had at least one introgressed haplotype were considered introgressed and the results across the chromosomes are illustrated in Figure 4.3 and their details are summarized in supplementary Table 11.

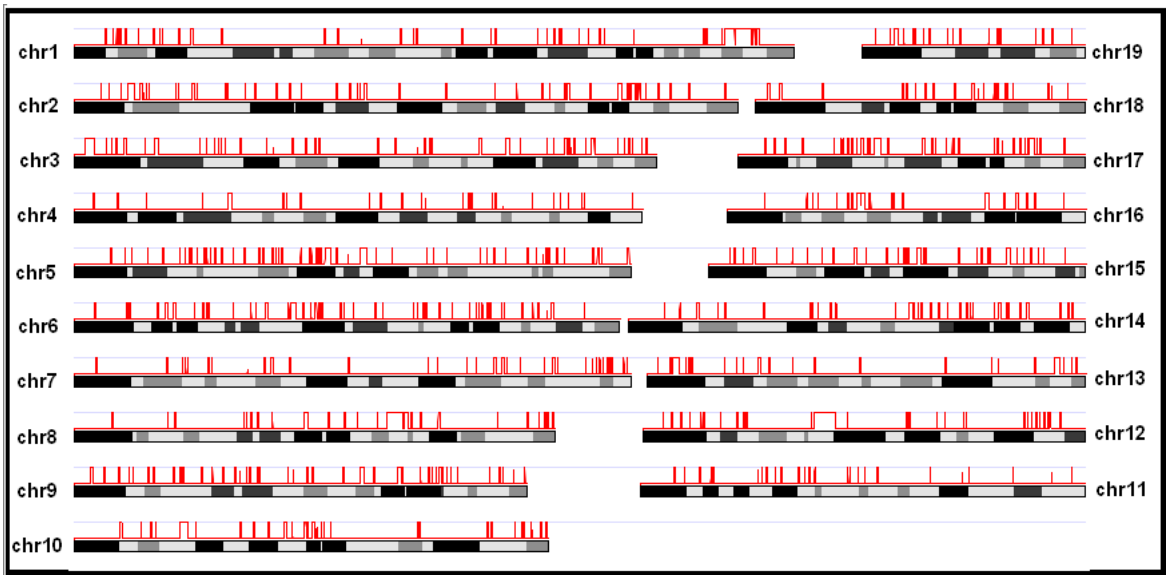


Figure 4.3: Patterns of introgression into the genome of *M. m. helgolandicus* from Heligoland visualised with the Genome Graphs utility of the UCSC Genome Browser (Kent et al., 2002). The elevated red bars are introgressed regions from *M. m. musculus* subspecies.

The genomic introgression into *M. m. helgolandicus* from *M. m. musculus* have on average large introgressed regions, with approximately 5MB across the genome (Table 4.3). Chromosome 17 shows the largest fraction of genome introgression across the genome. This pattern is expected given that this chromosome bears reduced local recombination due to the presence of several inversions most likey due to the t-haplotype. This finding corporate previous finding on the putatively pure *M. m. musculus* and *M. m. domesticus* populations with also high fraction of genome introgression on chromosome 17 (Staubach et al., 2012). However,this is not yet confirmed for the Heligoland mice.

GO analysis of genes covered by introgressed regions showed some enrichment terms due to the inclusion of gene clusters. Of interest, are the highly significant cluster of genes concerning the regulation of multi-cellular organismal process, regulation

Table 4.3: Genome regions affected by introgression across the genome of Heligoland house mouse

Introgression from *M. m. musculus*	
Hapmix (N)	694
average size (bp)	245,092
maximum size (bp)	5,796,932
fraction of genome (bp)	170,093,716
fraction of genome (%)	6.49%

of multi-cellular organismal development, regulation of interleukin-6 production and the regulation of organ formation p-values < 0.001. In addition, some gene clusters are correlated with the positive regulation of leukocytes, lymphocytes and T-cell differentiation due to various immune system responses. Besides that, some gene clustering is linked to the sensory perception, detection of stimulus. Go terms concerned with face morphogenesis and anatomical structure morphogenesis could be related to different processes where for example, the process in which the anatomical structures of the face are generated and organized. The GO terms for genes covered by introgression and their associated biological functions are illustrated in Figure 4.4 and their details are in Table 4.4.

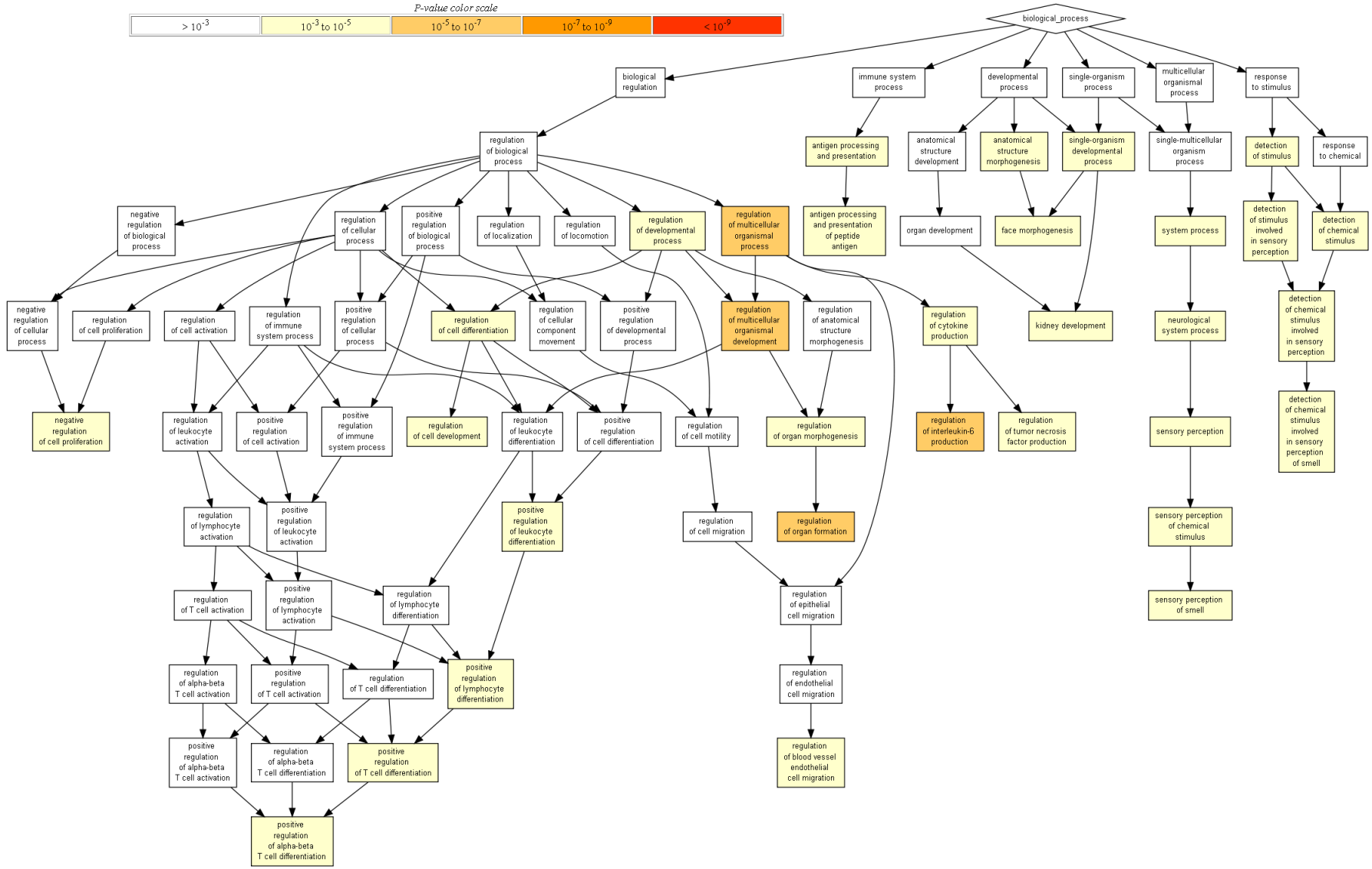


Figure 4.4: GO terms for biological processes of genes covered by introgression. Colors represent the P-value scale.

Table 4.4: Output of GOrilla showing Gene Ontology term enrichment for the biological process in the gene list overlapping the introgressed regions into *M. m. helgolandicus*.

GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)
GO:0051239	regulation of multicellular organismal process	1.77E-6	1.05E-2	4.46 (1952,200,35,16)
GO:2000026	regulation of multicellular organismal development	2.8E-6	8.28E-3	6.11 (1952,137,28,12)
GO:0003156	regulation of organ formation	6.3E-6	1.24E-2	488.00 (1952,2,4,2)
GO:0032675	regulation of interleukin-6 production	9.89E-6	1.46E-2	20.77 (1952,10,47,5)
GO:0001817	regulation of cytokine production	1.54E-5	1.83E-2	7.95 (1952,47,47,9)
GO:0007600	sensory perception	1.79E-5	1.77E-2	1.41 (1952,147,924,98)
GO:0050793	regulation of developmental process	2.52E-5	2.13E-2	5.01 (1952,167,28,12)
GO:0009593	detection of chemical stimulus	2.9E-5	2.14E-2	1.44 (1952,124,917,84)
GO:0048002	antigen processing and presentation of peptide antigen	3.36E-5	2.2E-2	2.53 (1952,20,656,17)
GO:0045595	regulation of cell differentiation	4.22E-5	2.49E-2	6.06 (1952,115,28,10)
GO:0050911	detection of chemical stimulus involved in sensory perception of smell	5.31E-5	2.86E-2	1.44 (1952,120,917,81)
GO:0050907	detection of chemical stimulus involved in sensory perception	5.31E-5	2.62E-2	1.44 (1952,120,917,81)
GO:0008285	negative regulation of cell proliferation	9.92E-5	4.51E-2	7.53 (1952,61,34,8)
GO:0003008	system process	1.08E-4	4.55E-2	1.32 (1952,188,952,121)
GO:0009653	anatomical structure morphogenesis	1.14E-4	4.51E-2	1.89 (1952,99,427,41)
GO:0007608	sensory perception of smell	1.54E-4	5.69E-2	1.41 (1952,124,917,82)
GO:1902107	positive regulation of leukocyte differentiation	1.72E-4	6E-2	6.51 (1952,12,175,7)
GO:0045621	positive regulation of lymphocyte differentiation	2.03E-4	6.66E-2	7.44 (1952,9,175,6)
GO:2000027	regulation of organ morphogenesis	2.14E-4	6.66E-2	162.67 (1952,6,4,2)
GO:0007606	sensory perception of chemical stimulus	2.41E-4	7.13E-2	1.40 (1952,125,917,82)
GO:0050906	detection of stimulus involved in sensory perception	3E-4	8.43E-2	1.38 (1952,131,931,86)
GO:0060325	face morphogenesis	3.15E-4	8.46E-2	78.08 (1952,2,25,2)
GO:0001822	kidney development	3.15E-4	8.1E-2	139.43 (1952,7,4,2)
GO:0032680	regulation of tumor necrosis factor production	3.77E-4	9.28E-2	16.61 (1952,10,47,4)
GO:0019882	antigen processing and presentation	4.65E-4	1.1E-1	2.17 (1952,26,656,19)
GO:0045582	positive regulation of T cell differentiation	4.78E-4	1.09E-1	7.97 (1952,7,175,5)
GO:0046638	positive regulation of alpha-beta T cell differentiation	5.41E-4	1.18E-1	15.02 (1952,3,130,3)
GO:0051606	detection of stimulus	6.3E-4	1.33E-1	1.35 (1952,140,931,90)
GO:0050877	neurological system process	6.52E-4	1.33E-1	1.31 (1952,170,937,107)
GO:0060284	regulation of cell development	8.81E-4	1.73E-1	6.62 (1952,59,35,7)
GO:0044767	single-organism developmental process	9.47E-4	1.8E-1	1.68 (1952,319,171,47)
GO:0043535	regulation of blood vessel endothelial cell migration	9.54E-4	1.76E-1	7.28 (1952,4,268,4)

Enrichment=(b/n)/(B/N), where N, is the total number of genes; B, is the total number of genes associated with a specific GO term; n, is the number of genes in the top of the user's input list or in the target set when appropriate; b, is the number of genes in the intersection. *P*-value is the enrichment *p*-value computed according to the mHG or HG model (Eden et al., 2007). FDR *q*-value is the correction of the above *p*-value for multiple testing using the Benjamini and Hochberg (1995) method (Benjamini and Hochberg, 1995). The FDR *q*-value is (*p*-value * number of GO terms)/i.

4.4 Discussion

Next generation sequencing technology has made significant development for genome sequencing of the house mouse, both in a time- and cost-effective fashion. The shift from conventional methodologies to entire genome analysis put forward great efforts in unraveling the evolutionary history at subspecies level and to some degree at different population level. The patterns of introgression observed in the genome of *M. m. helgolandicus* can be explained by two differnet scenarios. First, the introgressed regions from *M. m. musculus* are the results of an ongoing gene flow from some populations of *M. m. musculus* subspecies. Alternatively, the introgressed regions from *M. m. musculus* are relict of ancestral genomes which were mainly transported from the hybrid zone since first colonization of the island.

The results from this chapter corroborate previous findings on wild populations of the house mouse, that the genome of natural populations is shaped by introgression from sister-species (Staubach et al., 2012). However, while mainland populations show 3.5% introgression, the Heligoland population shows at least 6.5% and possibly more, when one takes into account that only 3 individuals were analysed so far.

The interpretation of observed patterns of introgression by simulations for wild populations is of great importance for such studies where incompatibilities with a neutral model could be proved or rejected. For example Staubach et al. (2012) conducted such analysis on simulated data using Hapmix software (Price et al., 2009) which was used for natural population data analysis and revealed that the frequency and size distribution of the introgressed fragments observed could not be explained by a neutral model of introgression. The population history of the Heligoland mice is too short to do similar simulations, but it is possible that the introgression of alleles may also have contributed to adaptations.

The genome sequences and the population of the house mouse of Heligoland

to which they belonged carry haplotypes from the subspecies *M. m. musculus*. Nevertheless, the picture that emerges from the molecular analysis of the nuclear genome is one where the population of Heligoland is of *M. m. domesticus* origin.

Three different scenarios could account for how such an introgression of haplotypes has come to be present in this population. One scenario is that these haplotypes were retained in the founders of this population which have been broken by recombination events. A second scenario is that they entered the population of Heligoland through gene flow from house mouse of the *M. m. musculus* subspecies. Although such gene flow cannot be detected with the current mtDNA and the analysed molecular DNA data, further sequencing of fossil material may unravel the cryptic history behind it. The third scenario that could account for the apparently introgressed haplotypes is likely that these were adaptively acquired by selection.

On a similar perspective, a study on island populations in New Zealand revealed patterns of genome intermixing in the house mouse inhabiting the different islands as a result to an ongoing hybridization among the three subspecies *M. m. domesticus*, *M. m. musculus* and *M. m. castaneus* which resulted in combinations of subspecies nuclear DNA and mtDNA (Searle et al., 2009a).

Here, I show that patterns of introgression from the genome sequences of the invasive house mouse *M. m. helgolandicus* can be reliably recovered. The assigned patterns of introgressions and the genes involved within these regions across the genome of *M. m. helgolandicus*, not only shed light on their importance, but also provide insights into population specific adaptations (Staubach et al., 2012; Teeter et al., 2007).

The assigned patterns point to a number of genomic regions and genes as candidates for positive selection in *M. m. helgolandicus*. For example, those involved in face morphogenesis and anatomical structure morphogenesis could have played a role in cranial morphology. And more likely to support the findings from chapter 3

on the distinct mandible shape of these mice. However, extended analysis is required to quantitatively confirm these findings. In addition those involved in detection of stimulus, sensory perception, regulation of immune cells (T cell) and developmental processes are likely acquired by adaptation to the novel environment on the island.

More interestingly, these findings coin with recent studies (Mallet, 2007; Nolte and Tautz, 2010; Staubach et al., 2012) pointing towards the important role of introgression in shaping the genome. More over, introgresssion as a creative force can be found not only, in shortly distant populations, but also in those isolated and hence provide the potential for diversification and the formation of new species.

I expect that further analyses of the house mouse genome as well as the genomes from fossil remains will provide further insights into the origins and early history of the invasive house mouse on Heligoland. In addition, sequencing of extra samples from Heligoland and reference populations representing wild mice from *M. m. musculus* and *M. m. domesticus* subspecies from the mainland might provide deeper insights into the level of adaptive evolution these mice had experienced and will also give insights into a better layout of their origin. It is noteworthy to mention that the use of the Mouse Genome Diversity microarray is limited to less than half million SNPs, and hence the application of admixure analysis on designed assumptions from this study should include whole genome sequence and hence large number of SNPs that could be analysed for patterns of introgression attempting to cover large regions across the admixed genome.

5 Concluding remarks

The small size of Heligoland along with the effect of isolation from the mainland, have influenced the genetic composition of the house mouse on the island. Genetic drift is suspected to have much influence on such a small population and as a sum the population of Heligoland represented by samples here exhibited a low genetic diversity from the perspective of microsatellites and a major mtDNA haplotype only found on Heligoland.

The first possible colonization of house mouse on Heligoland estimated from the mitochondrial DNA mutation frequency dates back to four centuries ago, which is in line with the documented evidence of the first humans on the island in the fifteenth century. This finding was supported by the presence of a major mtDNA haplotype specific to Heligoland, which also comprises of a distinct pattern (11-bp direct repeat) mostly observed in Western Europe and hence, explains the notion that Heligoland was colonized by house mouse from this region. Despite the commensal activity of the house mouse, the persistence of one major mtDNA haplotype on Heligoland suggests refractory to immigration and a non favorable ecological conditions for the migrants to establish a large population that could contribute significantly to the local established gene pool.

The analysis of geometric morphometrics supports the findings from previous morphological analysis on this population by the more advanced morphometric analysis of the mandible. Significant differences in the two-dimentional data of the house mouse mandible were documented here between Heligoland and continental populations. The house mouse of Heligoland features an elongated mandible which suggests a carnivorous or insectivorous adaptation to the prevalent diet on the island.

More interestingly, the finding from this study mainly the distinct mandible size and shape observed from geometric morphometrics analysis, confirm previous

assumptions that evolution of mammals is accelerated on small islands and that the "island rule" is evident for such invasive mammals. These fascinating results shed light on the importance of morphological adaptation to environmental changes, mostly influenced by the exploitation of food resources available on the newly colonized niche.

Although the molecular data and colonization history of the house mouse on Heligoland assign these to *M. m. domesticus*, the results from introgression analysis show high levels of genome intermixing from *M. m. musculus* subspecies represented by (6.49%). These patterns of introgression might explain part of the complex adaptation processes that could have shaped the genome of the house mouse on Heligoland.

Application of recent advanced statistical analysis on wide genome data for sequences from Heligoland and wild populations representing *M. m. musculus* and *M. m. domesticus* subspecies might unravel regions of introgresion not observed in this study due to the limited coverage of genomic variants which was based on a dataset from the mouse genome diversity array. Genome re-sequencing data for wild mouse populations will soon be available, such that full genome comparisons will be once more complete.

References

Akey JM. 2009. Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Res.* 19:711–722.

Arnold ML. 1992. Natural hybridization as an evolutionary process. *Annu. Rev. Ecol. Syst.* 23:237–261.

Arnold ML, Bulger MR, Burke JM, Hempel AL, Williams JH. 1999. Natural hybridization: how low can you go and still be important? *Ecology.* 80:371–381.

Atchley WR, Hall BK. 1991. A model for development and evolution of complex morphological structures. *Biol. Rev.* 66:101–157.

Auffray JC, Alibert P, Latieule C. 1996. Relative warps analysis of skull shape across the hybrid zone in the house mouse *Mus musculus* in denmark. *J Zool Lond.* 240:441–455.

Auffray JC, Vanlerberghe F, Britton-Davidian J. 1990. The house mouse progression in eurasia: a palaeontological and archaeozoological approach. *Bio Jour Lin Soc.* 41:13–25.

Avice JC. 2000. Phylogeography: The history and formation of species. Cambridge MA: Harvard University press.

Ballard JWO, Whitlock MC. 2004. The incomplete natural history of mitochondria. *Molecular Ecology.* 13:729–744.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological).* 57:289–300.

Berry RJ. 1996. Small mammal differentiation on islands. *Phil. Trans. R. Soc. Lond. B.* 351:753–764.

Berry RJ. 1998. Evolution of small mammals. In: Grant PR, editor, Evolution on islands. Oxford university press, pp. 35–50.

Berry RJ, Bronson FH. 1992. Life history and bioeconomy of the house mouse. *Biological reviews of the Cambridge Philosophical Society.* 67:519–550.

Berry RJ, Guthbert A, Peters J. 1982. Colonization by house mice: an experiment. *J. Zool.* 198:329–336.

Berry RJ, Peters J, Van Aarde RJ. 1978. Sub-antarctic house mice: colonization, survival and selection. *J. Zool.* 184:127–141.

Berry RJ, Scriven PN. 2005. The house mouse: a model and motor for evolutionary understanding. *Bio Jour of the Linne Socie.* 84:335–347.

Bibb MJ, Van Etten RA, Write CT, Walberg MW, Clayton DA. 1981. Sequence and gene organization of mouse mitochondrial dna. *Cell.* 26:167–180.

Boell L, Pallares LF, Brodski C, et al. (11 co-authors). 2013. Exploring the effects of gene dosage on mandible shape in mice as a model for studying the genetic basis of natural variation. *Dev Genes Evol.* 223:279–287.

Boell L, Tautz D. 2011. Micro-evolutionary divergence patterns of mandible shapes in wild house mouse (mus musculus) populations. *BMC Evolutionary Biology.* 11.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: Aflexible trimmer for illumina sequence data. *Bioinformatics.* .

Bonhomme F, Orth A, Cucchi T, Rajabi-Maham H, Catalan J, Boursot P, Auffray JC, Britton-Davidian J. 2010. Genetic differentiation of the house mouse around

the mediterranean basin: matrilineal footprints of early and late colonization. *Proceedings of the Royal Society B: Biological Sciences*. .

Bookstein FL. 1991. Morphometric tools for landmark data: Geometry and biology. Cambridge University Press.

Boursot P, Auffray J, Britton-Davidian J, Bonhomme F. 1993. The evolution of house mice. *Ann Rev Ecol Syst*. 24:119–152.

Cavalli-Sforza LL, Edwards AWF. 1967. Phylogenetic analysis: models and estimation procedures. *American Journal of Human Genetics*. 19:233–257.

Cingolani P, Platts A, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*. 6:80–92.

Corti M, Rohlf FJ. 2001. Chromosomal speciation and phenotypic evolution in the house mouse. *Biol J Linn Soc*. 73:99–112.

Cucchi T, Vigne JD, Auffray JC. 2005. First occurrence of the house mouse (*Mus musculus domesticus* schwarz & schwarz, 1943) in the western mediterranean: A zooarchaeological revision of subfossil occurrences. *Biol Jour Linn Soc Lond*. 84:429–445.

Danecek P, Auton A, Abecasis G, et al. (13 co-authors). 2011. The variant call format and vcftools. *Bioinformatics*. 27:2156–2158.

Dieringer D, Schlötterer C. 2003. Microsatellite analyser (msa): a platform independent analysis tool for large microsatellite data sets. *Molecular Ecology Notes*. 3:167–169.

Dierschke J, Dierschke V, Hüppop K, Hüppop O, Jachmann KF. 2010. Die Vogelwelt der Insel Helgoland. Bremen: OAG Helgoland.

Din W, Anand R, Boursot P, Darviche D, Dod B, Jouvin-Marche E, Orth A, Talwar GP, Cazenave PA, Bonhomme F. 1996. Origin and radiation of the house mouse: clues from nuclear genes. *Journal of Evolutionary Biology*. 9:519–539.

Dod B, Jermiin LS, Boursot P, Chapman VH, Tonnes-Nielsen J, Bonhomme F. 1993. Counterselection on sex chromosomes in the mus musculus european hybrid zone. *Journal of Evolutionary Biology*. 6:529–546.

Dod B, Smadja C, Karn RC. 2005. Testing for selection on the androgen-binding protien in the danish mouse hybrid zone. *Biological journal of the Linnean society*. 84:447–459.

Domingues VS, Poh YP, Peterson BK, Pennings PS, Jensen JD, Hoekstra HE. 2012. Evidence of adaptation from ancestral variation in young populations of beach mice. *Evolution*. 66:3209–3223.

Drower GMF. 2002. Heligoland: the true story of German Bight and the island that Britain betrayed. Sutton Pub.

Eden E, Lipson D, Yogev S, Yakhini Z. 2007. Discovering motifs in ranked lists of dna sequences. *PLoS Comput Biol*. 3:e39.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinformatics*. 10:48.

Egan SP, Nosil P, Funk DJ. 2008. Selection and genomic differentiation during ecological speciation: isolating the contributions of host-association via a comparative genome scan of neochlamisus bebbianae leaf beetles. *Evolution*. 62:1162–1181.

Ellegren H. 2004. Microsatelliets: Simple sequences with complex evolution. *Genetics*. 5:435–445.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology*. 14:2611–2620.

Felsenstein J. 1991. Phylip (phylogeny inference package) version 3.6. distributed by the author.

Figueroa F, Kasahara M, Tichy H, Neufeld E, Ritte U, Klein J. 1987. Polymorphism of unique noncoding dna sequences in wild and laboratory mice. *Genetics*. 117:101–108.

Forejt J. 1996. Hybrid sterility in the mouse. *Phil Trans R Soc Lond B*. 12:412–417.

Förster DW, Gündüz I, Nunes AC, Gabriel S, Ramalhino MG, Mathias ML, Britton-Davidian J, Searle JB. 2009. Molecular insights into the colonization and chromosomal diversification of madeiran house mice. *Molecular Ecology*. 18:4477–4494.

Foster JB. 1964. Evolution of mammals on islands. *Nature*. 202:234–235.

Frazer KA, Eskin E, Kang HM, et al. (16 co-authors). 2007. A sequence-based variation map of 827 million snps in inbred mouse strains. *Nature*. 448:1050–1055.

García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, Dopazo J, Meyer TF, Conesa A. 2012. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics*. 28:2678–2679.

Guénet JL, Bonhomme F. 2003. Wild mice: an ever-increasing contribution to a popular mammalian model. *Trends Genet*. 19:24–31.

Gündüz I, Auffray JC, Britton-Davidian J J Catalan, Ganem G, Ramalhinho MG, Mathias ML, Searle JB. 2001. Molecular studies on the colonization of the madeiran archipelago by house mice. *Molecular Ecology*. 10:2023–2029.

Gündüz I, Rambau RV, Tez C, Searle JB. 2005. Mitochondrial dna variation in the western house mouse (*Mus musculus domesticus*) close to its site of origin: studies in turkey. *Biological Journal of the Linnean Society*. 84:473–485.

Hardouin EA, Chapuis JL, Stevans MI, Van Vuuren JB, Quillfeldt P, Scavetta RJ, Teschke M, Tautz D. 2010. House mouse colonization patterns on the sub-antarctic kerguelen archipelago suggest singular primary invasions and resilience against re-invasion. *BMC Evolutionary Biology*. 10:1471–2148.

Hardouin EA, Tautz D. 2013. Increased mitochondrial mutation frequency after an island colonization: positive selection or accumuliation of slightly deleterious mutations? *Biol Lett*. 9:20121123.

Harr B, Weiss S, David JR, Brem G, Schlötterer C. 1998. A microsatellite-based multilocus phylogeny of the drosophila melanogaster species complex. *Current Biology*. 8:1183–1186.

Heany LR. 1978. Island area and body size of insular mammals: evidence from the tri-colored squirrel *Callosciurus prevosti* of southeast asia. *Evolution*. 32:29–44.

Hubisz MJ, Falush D, Stephens M, Pritchard JK. 2009. Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*. 9:1322–1332.

Ihle S, Ravaoarimanana I, Thomas M, Tautz D. 2006. Tracing signatures of selective sweeps in natural populations of the house mouse. *Molecular Biology and Evolution*. 23:790–797.

Jakobsson M, Rosenberg NA. 2007. Clumpp:a cluster matching permutation program for dealing with label switching multimodality in analysis of population structure. *Bioinformatics*. 23:1801–1806.

Joly S, Stevens MI, van Vuuran BJ. 2007. Haplotype networks can be misleading in the presence of missing data. *Systematic Biology*. 56119:857–862.

Jones EP, Jensen JK, Magnussen E, Gregersen N, Hansen H, Searle JB. 2011. The molecular characterization of the charismatic faroe house mouse. *Biol J Linn Soc*. 102:471–482.

Jones EP, Skirnisson K, McGovern TH, Gilbert MTP, Willerslev E, Searle JB. 2012. Fellow travellers: a concordance of colonization patterns between mice and men in the north atlantic region. *MC Evolutionary Biology*. 12:1–8.

Jones EP, van der Kooji J, Solheim R, Searle JB. 2010. Norwegian house mice (*Mus musculus musculus/ domesticus*): distributions, routes of colonization and patterns of hybridization. *Molecular Ecology*. 19:5252–5264.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The ucsc table browser data retrieval tool. *Nucleic Acids Research*. 32:D493–D496.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at ucsc. *Genome Research*. 12:996–1006.

Kim KM, Park JH, Bhattacharya D, Yoon HS. 2014. Applications of next generation sequencing to unravelling the evolutionary history of algae. *International Journal of Systematic and Evolutionary Microbiology*. 64:333–345.

Kirkness EF, C NP. 2010. Whole genome sequencing. In: Barnes MR, Breen G,

editors, Genetic variation, Humana Press, volume 628 of *Methods in Molecular Biology*, pp. 215–226.

Klingenberg CP. 2010. Evolution and development of shape: integrating quantitative approaches. *Nat Rev Genet*. 11:623–635.

Klingenberg CP. 2011. Morphoj: An integrated software package for geometric morphometrics. *Mol Ecol Resour*. 11:353–357.

Klingenberg CP, Leamy LJ, Routman EJ, Cheverud JM. 2001. Genetic architecture of mandible shape in mice: Effects of quantitative trait loci analyzed by geometric morphometrics. *Genetics*. 157:785–802.

Klingenberg CP, McIntyre GS. 1998. Geometric morphometrics of developmental instability: analyzing patterns of fluctuating asymmetry with procrustes methods. *Evolution*. 52:1363–1375.

Kraft R. 1985. Morphological charcteristics and distribution of the house mouse *Mus musculus musculus L.*, 1758, and *Mus musculus musculus* rutty, 1772 (rodentia, muridae) in bavaria. *Säugetierkd. Mitt.* 32:1–12.

Lanneluc I, Desmarais E, Boursot P, Dod B, Bonhomme F. 2004. Characterization of a centromeric marker on mouse chromosome 11 and its introgression in a domesticus/musculus hybrid zone. *Mammalian Genome*. 15:924–934.

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*. 8:46–62.

Li H. 2011. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 27:2987–2993.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with burrowsś wheeler transform. *Bioinformatics*. 26:589–595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. 2009. The squence alignment/map format and samtools. *Bioinformatics*. 25:2078–2079.

Linnenbrink M, Wang J, Hardouin EA, Künzel S, Metzler D, Baines JF. 2013. The role of biogeography in shaping diversity of the intestinal microbiota in house mice. *Molecular Ecology*. 22:1904–1916.

Lomolino MV. 1985. Body size of mammals on islands: The island rule reexamined. *American Nat.* 125:310–316.

Lomolino MV. 2005. Body size evolution in insular vertebrates: Generality of the island rule. *Journal of Biogeography*. 32:1683–1699.

Losos JB, Jackman TR, Larson A, Queiroz Kd, Rodríguez-Schettino L. 1998. Contingency and determinism in replicated adaptive radiations of island lizards. *Science*. 279:2115–2118.

Losos JB, Ricklefs RE. 2009. Adaptation and diversification on islands. *Nature*. 457:830–836.

Macholán M, Munclinger P, Šugerková M, Dufková P, Bímová B, Boíová E, Zima J, Piálek J. 2007. Genetic analysis of autosomal and x-linked markers across a mouse hybrid zone. *Evolution*. 61:746–771.

Mallet J. 2005. Hybridization as an invasion of the genome. *TRENDS in Ecology and Evolution*. 20:229–237.

Mallet J. 2007. Hybrid speciation. *Nature*. 446:279–283.

Marshall JT, Sage RD. 1981. Taxonomy of the house mouse. *Symp Zool Soc Lon.* 47:15–25.

Mayr E. 1942. Systematics and the Origin of Species, from the Viewpoint of a Zoologist. Harvard University Press.

Mayr E. 1967. The challenge of island faunas. *Aust Nat Hist.* 15:369–374.

Metzker ML. 2010. Sequencing technologies-the next geration. *Nature Reviews Genetics.* 11:31–46.

Millien V. 2011. Mammals evolve faster on small islands. *Evolution.* 65:1935–1944.

Minoche AE, Dohm JC, Himmelbaue H. 2011. Evaluation of genomic high-throughput sequencing data generated on illumina hiseq and genome analyzer systems. *Genome Biology.* 12:1–15.

Moriwaki K, Miyashita N, Suzuki H, Kurihara Y, Yonekawa H. 1986. Genetic features of major geographical isolates of mus musculus. *Curr. Top. Microbiol. Immunol.* 127:55–61.

Mullen LM, Vignieri SN, Gore JA, Hoekstra HE. 2009. Adaptive basis of geographic variation: genetic, phenotypic and environmental differences among beach mouse populations. *Proceedings of the Royal Society B: Biological Sciences.* 276:3809–3818.

Munclinger P, Boursot P, Dod B. 2003. B1 insersions as an easy markers for mouse population studies. *Mammalian Genome.* 14:359–366.

Munclinger P, Boíková E, M, Piálek J, Macholán M. 2002. Genetic variation in house mice (*Mus*, muridae, rodentia) from the czech and slovac republics. *Folia Zoo.* 51:81–92.

Nielsen R, Beaumont MA. 2009. Statistical inferences in phylogeography. *Molecular Ecology*. 18:1034–1047.

Nolte AW, Tautz D. 2010. Understanding the onset of hybrid speciation. *Trends in genetics*. 26:45–58.

Nosil P, Funk DJ, Ortiz-Barrientos D. 2009. Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*. 18:375–402.

Nunome M, Ishimori C, Aplin KP, Tsuchiya K, Yonekawa H, Moriwaki K, Suzuki H. 2010. Detection of recombinant haplotypes in wild mice (*Mus musculus*) provides new insights into the origin of japanese mice. *Molecular Ecology*. 19:2474–2489.

Ogden R, Thorpe RS. 2002. Molecular evidence for ecological speciation in tropical habitats. *PNAS*. 99:13612–13615.

Oleksyk TK, Smith MW, OḂrien SJ. 2010. Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 365:185–205.

Olson JS, Shadle R. 1991. Historical Dictionary of European Imperialism. Greenwood Press.

Orth A, Lyapunova E, Kandaurov A, Boissinot S, Boursot P, Vorontsov N, Bonhomme F. 1996. Polytypic species *Mus musculus* in transcaucasia. *Biologies*. 319:435–441.

Payseur BA, Krenz JG, Nachman MW. 2004. Differential patterns of introgression across the x chromosome in a hybrid zone between two species of house mice. *Evolution*. 58:2064–2078.

Payseur BA, Nachman MW. 2005. The genomics of speciation: investigating the molecular correlates of x chromosome introgression across the hybrid zone between

Mus domesticus and *Mus musculus*. *Biological journal of the Linnean society*. 84:523–534.

Peltonen A, Hanski I. 1991. Patterns of island occupancy explained by colonization and extinction rates in shrews. *Ecology*. 72:1698–1708.

Pergams ORW, Ashley MV. 2001. Microevolution in island rodents. *Genetica*. 112-113:245–256.

Pocock MJO, Hauffe HC, Searle JB. 2005. Dispersal in house mice. *Biol. jour. of the Lin. soc*. 84:565–583.

Prager EM, Sage RD, Gyllensten U, Thomas WK, Hübner R, Jones CS, Noble L, Searle JB, Wilson AC. 1993. Mitochondrial dna sequence diversity and the colonisation of scandinavia by house mice from east holstein. *Biological journal of the Linnean society*. 30:85–122.

Prager EM, Tichy H, Sage RD. 1996. Mitochondrial dna sequence variation in the eastern house mouse, mus musculus: Comparison with other house mice and report of a 75-bp tandem repeat. *Genetics*. 143:427–446.

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski TH I Beaty, Mathias R, D R, S M. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PloS Genetics*. 5:e1000519.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics*. 155:945–959.

Rajabi-Maham H, Orth A, Bonhomme F. 2008. Phylogeography and postglacial expansion of *Mus musculus domesticus* inferred from mitochondrial dna coalescent, from iran to europe. *Molecular Ecology*. 17:627–641.

Raufaste N, Orth A, Belkhir K, Senet D, Smadja C, Baird SJE, Bonhomme F, Dod B, Boursot P. 2005. Inferences of selection and migration in the danish house mouse hybrid zone. *Biological Journal of the Linnean Society*. 84:593–616.

Reichstein H, Vauk G. 1968. Beitrag zur kenntnis der helgoländer hausmaus, (*Mus musculus helgolandicus* zimmmermann), 1953. *Zoologischer Anzeiger, Supplementband, 31 (zugl: Verhandlungen der Deutschen Zoologischen Gesellschaft, [61]):, 1 Abb, 3 Tab, Leipzig*. pp. 386–394.

Renaud S, Alibert P, Auffray JC. 2009. Mandible shape in hybrid mice. *Naturwissenschaften*. 96:1043–1050.

Renaud S, Auffray JC. 2010. Adaptation and plasticity in insular evolution of the house mouse mandible. *J Zool Evol Res*. 48:138–150.

Renaud S, Auffray JC, de la Porte S. 2010. Epigenetic effects on the mouse mandible: common features and discrepancies in remodeling due to muscular dystrophy and response to food consistency. *BMC Evol Biol*. 10.

Renaud S, Chevret P, Michaux J. 2007. Morphological vs. molecular evolution: ecology and phylogeny both shape the mandible of rodents. *Zoologica Scripta*. 36:525–535.

Renaud S, Hardouin EA, Pisanu B, Chapuis JL. 2013. Invasive house mice facing a changing environment on the sub-antarctic guillou island (kerguelen archipelago). *Evol Biol*. 26:612–624.

Renaud S, Millien V. 2001. Intra- and interspecific morphological variation in the field mouse species apodemus argenteus and a. speciosus in the japanese archipelago: the role of insular isolation and biogeographic gradients. *Bio Jour Lin Soc*. 74:557–569.

Ritsema A. 2007. Heligoland, Past and Present. Lulu Enterprises Incorporated.

Rohlf FJ. 2004. tpsUtil, file utility program version 1.26 Department of Ecology and Evolution. Department of Ecology and Evolution, State University of New York at Stony Brook. Software.

Rohlf FJ. 2005. tpsDig, digitize landmarks and outlines, version 2.05. Department of Ecology and Evolution, State University of New York at Stony Brook. Software.

Rosenberg NA. 2004. Distruct:a program for the graphical display of population structure. *Mol Ecol Notes*. 4:137–138.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. 2003. Dna polymorphism analyses by the coalescent and other methods. *Bioinformatics*. 119:2496–2497.

Sabeti PC, Schaffner SF, Fry B, Lohmueler J, Varilly P, *et al.* 2006. Positive natural selection in the human lineage. *Science*. 312:1614–1620.

Sage RD, Atchley WR, Capanna E. 1993. House mice as models in systematic biology. *Systematic Biology*. 42:523–561.

Sage RD, Heyneman D, Lim KC, Wilson AC. 1986. Wormy mice in a hybrid zone. *Nature*. 324:60–63.

Sage RD, Prager EM, Tichy H, Wilson AC. 1990. Mitochondrial dna variation in house mice, *Mus domesticus* (rutty). *Biological Journal of the Linnean Society*. 41:105–123.

Savolainen V, Anstett MC, Lexer C, Hutton I, Clarkson JJ, Norup MV, Powell MP, Springate D, Salamin N, Baker WJ. 2006. Sympatric speciation in palms on an oceanic island. *Nature*. 441:210–213.

Schatz MC, Delcher AL, Salzberg SL. 2010. Assembly of large genomes using second-generation sequencing. *Genome Research*. 20:1165–1173.

Searle JB, Jamieson PM, Gündüz I, Stevans MI, Jones EP, Gemmill CE, King CM. 2009a. The diverse origins of new zealand house mice. *Pro. R. Soc. B.* 276:209–217.

Searle JB, Jamieson PM, Gündüz I, et al. (11 co-authors). 2009b. Of mice and (viking?) men: phylogeography of british and irish house mice. *Pro. R. Soc. B.* 276:201–207.

Siahsarvie R, Auffray JC, Darvish J, Rajabi-Maham H, Yu HT, Agret S, Bonhomme F, Claude J. 2012. Patterns of morphological evolution in the mandible of the house mouse *Mus musculus* (rodentia: Muridae). *Biological Journal of the Linnean Society.* 105:635–647.

Smith VR, Avenant NL, Chown SL. 2002. The diet and impact of house mice on a sub-antarctic island. *Polar Biology.* 25:703–715.

Sondaar PY. 1991. Island mammals of the past. *Sci Progr.* 75:249–264.

Spaeth C. 1990. Zur geologie der insel helgoland. *Küste.* 49:1–32.

Staubach F, Lorenc A, Messer P, Tang K, Petrov DA, Tautz D. 2012. Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse *Mus musculus*. *PloS Genetics.* 8:e1002891.

Stewart JB, Freyer C, Elson JL, Wredenberg A, Cansu Z, Trifunovic A, Larsson NG. 2008. Strong purifying selection in transmission of mammalian mitochondrial dna. *PLoS Biol.* 6:63–71.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. Mega5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution.* 28:2731–2739.

Tang K, Thornton KR, Stoneking M. 2007. A new approach for using genome scans to detect recent positive selection in the human genome. *PLOS Biol.* 5.

Tautz D. 1989. Hypervariability of simple sequences as a general source for polymorphic dna markers. *Nucleic Acids Res.* 17:6463–6471.

Teeter KC, Payseur BA, Harris LW, Bakewell MA, Thibodeau LM, ÓBrien JE, Krenz JG, Sans-Fuentes MA, Nachman MW, Tucker PK. 2007. Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome Research.* 18:67–76.

Teschke M, Mukabayire O, Wiehe T, Tautz D. 2008. Identification of selective sweeps in closely related populations of the house mouse based on microsatellite scans. *Genetics.* 180:1537–1545.

Thomas M, Moller F, Wiehe T, Tautz D. 2007. A pooling approach to detect signatures of selective sweeps in genome scans using microsatellites. *Molecular Ecology Notes.* 7:400–403.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics.* 14:178–192.

VanValen L. 1973. Pattern and the balance of nature. *Evol Theory.* 1:31–49.

Ďureje L, Macholán M, E BSJ, Piálek J. 2012. The mouse hybrid zone in central europe: from morphology to molecules. *Folia Zool.* 61:308–318.

Wade CM, Kulbokas EJ, Kirby AW, Zody MC, Mullikin JC, Lander ES, Lindblad-Toh K, Daly MJ. 2002. The mosaic structure of variation in the laboratory mouse genome. *Nature.* 420:574–578.

Wang K, Li M, Hakonarson H. 2010. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*. 38:e164.

Webster M, Sheets HD. 2001. A practical introduction to landmark-based geometric morphometrics. *Quantitative Methods in Paleobiology*. 16:163–188. Paleontological Society Papers.

Yalcin B, Adams DJ, Flint J, Keane TM. 2012. Next-generation sequencing of experimental mouse strains. *Mammalian Genome*. 23:490–498.

Yang H, Bell TA, Churchill GA, Pardo-Manuel De Villena F. 2007. On the subspecific origin of the laboratory mouse. *Nature Genetics*. 39:1100–1107.

Yeh FC, Yang RC, Boyle T, Ye ZH, Mao JX. 1997. Popgene, the user-friendly shareware for populations genetic analysis. *Molecular Ecology Notes*. .

Yonekawa H, Moriwaki K, Gotoh O, Miyashita N, Matsushima Y, Shi LM, Cho WS, Zhen XL, Tagashira Y. 1988. Hybrid origin of japanese mice *Mus musculus molossinus*: evidence from restriction analysis of mitochondrial dna. *Mol Biol Evol*. 5:63–78.

Zimmermann K. 1953. Die hausmaus von helgoland, *Mus musculus* sspec. *Zeitschrift feur Seaugetierkunde*. 17:163–166.

Appendix

- **Supplementary Table 1:** Details of Nuclear markers and the primers used for typing.
- **Supplementary Table 2:** Details of the primers used for microsatellite loci typing.
- **Supplementary Table 3:** Primers used for mtDNA genome amplification and sequencing.
- **Supplementary Table 4:** List of mice skull specimens for the mice collected for this study.
- **Supplementary Table 5:** List of mice skull specimens collected during 1950s-1970s.
- **Supplementary Table 6:** List of mice skull specimens collected during 1935-1936.
- **Supplementary Table 7:** Summary statistics for 3 genome sequences from Heligoland.
- **Supplementary Table 8:** Details of chromosome statistics for HG_06 genome.
- **Supplementary Table 9:** Details of chromosome statistics for HG_08 genome.
- **Supplementary Table 10:** Details of chromosome statistics for HG_13 genome.
- **Supplementary Table 11:** Introgressed regions in the genome of house mouse from Heligoland and their frequencies
- **Supplementary Figure 1:** Histogram of the mean sequencing read depth per genome for the 3 sequenced samples.
- **Supplementary Figure 2:** IGV snapshot illustrating the mapped reads of *M. m. helgolandicus* genome to the mouse reference genome (NCBI/37) mm9.
- **Supplementary Figure 3:** IGV snapshot illustrating variants in the genome of *M. m. helgolandicus*.

Supplementary table 1: Details of the nuclear markers and the primers used for typing.

Gene/marker	Gene product/ function	Primer sequences
Abpa	Androgen binding protein	AbpFd 5'GAAACAATTCAATGAAAACACTAAAG AbpRd 5'TGTGCCACTGCTCTGTATTC
Abpb	Androgen binding protein	AbpFm 5'ACAATTCAATGAAAACCGTGA AbpRm 5'AAACTTGGGCAGGGATTTAG
D11 cenB2	Centromeric structure	D11 cenB2F 5'GTAACTCGCTGGTCTCTTCAT D11 cenB2Rd 5'CCACTCCTGCTTAGGACTGA D11 cenB2Rm 5'CCACTCCTGCTTAGGACCGTC
Btk	Bruton agammaglobulinemia tyrosine kinase	BtkF 5'AATGGGCTAGCGTAGTGCAG BtkR 5'AGGGGACGTACACTCAGCTTT
Zfy2	Zinc finger protein 2	Zfy2F 5'CATTAAGACAGAAAAGACCACCG Zfy2R 5'GTGAGGAAATTTCTTCCTGTGG

Supplementary table 2: Details of the primers used for microsatellite loci typing. Primers were designed by (Teschke et al., 2008)

Marker name	Chromosome	Primer Info	Forward Primer	Reverse Primer
D5Mit149	5	D5Mit149	TCAGGAAGTGATCTTCCAACG	ATCTGATGCCTCTGACCTTCA
Chr16_09	16	PP7B08	ACTAGCAGAGCCACTTCGGGAG	TCCCAGGATCGTCCAGCTCAGG
Chr11_31	11	PP4A02	TCCAGTCCCAGTTGCCAGACAC	AACCTGGTCTCAGAGGCTGTCG
Chr17_14	17	PP8A05	TCTATCAGATGGGCTCTGGGAG	TCATCCATGCAAAGTAGGCCTC
D9Mit330	9	D9Mit330	GAAATGAGGCTACTTACCACGG	GATATGTACATTCAGATGCATCCA
D1/EnsmusG22992	3	D1/EnsmusG22992	TAGAGCGTGAAGAAGAGGGC	AAGACTGACCAACATTCACGG
Chr3_38	3	PP6E9	GAGGCCAAGTGTCCAGCAGTGG	ACCAATCTGCATGCCATCATGG
E1/EnsmusG46849	15	E1/EnsmusG46849	GACAAGAAGGAGAGGGGGAG	ATCTGGCCACTTTGCTGAAG
D15Mit98	15	D12/D15Mit98	AGCACATTCTCCCAACAACC	CAAAACAAGCACAAAACAAATACA
D19Mit39	19	D19Mit39	GGAGGTCTCAGGAAATATTACTCC	ATTCCTGTGTAAAGGTGGATGG
Chr10_9	10	PP3A02	GCCGATAGCCTTGCTCTGGCTG	CCACTTGGTAAGGTGTCCATGC
D9Mit54	9	B9/D9Mit54	TGGGGATACTATGCCTTCTACTG	CAGGTCAAGGCTACTTTTATTTTC
D13Mit61	13	D13Mit61	TGCTCCAATACAACAAGGTCC	CCAGCCAAGGTGTGTTGAC
Chr08_22	8	PP10A02	AGAAAGCAGCTTACAGTCCCAG	TGGATTGGGATGAGCCTTGAAC
Chr04_07	4	PP10E08	AGGCTCCACCACGGAGCACCTC	GGAGGTTGAGTCCAGCCTAGTC
D14Mit203	14	P140a/D14Mit203	GTTAGCCAATTTAGAGGAGAGCC	CAGAACTCCAGTCTAACTATCACACA
X_2	X	PP2A01	ACTCTCAGGAATGTGTTGCGTC	TGGAGAGACCTTGGGCTACCCA
Chr02_31	2	PP8E11	ACTCGTGAGACACACAGTCCTG	GTGTCAGCCAACAGGCACTTGC
Chr12_04	12	PP9B08	ACACACGAGAAACCTCCTCTGC	ACCGTTGCTAGGTGAACGCGAC
D6Mit309	6	D6Mit309	TATGCTTTTTTTCAAATCTGTTGC	CACTAGGAAACCCACCCTGA
Chr18_06	18	PP9B10	TGACCATACCCGGCTATTGGCA	CCAAGGTTCCCAACAGAGCCTG

Supplementary table 3: Primers used for mtDNA genome amplification and sequencing. The primers were obtained from (Stewart et al., 2008) and the additional list was designed by (Hardouin and Tautz, 2013)

Oligo Name	Sequence 5'-3'	annealing temperature	Oligo Name	Sequence 5'-3'	annealing temperature
mtF1	agagaactactagccatagc	52°C	mtF15	catacatttacacctactacc	60°C
mtR1	tgcggtactagttctatagc	52°C	mtR15	tgtacaataggagtgtggtgg	60°C
mtF2	atgaacactctgaactaatcc	52°C	mtF16	ctactcaccaatatcctcac	62°C
mtR2	tactcatactaacagtgttgc	52°C	mtR16	atgaagataacagtgtacagg	62°C
mtF3	agaaagcgttcaagctcaac	62°C	mtF17	ttgatgaggatcttactccc	64°C
mtR3	actttgacttgtaagtctagg	62°C	mtR17	gtaggttgagattttggacg	64°C
mtF4	ttgacctttcagtgaagagg	62°C	mtF17b	aaaaaaattaatgatttcgactc	60°C
mtR4	agaacactattagggagagg	62°C	mtR17b	ttttaaactaattaccatttactctg	60°C
mtF5	agagaaggttattagggtgg	62°C	mtF19	aatcggttctattccactgc	62°C
mtR5	ttgtttctgctagggttgag	62°C	mtR19	gctagattagctagacttgc	62°C
mtF6	tcactattcggagctttacg	62°C	mtF20	catcatcactcctattctgc	62°C
mtR6	ggataaggtgtttaggtagc	62°C	mtR20	agaatttgattgatgtggtgg	62°C
mtF7	caagccctcttatttctagg	62°C	mtF21	tcttcattcttctactatccc	62°C
mtR7	gatagtagagttgagtagcg	62°C	mtR21	gtacttgagtgtagtagtgct	62°C
mtF8	ccattccacttctgattacc	62°C	mtF22	aggaaaatcagcacaatttgg	62°C
mtR8	tgcttatgatagctagggtg	62°C	mtR22	gacaaatcctgcaaagatgc	62°C
mtF9	cactcatagcaataatagctc	62°C	mtF23	acagctatgtacagcatacg	62°C
mtR9	aaaagcatgggcagttacg	62°C	mtR23	atgatgttggagttatgttgg	62°C
mtF9b	aatggcggtagaagtcttag	60°C	mtF24	tactaccatcattcaagtagc	62°C
mtR9b	Tttttcggcggtagaagtag	60°C	mtR24	tctgatgtgtagtgtatggc	62°C
mtF10	ggaatagtgggtactgcac	62°C	mtF25	cactcattcattgacctacc	62°C
mtR10	gctgatgtaaagtaagctcg	62°C	mtR25	gtatagtaggggtgaaatgg	62°C
mtF11	ctgctcctattatcactacc	62°C	mtF26	gctttccacttcatcttacc	62°C
mtR11	ttgctcatgtgtcatctagg	62°C	mtR26	tcatttcaggtttacaagacc	62°C
mtF12	gatacatactatgtagtagcc	60°C	mtF27	cttatcttaacctgaattggg	62°C
mtR12	gatatagaggactaaggagc	60°C	mtR27	cacagttatgttggtcatgg	62°C
mtF13	tccaacttggtctacaagac	58°C	mtF28	gaaactttatcagacatctgg	62°C
mtR13	ttgatgtatctagttgtgg	58°C	mtR28	tctatggaggtttgcatgtg	62°C
mtF14	tgtcctagaaatggttccac	60°C	mtF29	tgtatcccataaacacaaagg	62°C
mtR14	tttagtttgtgtcggaagcc	60°C	mtR29	gctgaattagcaagagatgg	62°C

Additional sequencing primers

Oligo Name	Sequence 5'-3'	annealing temperature
mtF7b	gatccgagcatcttatccacgct	64°C
mtR7b	tctggtaatcagaagtggaatgg	64°C
mtF7t	gggcccataccccgaaaacgt	64°C
mtR7t	ggttgagtagagtgagggatgg	64°C
mt18bis_F	cataagctccataccaatcccca	60°C
mt18bis_R	atgagggcaattagcagtgga	60°C
mt21bis_F	cttaggaaccaaaaaccttggtg	68°C
mt21bis_R	gtagtgctgaaactggtgtagg	68°C

Sublementary table 4: List of mice skull specimens for the mice collected for this study.

No.	Location of samples	Land	Location of skull collection	Species	Date of collection	Collection number	Sex
1	North see, Heligoland	Germany	Max-Planck Institute	*M. musculus*	22-07-2008	HG-1450_1	M
2	North see, Heligoland	Germany	Max-Planck Institute	*M. musculus*	23-07-2008	HG-1450_2	F
3	North see, Heligoland	Germany	Max-Planck Institute	*M. musculus*	06-06-2012	HG_01	M
4	North see, Heligoland	Germany	Max-Planck Institute	*M. musculus*	06-06-2012	HG_03	M
5	North see, Heligoland	Germany	Max-Planck Institute	*M. musculus*	06-06-2012	HG_02	M
6	North see, Heligoland	Germany	Max-Planck Institute	*M. musculus*	07-06-2012	HG_04	M
7	North see, Heligoland	Germany	Max-Planck Institute	*M. musculus*	07-06-2012	HG_05	M
8	North see, Heligoland	Germany	Max-Planck Institute	*M. musculus*	07-06-2012	HG_06	F
9	North see, Heligoland	Germany	Max-Planck Institute	*M. musculus*	08-06-2012	HG_07	M
10	North see, Heligoland	Germany	Max-Planck Institute	*M. musculus*	14-06-2004	HG_08	M
11	North see, Heligoland	Germany	Max-Planck Institute	*M. musculus*	28-08-2004	HG_09	M
12	North see, Heligoland	Germany	Max-Planck Institute	*M. musculus*	19-04-2008	HG_10	M
13	North see, Heligoland	Germany	Max-Planck Institute	*M. musculus*	05-03-2012	HG_11	F
14	North see, Heligoland	Germany	Max-Planck Institute	*M. musculus*	19-10-2010	HG_12	M
15	North see, Heligoland	Germany	Max-Planck Institute	*M. musculus*	04-12-2011	HG_13	F
16	North see, Heligoland	Germany	Max-Planck Institute	*M. musculus*	17-05-2012	HG_14	M
17	North see, Heligoland	Germany	Max-Planck Institute	*M. musculus*	17-05-2012	HG_15	M

Supplementary table 5: List of mice skull specimens collected during 1950s-1970s.

No.	Location of samples	Land	Location of skull collection	Species	Date of collection	Collection number	Sex
1	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	29/9/1964	4657	F
2	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	7/12/1963	4661	M
3	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	12/11/1964	4668	F
4	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	12/14/1957	9180	M
5	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	10/4/1959	9188	M
6	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	22/4/1959	9190	F
7	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	23/4/1959	9191	F
8	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	2/5/1959	9192	F
9	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	9/12/1960	9201	M
10	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	9/12/1960	9202	-
11	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	26/4/1963	9203	M
12	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	12/6/1969	10008	M
13	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	26/6/1969	10011	F
14	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	15/5/1969	10013	M
15	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	24/11/69	10418	M
16	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	24/11/1969	10420	M
17	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	29/5/1970	11248	F
18	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	30/6/1970	11250	M
19	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	2/7/1970	11251	M
20	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	30/6/1970	11255	F
21	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	25/11/1970	11686	M
22	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	6/8/1970	11690	F
23	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	12/11/1970	11691	F
24	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	18-11-1970	11698	M
25	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	10/9/1970	11699	M
26	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	12-02-1970	11703	F
27	North see, Heligoland	Germany	Haustierkd. Univ. Kiel	*M. musculus*	26-05-1971	12387	M

Supplementary table 6: List of mice skull specimens collected during 1935-1936.

No.	Location of samples	Land	Location of skull collection	Species	Date of collection	Collection number	Sex
1	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105261	M
2	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105262	F
3	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105263	-
4	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105264	-
5	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105265	-
6	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105266	-
7	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105267	F
8	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105268	M
9	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105269	F
10	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105270	F
11	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105271	F
12	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105272	M
13	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105273	M
14	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105274	M
15	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105275	F
16	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105276	F
17	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105277	F
18	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105278	M
19	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105279	-
20	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105280	M
21	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105281	M
22	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105282	F
23	North see, Heligoland	Germany	Zoological Museum Berlin	*M. musculus*	1935-1936	ZMB_105283	-

Details of whole genome sequence analysis

Reference genome indexing

bwa index -p mm9_genome -a bwtsw mm9_genome.fasta > index.log

Trimming of the reads

java -jar trimmomatic-0.30.jar PE -threads 4 -phred33 -trimlog tim06log HG_1_sequence.fq.gz HG_2_sequence.fq.gz HG_1_paired.fq.gz HG_1_unpaired.fq.gz HG_2_paired.fq.gz HG_2_unpaired.fq.gz ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:60

Mapping reads to the genome

bwa aln -n 2 -t 4 ./index/mm9_genome HG_1_paired.fq > HG_1.sai

bwa aln -n 2 -t 4 ./index/mm9_genome HG_2_paired.fq > HG_2.sai

bwa sampe ./index/mm9_genome HG_seq_1.sai HG_seq_2.sai HG_1_paired.fq HG_2_paired.fq > HG.sam

Converting SAM to BAM, sorting and indexing mapped reads

samtools view -bS -o HG_unstr.bam HG.sam

samtools sort HG_unstr.bam HG_str | samtools rmdup -S HG_str.bam HG_ind.bam | samtools index HG_ind.bam

Whole genome sequence summary statistics

Supplementary table 7: Summary statistics for 3 genome sequences from Heligoland

Globals	HG_06	HG_08	HG_13
Reference size	2725765481	2725765481	2725765481
Number of reads	326010421	419536531	360615570
Mapped reads	296,163,319 / 90.84%	375,493,228 / 89.5%	326,893,306 / 90.65%
Unmapped reads	29,847,102 / 9.16%	44,043,303 / 10.5%	33,722,264 / 9.35%
Paired reads	296,163,319 / 90.84%	375,493,228 / 89.5%	326,893,306 / 90.65%
Mapped reads, only first in pair	148,053,652 / 45.41%	187,744,613 / 44.75%	163,421,273 / 45.32%
Mapped reads, only second in pair	148,109,667 / 45.43%	187,748,615 / 44.75%	163,472,033 / 45.33%
Mapped reads, both in pair	290,210,206 / 89.02%	366,366,661 / 87.33%	320,285,413 / 88.82%
Mapped reads, singletons	5,953,113 / 1.83%	9,126,567 / 2.18%	6,607,893 / 1.83%
Read min/max/mean length	60 / 101 / 99.17	60 / 101 / 99.2	60 / 101 / 99.19
Clipped reads	8,661,600 / 2.66%	11,712,402 / 2.79%	9,698,719 / 2.69%
Duplication rate	6.80%	8.71%	7.57%
ACGT Content			
Number/percentage of A's	8,681,594,427 / 29.74%	10,856,889,653 / 29.35%	9,411,928,542 / 29.21%
Number/percentage of C's	5,955,396,040 / 20.4%	7,671,004,497 / 20.74%	6,719,402,481 / 20.86%
Number/percentage of T's	8,594,743,247 / 29.44%	10,755,631,706 / 29.08%	9,341,118,985 / 28.99%
Number/percentage of G's	5,964,578,346 / 20.43%	7,701,577,171 / 20.82%	6,746,410,555 / 20.94%
Number/percentage of N's	0 / 0%	0 / 0%	0 / 0%
GC Percentage	40.83%	41.56%	41.79%
Coverage			
Mean	10.72	13.58	11.83
Standard Deviation	720.6	1053.72	793.86
Mapping Quality			
Mean Mapping Quality	48.06	47.88	48.06
Indels			
Total reads with indels	9618258	11809533	10145668
Insertions	4233810	5120586	4439914
Deletions	5384448	6688947	5705754
Homopolymer indels	61.57%	61.14%	60.94%
Insert size			
Mean	254.82	236.42	241.38
Median	238	220	224

SNPs calling

samtools mpileup -I -F 0.0005 -gf ./index/mm9_genome.fasta HG_06.bam HG_08.bam HG_13.bam | bcftools view -bcvg -> HG_samples.bcf

bcftools view HG_samples.bcf | perl vcfutils.pl varFilter -Q 20 -d 15 -D 2000 > HG_SNPs.vcf

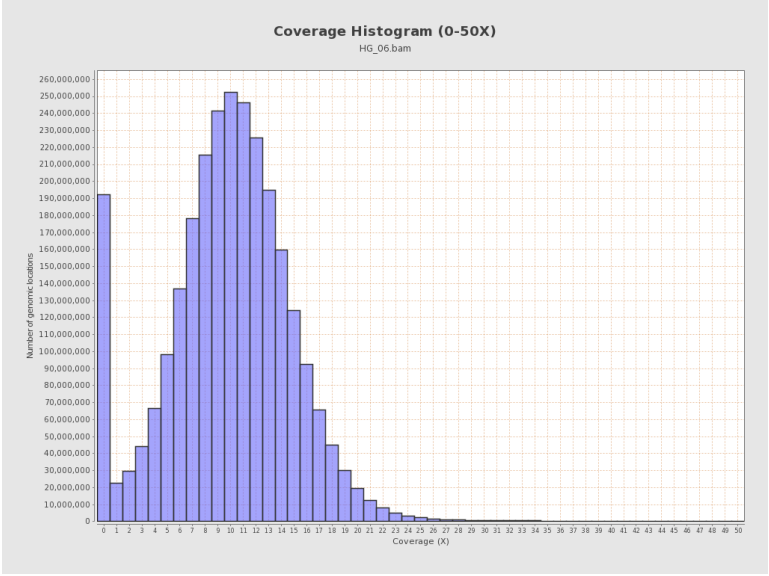
Detection of the variants

java -Xmx4g -jar snpEff.jar eff -c snpEff.config -v NCBIM37.64 HG_SNPs.vcf -s snpEff1 > HG_SNPs_trim_1.vcfs

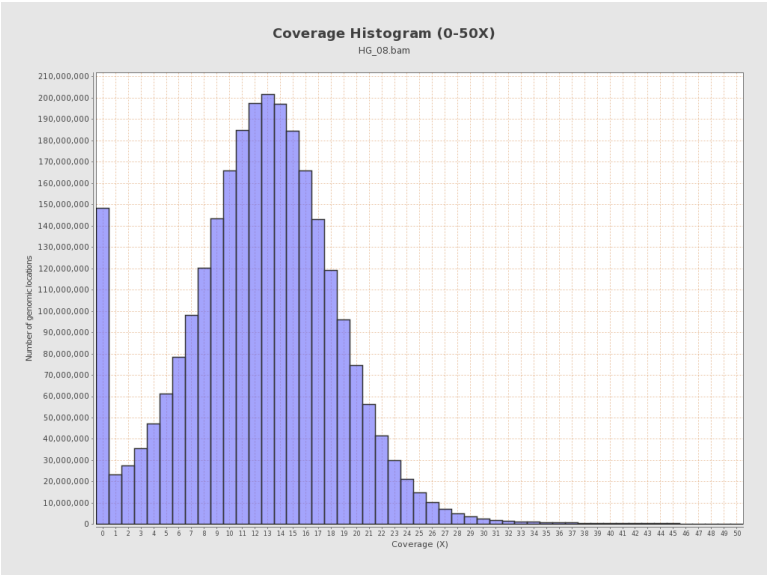
Detection of variants using ANNOVAR and dbSNP128 for detection of overlapping with reference data

perl convert2annovar HG_SNPs.vcf -format vcf4 > HG_SNPs.annovar

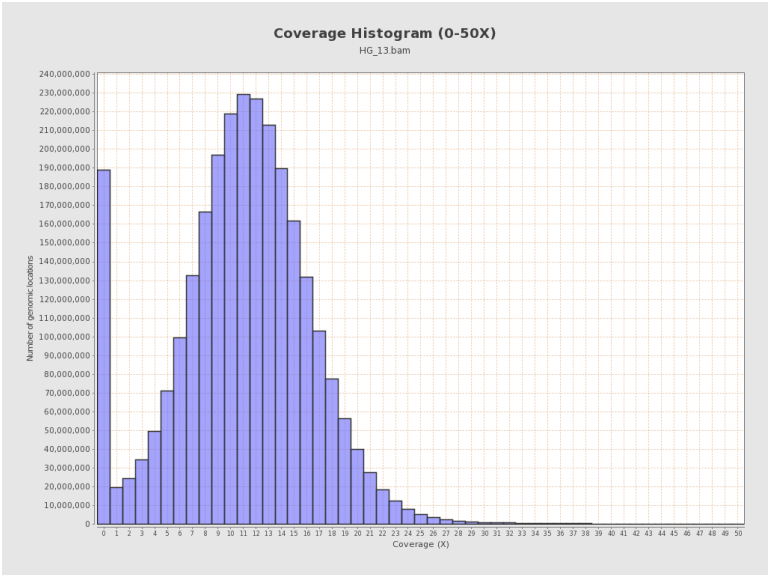
perl annotate variation.pl –filter –buildver mm9 –dbtype snp128 HG_SNPs.annovar mousedb/



[I]



[II]



[III]

Supplementary figure 1: Histogram of the mean sequencing read depth per genome for the 3 sequenced samples.

Supplementary table 8: Details of chromosome statistics for HG_06 genome

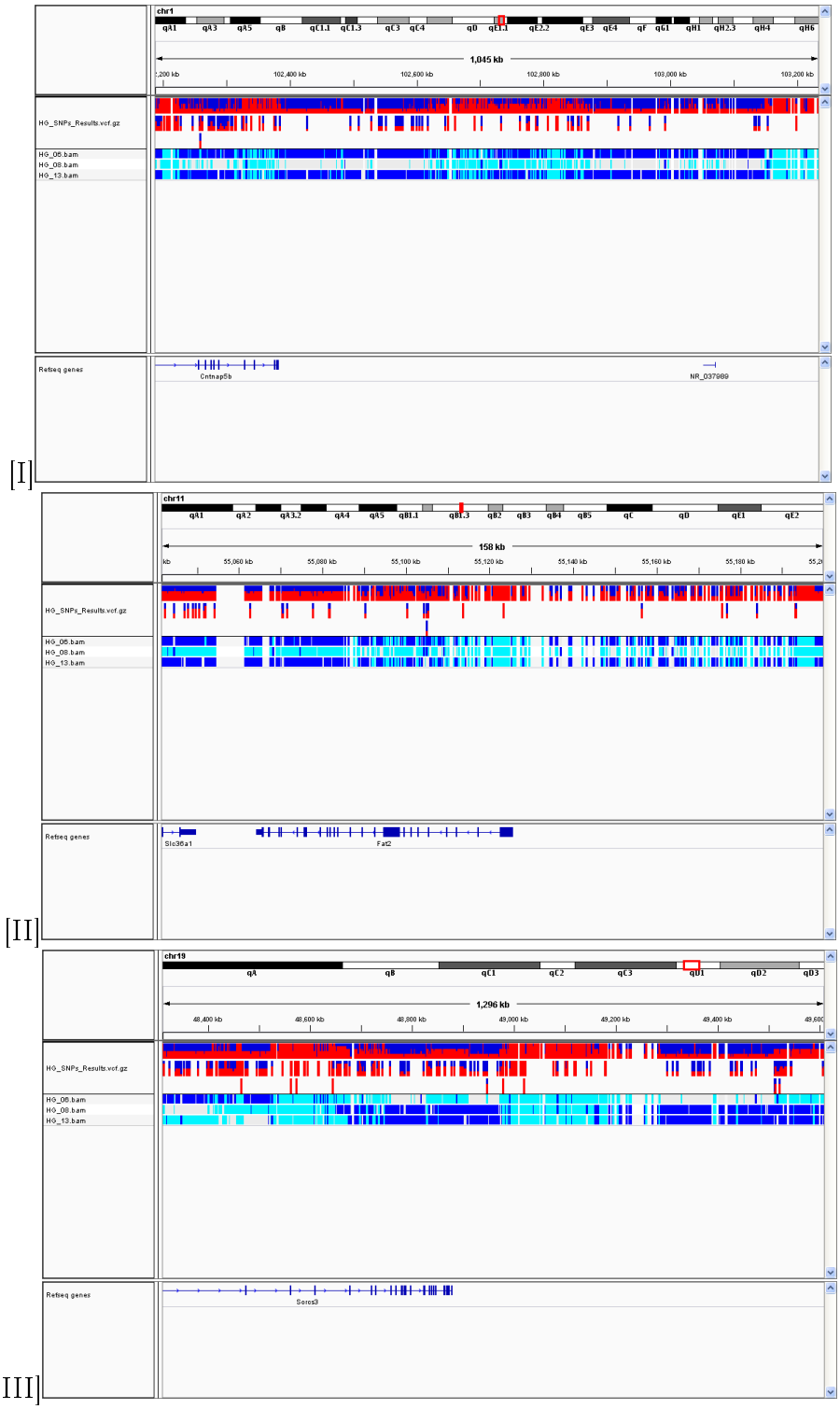
Chromosom stats			HG_06	
Name	**Length**	**Mapped bases**	**Mean coverage**	**Standard deviation**
chr1	197195432	2101943208	10.66	56.45
chr1_random	1231697	19619340	15.93	41.23
chr2	181748087	2844429013	15.65	2619.98
chr3	159599783	1678024489	10.51	6.2
chr3_random	41899	221665	5.29	2.7
chr4	155630120	1581504836	10.16	8.78
chr4_random	160594	1015936	6.33	12.38
chr5	152537259	1563743830	10.25	7.75
chr5_random	357350	2485189	6.95	5.04
chr6	149517037	1601479202	10.71	390.27
chr7	152524553	1454702672	9.54	6.17
chr7_random	362490	2351813	6.49	6.46
chr8	131738871	1329947988	10.1	11.34
chr8_random	849593	32268068	37.98	72.34
chr9	124076172	1773168993	14.29	987.81
chr9_random	449403	4659657	10.37	11.15
chr10	129993255	1355672013	10.43	7.61
chr11	121843856	1232562994	10.12	15.02
chr12	121257530	1268365664	10.46	373.27
chr13	120284312	1252184250	10.41	26.01
chr13_random	400311	2048154	5.12	5.17
chr14	125194864	1303438312	10.41	9.55
chr15	103494974	1066870005	10.31	12.58
chr16	98319150	1019156663	10.37	6.12
chr16_random	3994	60732	15.21	13.51
chr17	95272651	962785504	10.11	9.62
chr17_random	628739	2906487	4.62	3.48
chr18	90772031	946155042	10.42	28.43
chr19	61342430	601048317	9.8	5.18
chrM	16299	301953549	18525.89	5070.52
chrUn_random	5900358	35583103	6.03	70.47
chrX	166650296	1844886851	11.07	27.45
chrX_random	1785075	10709641	6	9.42
chrY	15902555	1931464	0.12	2.17
chrY_random	58682461	10817192	0.18	2.83

Supplementary table 9: Details of chromosome statistics for HG_08 genome

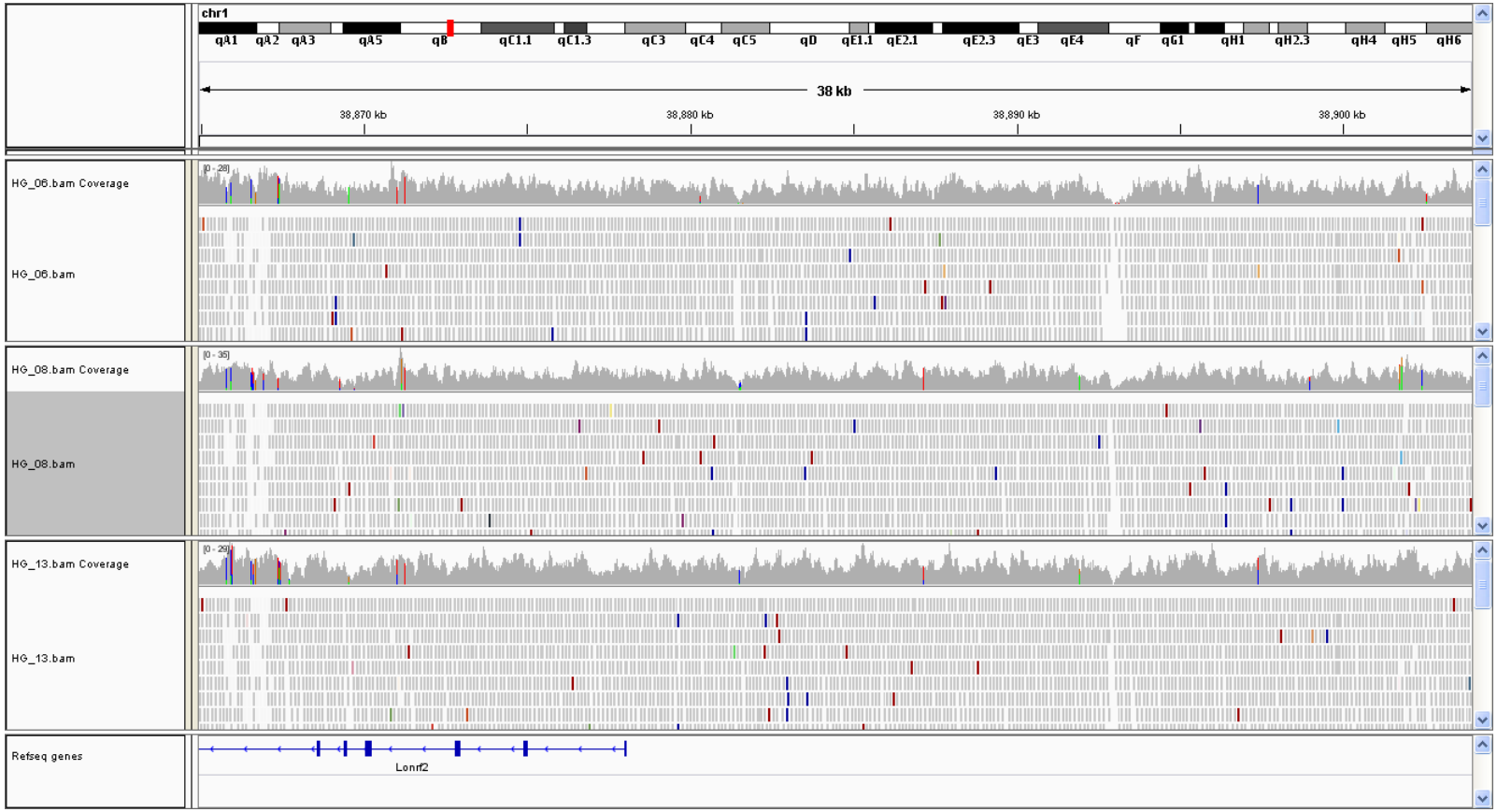
Chromosome stats	HG_08			
Name	**Length**	**Mapped bases**	**Mean coverage**	**Standard deviation**
chr1	197195432	2636210272	13.37	23.31
chr1_random	1231697	26117643	21.2	48.55
chr2	181748087	3800652782	20.91	3824.95
chr3	159599783	2116829520	13.26	7.34
chr3_random	41899	300441	7.17	2.9
chr4	155630120	2044813731	13.14	9.59
chr4_random	160594	1260839	7.85	15.43
chr5	152537259	2028047492	13.3	9.71
chr5_random	357350	3291529	9.21	6.51
chr6	149517037	2068608688	13.84	537.66
chr7	152524553	1945208937	12.75	7.71
chr7_random	362490	3079354	8.5	8.56
chr8	131738871	1724381371	13.09	14.7
chr8_random	849593	42820236	50.4	98.34
chr9	124076172	2427490056	19.56	1519.66
chr9_random	449403	5782002	12.87	17.21
chr10	129993255	1724776635	13.27	10.12
chr11	121843856	1623443240	13.32	14.57
chr12	121257530	1629163953	13.44	545.5
chr13	120284312	1606695001	13.36	30.29
chr13_random	400311	2655232	6.63	6.59
chr14	125194864	1668398414	13.33	10.94
chr15	103494974	1375836774	13.29	14.87
chr16	98319150	1290563857	13.13	7.34
chr16_random	3994	80259	20.09	17.49
chr17	95272651	1249657031	13.12	10.07
chr17_random	628739	3850149	6.12	4.51
chr18	90772031	1208006825	13.31	23.86
chr19	61342430	783432136	12.77	6.4
chrM	16299	116745148	7162.72	2149.86
chrUn_random	5900358	46524432	7.89	127.47
chrX	166650296	1266073049	7.6	26.94
chrX_random	1785075	9508818	5.33	14.75
chrY	15902555	17003137	1.07	8.28
chrY_random	58682461	505236608	8.61	24.41

Supplementary table 10: Details of chromosome statistics for HG_13 genome

Chromosome stats	HG_13			
Name	**Length**	**Mapped bases**	**Mean coverage**	**Standard deviation**
chr1	197195432	2275499001	11.54	17.22
chr1_random	1231697	23266414	18.89	54.5
chr2	181748087	3160981512	17.39	2881.17
chr3	159599783	1823474566	11.43	6.64
chr3_random	41899	262187	6.26	2.76
chr4	155630120	1776713254	11.42	8.49
chr4_random	160594	1001487	6.24	12.77
chr5	152537259	1745956205	11.45	8.4
chr5_random	357350	1648523	4.61	3.69
chr6	149517037	1778534691	11.9	418.72
chr7	152524553	1668181668	10.94	6.9
chr7_random	362490	2635064	7.27	6.95
chr8	131738871	1485072940	11.27	10.61
chr8_random	849593	28991873	34.12	61.67
chr9	124076172	2025977080	16.33	1143.7
chr9_random	449403	4901149	10.91	13.24
chr10	129993255	1497805199	11.52	8.66
chr11	121843856	1425126777	11.7	13.78
chr12	121257530	1409439417	11.62	402.62
chr13	120284312	1414431282	11.76	24.11
chr13_random	400311	2350226	5.87	6.07
chr14	125194864	1440405459	11.51	10.27
chr15	103494974	1194595194	11.54	16.45
chr16	98319150	1118081228	11.37	6.69
chr16_random	3994	64316	16.1	15.49
chr17	95272651	1090049474	11.44	10.94
chr17_random	628739	3427310	5.45	4.09
chr18	90772031	1047996934	11.55	24.93
chr19	61342430	680547945	11.09	5.79
chrM	16299	78256590	4801.31	1431.65
chrUn_random	5900358	36093929	6.12	96.53
chrX	166650296	1966805944	11.8	25.71
chrX_random	1785075	11388783	6.38	10.43
chrY	15902555	2177161	0.14	2.22
chrY_random	58682461	12037159	0.21	2.54



Supplementary figure 2: IGV snapshot illustrating the mapped reads of *M. m. helgolandicus* genome to the mouse reference genome (NCBI/37) mm9, (I) variants across chromosome 1, (II) variants across chromosome 11, (III) variants across chromosome 19.



Supplementary figure 3: IGV snapshot illustrating variants in the genome of *M. m. helgolandicus*.

Supplementary table 11: Introgressed regions in the genome of house mouse from Heligoland and their frequencies.

Chromosome	Chrom_start	Chrom_end	Frequency
chr1	8433985	9042470	2
chr1	10828106	10865227	1
chr1	11689414	11824165	2
chr1	11973791	12107080	2
chr1	12357962	12743248	2
chr1	12944072	12984882	1
chr1	13063422	13088572	1
chr1	14217106	14474489	1
chr1	17648838	17743690	4
chr1	21598555	21631121	1
chr1	21909864	22026326	4
chr1	23430986	23596966	1
chr1	24299499	24344816	1
chr1	28543693	29016450	4
chr1	31973240	32697100	5
chr1	40434599	40598251	6
chr1	68578242	68734285	2
chr1	72308343	72416529	2
chr1	78616482	78622730	6
chr1	89365399	89547837	2
chr1	92829615	93034319	2
chr1	93635932	94148004	3
chr1	105987750	106537297	2
chr1	113990281	114330676	2
chr1	115102642	115488397	2

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr1	126818336	126968184	4
chr1	129454638	129474853	4
chr1	131495900	131582613	4
chr1	132801232	133086383	2
chr1	133344779	133420666	2
chr1	138226955	138280510	4
chr1	140729929	140853457	2
chr1	143879984	144157359	3
chr1	145250995	145335981	1
chr1	153142815	153158653	4
chr1	172389694	172411831	2
chr1	172713357	172769188	1
chr1	173087045	173368509	2
chr1	177198221	177249525	6
chr1	177980609	180832210	6
chr1	180905515	185454649	6
chr1	185456099	185561990	2
chr1	185760024	186313657	2
chr1	186359882	186838505	1
chr1	186862350	187420604	1
chr2	7444571	7510347	6
chr2	9593507	9727450	4
chr2	13099567	13137526	3
chr2	14897218	16582146	3
chr2	17943536	18672986	1
chr2	18786342	18880319	1

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr2	19267933	19274150	4
chr2	19789590	19812970	1
chr2	20564108	20790854	5
chr2	27851721	27901113	3
chr2	28517705	28744004	4
chr2	32246747	33022044	3
chr2	33837690	33889819	3
chr2	41513100	41620660	1
chr2	41998137	42071053	2
chr2	42175339	42356100	1
chr2	46824813	46928401	2
chr2	49560831	49609124	2
chr2	49926481	49977747	1
chr2	55230701	55841462	1
chr2	59064798	59257460	4
chr2	62353077	62458581	5
chr2	71959056	72183538	1
chr2	75523701	75855833	1
chr2	77593271	77634864	1
chr2	78199181	78225159	1
chr2	78262444	78431293	2
chr2	78472238	78524663	1
chr2	79615443	80074298	1
chr2	88874705	89265380	1
chr2	101798383	101979456	2
chr2	115098703	115276355	1

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr2	120996444	121277588	2
chr2	128196628	128342304	1
chr2	130070957	130112609	2
chr2	134242734	134609018	4
chr2	134800003	134874518	4
chr2	134907909	135727324	6
chr2	141239260	141529332	6
chr2	148724260	149810230	6
chr2	151586829	151646244	5
chr2	152309358	152330044	5
chr2	152547905	153776379	1
chr2	153830685	154194644	3
chr2	154228322	154510928	1
chr2	154533382	155060764	1
chr2	156460495	156574887	6
chr2	159977214	160312345	6
chr2	163080480	163387628	6
chr2	168356303	168421473	3
chr2	178163683	178301463	2
chr2	178355341	178389525	2
chr3	3033781	5572503	4
chr3	8928455	9114091	1
chr3	10334638	10564221	2
chr3	11640951	11842859	2
chr3	13708786	14318529	4
chr3	19570358	19581264	4

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr3	22167397	23025981	3
chr3	34367164	34491272	3
chr3	36466562	36498280	1
chr3	38582304	38612110	1
chr3	39684221	39808041	2
chr3	40231643	40510696	4
chr3	41442459	41485191	1
chr3	45573108	45690459	3
chr3	52870572	52945174	5
chr3	54735926	54741463	1
chr3	59778759	60219300	4
chr3	63273551	63827241	6
chr3	65858672	66099202	2
chr3	66200790	66292705	3
chr3	67048884	67372984	1
chr3	83168905	83307617	6
chr3	88017469	88222147	6
chr3	94152913	94157487	2
chr3	95737746	95850247	2
chr3	96098617	96145240	1
chr3	97502613	97551583	3
chr3	97938531	97981701	3
chr3	110764022	111538513	6
chr3	118283851	119331296	1
chr3	122117819	122291160	1
chr3	129543353	129616386	4

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr3	131260005	131411287	4
chr3	134277106	134827514	5
chr3	135430998	135703969	6
chr3	136059720	136064022	1
chr3	136949405	137015180	1
chr3	139996234	140039368	1
chr3	141728392	142467966	6
chr3	142594211	142679328	1
chr3	148979077	148989395	1
chr3	150663921	150736708	1
chr3	151734381	152023961	2
chr3	152904420	152966171	6
chr3	155284617	155468521	1
chr3	156540580	156840721	1
chr4	5180616	5580478	1
chr4	11939793	12091924	6
chr4	19754815	19802437	1
chr4	35177287	35209827	6
chr4	42229812	43284658	6
chr4	57217299	57586983	6
chr4	58358371	58412371	4
chr4	61913679	62311190	6
chr4	80937909	80961094	6
chr4	83988281	84035580	2
chr4	84150884	84181788	2
chr4	89744861	89961836	6

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr4	95009141	95019711	6
chr4	96189063	96196229	1
chr4	106338827	106492492	4
chr4	108009667	108350681	4
chr4	108611059	108820662	1
chr4	109163834	109198308	1
chr4	114555979	114665911	1
chr4	115320514	115611586	5
chr4	117428206	117432098	1
chr4	125386000	125432063	3
chr4	131204344	131385649	1
chr4	133119279	133290807	5
chr4	135185325	135252551	6
chr4	138917424	139062664	1
chr4	139112406	139172219	1
chr4	139606139	139621439	2
chr4	152935844	152969777	3
chr5	10108412	10333704	1
chr5	13893745	13982394	6
chr5	14101621	14190086	6
chr5	15739325	15808957	6
chr5	20164025	20446377	6
chr5	24323716	24382068	6
chr5	24635173	24650568	6
chr5	28987428	29054407	4
chr5	30586583	30613889	1

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr5	31638650	31888739	2
chr5	32240489	32329548	2
chr5	33683980	33713777	1
chr5	35760555	35862373	3
chr5	35955100	36043094	2
chr5	36167878	36195355	1
chr5	36486697	36524542	3
chr5	37481865	37549849	1
chr5	38567873	38757333	1
chr5	40345343	40571420	2
chr5	43762573	43792724	2
chr5	46732706	46936312	5
chr5	50719327	50812880	6
chr5	51784223	52157244	6
chr5	54214108	54718477	5
chr5	55305154	55386418	2
chr5	55405225	55496419	2
chr5	55940564	56097784	2
chr5	56831661	57084689	4
chr5	62095445	62122523	1
chr5	62355696	62434233	1
chr5	64602875	64805942	1
chr5	64919537	65024819	1
chr5	66193881	66453726	1
chr5	66726544	66875339	1
chr5	66980550	67544007	1

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr5	67821286	67857705	1
chr5	68686566	70180973	2
chr5	71688732	72106810	2
chr5	75004153	75012875	2
chr5	78456130	80111297	5
chr5	82652837	82725250	1
chr5	87019705	87260201	5
chr5	96583740	96772353	1
chr5	98338604	98375508	4
chr5	102114467	102149535	6
chr5	111314735	111396525	1
chr5	113138269	113176836	2
chr5	114734761	114908636	1
chr5	115005114	115245561	1
chr5	118001596	118198459	6
chr5	124073775	124461993	3
chr5	126656984	126677476	1
chr5	128253893	128262111	3
chr5	129224868	129240744	2
chr5	129263248	129449441	3
chr5	129706651	129817889	4
chr5	130015943	130147279	4
chr5	132173070	132297668	1
chr5	134170285	134266194	4
chr5	141991234	142073001	6
chr5	142092609	142126318	6

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr5	143204048	143628917	6
chr5	144483535	144521018	6
chr5	151313734	151972105	3
chr6	5728015	5934655	2
chr6	14471420	14969096	2
chr6	15085551	15519552	5
chr6	22936219	22950007	6
chr6	23310776	23334989	3
chr6	25030342	25686766	3
chr6	27386077	27882166	2
chr6	33146783	33175552	6
chr6	35376072	35752787	3
chr6	36359558	36750369	2
chr6	37084728	37121833	2
chr6	43636709	43760399	1
chr6	48542198	48585144	6
chr6	49803428	50009640	1
chr6	52398925	53202221	5
chr6	58573287	58816010	6
chr6	59212509	60797531	6
chr6	62718904	63177391	6
chr6	64751276	65012236	1
chr6	65871000	66493671	1
chr6	67079575	67283165	1
chr6	67722777	68095734	1
chr6	73034005	73087525	4

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr6	73987232	74005558	1
chr6	76419507	76481691	5
chr6	82445457	82674041	4
chr6	84485492	84506402	4
chr6	92855460	92955457	3
chr6	93946035	94045258	4
chr6	95477216	95870672	2
chr6	96256786	96556873	4
chr6	100034028	100128039	4
chr6	100346539	100408479	1
chr6	103697643	103724221	1
chr6	108168973	108214189	1
chr6	110153138	110299097	6
chr6	110774976	110816458	6
chr6	111638929	111788355	6
chr6	112910391	112951323	3
chr6	113496580	113602572	2
chr6	113967393	114039573	3
chr6	116926947	117006506	3
chr6	117718155	117758675	1
chr6	122645895	123076366	2
chr6	125629793	125669878	2
chr6	127094539	127227530	4
chr6	129495230	129504608	1
chr6	130600482	131470053	1
chr6	139665122	139684391	2

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr7	6096388	6379886	6
chr7	25254592	25298203	1
chr7	25594594	25923734	5
chr7	29606182	29671306	5
chr7	30434189	30559851	3
chr7	31055944	31205553	3
chr7	31259480	31279361	1
chr7	36813630	37004412	6
chr7	37746344	37778670	1
chr7	47498110	47502920	1
chr7	52043709	52373390	6
chr7	53766774	54475478	6
chr7	59137271	59269487	6
chr7	74920335	75294558	2
chr7	97108997	97120948	5
chr7	99550571	99566249	1
chr7	107418581	107685462	2
chr7	110038466	110066713	6
chr7	114841819	115511431	5
chr7	117088554	117775301	5
chr7	118388312	118408313	1
chr7	122211149	122264899	6
chr7	128632682	128907216	5
chr7	131129087	131538460	2
chr7	132325759	132352846	2
chr7	140077707	140116572	1

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr7	141951518	141963845	3
chr7	142918108	142989816	6
chr7	143711613	143938111	6
chr7	144873187	144923301	2
chr7	145854916	146136966	2
chr7	146526441	146639043	2
chr7	147192595	147286074	3
chr7	150461873	150515783	1
chr7	151350690	151365451	1
chr8	10362323	10719399	6
chr8	25674692	25765487	1
chr8	27615520	27871902	6
chr8	46579489	46612318	1
chr8	47327218	47407014	2
chr8	48517666	48831666	2
chr8	50157887	50203400	4
chr8	50420578	50449120	1
chr8	54235929	54425364	2
chr8	63095964	64001995	3
chr8	69727504	70071341	2
chr8	74061453	74289242	1
chr8	77594799	77697012	3
chr8	82929889	83006160	1
chr8	85812925	90190216	2
chr8	90206267	90436006	2
chr8	90469686	90509440	3

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr8	90516296	91618897	4
chr8	92340085	92407288	1
chr8	92708467	92774822	1
chr8	94805894	94823052	2
chr8	95132756	95266252	2
chr8	99726135	100258359	2
chr8	115889280	116040449	3
chr8	124034037	124232323	1
chr8	127185274	127216700	3
chr8	128693022	128744485	1
chr8	129098505	129390251	4
chr8	129629481	129905707	4
chr8	131014833	131467068	3
chr9	4732503	5366644	2
chr9	8057511	8503638	2
chr9	12235481	12688441	6
chr9	14990147	15171956	4
chr9	16149668	16207035	1
chr9	20351679	20556956	1
chr9	21868695	22013270	4
chr9	27061608	27259014	2
chr9	27333654	27479922	2
chr9	29015257	29239534	2
chr9	29307913	29450576	4
chr9	29877671	30125346	2
chr9	34869349	34919426	4

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr9	34937091	34954178	4
chr9	34981562	35110093	4
chr9	37869726	37993061	1
chr9	40858823	40928370	3
chr9	44040459	44520470	2
chr9	45492868	45505547	1
chr9	47254644	47282765	1
chr9	47295482	47333902	1
chr9	47965138	47992400	1
chr9	48291180	48325559	1
chr9	50282966	50470535	2
chr9	50653204	50952892	4
chr9	58587381	58645248	5
chr9	60011683	60019609	6
chr9	63607030	63734392	1
chr9	64039962	64142258	5
chr9	64155869	64204813	1
chr9	65312971	65356172	3
chr9	65582247	65635795	1
chr9	73472130	73554282	3
chr9	74047450	74072812	1
chr9	78265749	79363459	6
chr9	82122201	82458081	6
chr9	88631382	89562780	2
chr9	89851757	89895179	2
chr9	89913351	89929183	2

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr9	94952413	94972608	1
chr9	97115045	97222114	2
chr9	98017523	98054378	1
chr9	99287408	99323101	3
chr9	99784385	100578200	5
chr9	100952173	101073946	1
chr9	101971834	102102062	1
chr9	103275595	103320436	1
chr9	105039896	105101236	1
chr9	113724364	113790200	1
chr9	116505560	116559949	5
chr9	116575939	116669446	5
chr9	117537697	117572629	1
chr9	120451898	120467575	1
chr9	123352391	123469149	2
chr10	12673308	13158787	6
chr10	18576412	18599061	1
chr10	20856662	21240959	6
chr10	21336616	21456381	6
chr10	22414769	22498551	1
chr10	29121558	31319020	5
chr10	33370382	33499988	1
chr10	45442835	45949076	1
chr10	49287284	49570649	1
chr10	53094579	53474603	1
chr10	57670381	58385076	3

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr10	59860032	59901785	1
chr10	63090729	63566390	4
chr10	63916014	63999416	4
chr10	65176000	65907225	4
chr10	66484381	66496754	4
chr10	67778523	67848471	5
chr10	68428706	68530522	6
chr10	70189857	70278973	1
chr10	94042413	94291954	6
chr10	94760500	94842463	6
chr10	113097293	113317527	3
chr10	114260580	114459569	6
chr10	120874005	120899587	1
chr10	122719244	122817988	2
chr10	123257592	123366786	2
chr10	125008071	125021047	1
chr10	125364650	125369612	2
chr10	127696020	127874456	1
chr10	129130523	129576002	1
chr11	9254914	9654261	5
chr11	12359217	12376235	1
chr11	16316165	16704107	2
chr11	19366905	19388954	1
chr11	19745689	19826745	2
chr11	20138917	20373463	2
chr11	32251789	32260071	3

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr11	33477192	33516248	2
chr11	34530806	34839363	2
chr11	36989390	37120535	1
chr11	38638332	38936692	1
chr11	43363295	43560852	2
chr11	45173972	45259808	3
chr11	46343122	46498066	1
chr11	47852547	47929980	2
chr11	56715732	56783582	2
chr11	57354601	57471211	3
chr11	59673426	59847964	4
chr11	61529642	61649434	2
chr11	64953373	65080737	1
chr11	79354925	79477356	6
chr11	88074454	88086405	1
chr11	89930194	89945604	1
chr11	101957722	102041593	1
chr11	112795454	112806898	2
chr11	118123864	118200216	2
chr12	5607574	5825539	5
chr12	7934457	8065680	5
chr12	10368143	10664225	1
chr12	12865426	12882714	1
chr12	15761503	15907832	1
chr12	16439596	16484716	1
chr12	16848381	17088344	1

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr12	21227429	21293968	2
chr12	21301073	21540990	2
chr12	25602328	25635938	1
chr12	26178703	26213691	1
chr12	27681656	27843146	3
chr12	28274058	28525448	2
chr12	46050131	46107461	1
chr12	47004172	52801104	6
chr12	55904391	55944077	1
chr12	72230639	72399587	3
chr12	73033898	73094781	3
chr12	80625917	80693905	2
chr12	82678096	82761689	3
chr12	87802366	87854970	1
chr12	88455806	88510346	3
chr12	104182081	104549339	6
chr12	105261310	105283831	1
chr12	106494569	106616144	1
chr12	107471288	107535472	4
chr12	107990241	108053721	5
chr12	108733545	108809475	5
chr12	109889707	109991280	1
chr12	110822148	111011261	2
chr12	111300667	111770896	4
chr12	114144765	114355913	1
chr13	3006383	3154216	4

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr13	6519522	6602929	1
chr13	7270154	8852461	4
chr13	10621147	10768233	4
chr13	11627696	11758133	1
chr13	11910651	12162516	3
chr13	12572502	12710810	3
chr13	21202706	21341570	1
chr13	24489470	24608387	6
chr13	26430350	26494528	6
chr13	30564937	31151142	3
chr13	33296668	33946085	3
chr13	39851811	40127905	1
chr13	45810147	45856715	2
chr13	46137238	46315848	3
chr13	58415226	58472991	6
chr13	81734956	81889306	2
chr13	94396646	94593193	4
chr13	96411427	96420039	2
chr13	106021620	106451201	6
chr13	111542458	112988657	1
chr13	114114782	114271785	5
chr13	116207594	116372364	4
chr13	117594234	117705034	3
chr14	9997796	10082417	3
chr14	14244067	14299474	1
chr14	18812531	19665896	1

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr14	24528903	24569746	3
chr14	26584758	26682614	6
chr14	36946756	37546476	6
chr14	49991679	50087785	1
chr14	50614419	50660460	1
chr14	51943526	52100863	1
chr14	55447168	55657865	1
chr14	56726169	56999146	4
chr14	73127236	73212809	1
chr14	76798626	77943037	2
chr14	78006644	78041511	1
chr14	78054538	78229035	1
chr14	79439630	79465828	1
chr14	80075116	80605426	1
chr14	82886335	83052813	1
chr14	84929572	85436866	1
chr14	87430524	87502083	2
chr14	90843740	91143988	1
chr14	92747930	93076534	3
chr14	93827593	93872540	1
chr14	94371476	94492884	1
chr14	100169598	100382837	1
chr14	101429928	102056948	5
chr14	103423933	103588202	2
chr14	103741947	103814568	1
chr14	104309780	104319442	1

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr14	107101844	107543841	6
chr14	109216563	109678203	4
chr14	118028620	118059161	1
chr14	118387630	118616001	5
chr14	120200423	120341644	2
chr14	120739887	120908375	2
chr14	121628279	121694654	1
chr15	5893299	6133680	2
chr15	7262673	7355432	4
chr15	9556487	9582907	2
chr15	12229932	12861881	4
chr15	18691961	18823709	2
chr15	23017246	23169800	1
chr15	27306913	27355983	5
chr15	31163027	31282763	1
chr15	34310857	34502664	1
chr15	39913652	40666791	6
chr15	43326974	43373656	2
chr15	48668812	49199769	3
chr15	52011270	52214326	2
chr15	53490374	53694443	3
chr15	53952779	54139966	1
chr15	54674975	55112405	1
chr15	57234458	58260100	4
chr15	58860985	59271697	2
chr15	59301035	59453993	2

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr15	59555290	59674241	2
chr15	68521784	68574287	3
chr15	68889893	69007206	3
chr15	73569014	73630529	2
chr15	74717806	75058612	4
chr15	76406625	76453389	1
chr15	76874009	77439918	4
chr15	81820505	82796482	6
chr15	82805161	82826179	4
chr15	84538670	84650896	4
chr15	86349999	86413110	2
chr15	88319728	88347337	2
chr15	88377699	88417900	4
chr15	90263138	90569476	5
chr15	91985286	92021424	3
chr15	97753072	97801211	1
chr16	10421605	10546455	2
chr16	21914196	22115249	2
chr16	23510310	23558738	3
chr16	25075665	25104321	6
chr16	30100338	30253924	6
chr16	33221329	33276917	1
chr16	34293738	34340081	1
chr16	34498423	34533652	1
chr16	35590620	36614225	4
chr16	36627028	37742524	2

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr16	39004590	39049570	1
chr16	39055641	39113348	2
chr16	39326500	39445706	1
chr16	45147676	45215090	2
chr16	45560023	45577653	2
chr16	45638232	45682168	2
chr16	48190353	48473543	2
chr16	70663065	71717436	6
chr16	75870379	76177210	1
chr16	79024995	79852213	5
chr16	81930244	82434926	2
chr16	84304936	84510966	1
chr16	92163926	92197856	4
chr17	7494637	7980497	5
chr17	10040029	10263477	5
chr17	12982603	13214245	2
chr17	17889531	17919437	5
chr17	26401090	26473662	1
chr17	26605415	26638742	1
chr17	28382844	28499186	2
chr17	28982150	29043181	3
chr17	30131275	30518126	4
chr17	31333730	31445804	3
chr17	32791791	32857602	2
chr17	33715984	34578132	3
chr17	35783680	35977112	6

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr17	36268903	36502140	2
chr17	37541841	39436355	6
chr17	41231540	41318321	2
chr17	50600628	51383615	6
chr17	53096697	53139339	1
chr17	53608962	53629520	1
chr17	54336045	54381844	1
chr17	57041453	57200869	4
chr17	58810428	58896599	1
chr17	59204103	59353125	1
chr17	60505342	60730356	1
chr17	70226579	70246066	2
chr17	70937721	70999095	1
chr17	71434140	71474892	2
chr17	72582112	72591812	5
chr17	75289490	75322146	4
chr17	75880993	75927813	2
chr17	77420080	77655975	1
chr17	78661272	78691954	1
chr17	78783769	78869822	6
chr17	79297591	79360979	2
chr17	79715751	79741878	2
chr17	79923398	81111315	5
chr17	81917703	81938789	1
chr17	82637328	83132530	4
chr17	89381593	89585125	1

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr18	3544234	4170283	6
chr18	6592919	7443167	6
chr18	18575064	18863652	4
chr18	40375616	40464164	1
chr18	40493047	40853784	6
chr18	41481339	41843531	2
chr18	44018089	44073650	1
chr18	46580928	46915862	1
chr18	48688039	48774180	6
chr18	49028799	49136501	6
chr18	54761069	54882118	5
chr18	59214746	59326508	2
chr18	59429079	59988480	3
chr18	61102673	61108972	2
chr18	62627558	62707520	1
chr18	65396055	65412266	6
chr18	65963609	66229756	3
chr18	66477503	66618707	2
chr18	68651471	68748337	2
chr18	69107611	69123980	2
chr18	70759190	70808926	1
chr18	80270534	80392283	2
chr18	80435813	80525231	2
chr18	81340648	81348604	3
chr18	85573711	85605317	5
chr19	3839708	3941876	2

Continued on next page

Supplementary table 11: – *Continued from previous page*

Chromosome	Chrom_start	Chrom_end	Frequency
chr19	7712026	7755895	1
chr19	8842181	10131677	5
chr19	11655783	11681544	1
chr19	13894203	14008855	2
chr19	14260830	14755340	2
chr19	16866135	16905395	6
chr19	18334111	18809860	2
chr19	18865786	18981402	2
chr19	23423831	23470605	3
chr19	25078150	25134803	2
chr19	26704273	26740604	2
chr19	27020705	27093978	1
chr19	34708402	34807673	1
chr19	37241520	37457000	1
chr19	37761016	38039361	4
chr19	47637141	47716150	2
chr19	49525279	49965745	2
chr19	52177107	52528030	6
chr19	57493921	57579376	4

Affidavit

I here by confirm that this thesis constitutes my own work and that I wrote it independently except for advises given by my supervisor and that all sources applied are listed and specified in the thesis. Furthermore, I confirm that this work has never been submitted as part of another examination process and has not yet been submitted for publication. This work has been undertaken in compliance with the Max-Planck institute rules of good scientific practice.

Plön, 22.07.2014

Hiba Mohammed Ali Babiker

Curriculum Vitae

Personal Information

Name:	**Hiba Mohammed Ali Babiker**
Birthdate:	**13.10.1980**
Birthplace:	**Khartoum**
Nationality:	**Sudanese**
Address:	**Danziger Strasse 7a** **24306 Plön, Germany**

School Education

1986-1992	**Halfayat El-Muluk Primary school, Khartoum, Sudan**
1992-1995	**Hassan Basheer Elementary school, Khartoum, Sudan**
1995-1998	**High school certificate, Al-ain High School** **Al-Ain, United Arab Emirates**

Higher Education

1998-2002	**B.Sc Cell and Molecular Biology, United Arab Emirates University, Al-Ain, United Arab Emirates**
2008-2010	**M.Sc Biology, Uppsala University, Uppsala, Sweden**
2011-present	**PhD student, Max-Planck Institute for Evolutionary Biology and Kiel university, Germany**