Unveiling the murine t-haplotype’s extent and emergence of diversity in MHC class II genes

Bachelor's thesis

in the Biology single-subject bachelor program

of the Faculty of Mathematics and Natural Sciences of the
Christian-Albrechts-University of Kiel

submitted by

Meri Nehlsen

First examiner: Prof. Dr. Tal Dagan

Second examiner: Dr. Linda Odenthal-Hesse

Kiel, September 2022

Table of contents

Index of abbreviations

AH	Ahvaz-Region in Iran
bp	Base Pairs
CAST	*Mus musculus castaneus* subspecies
CB	Cologne-Bonn-Region in Germany
cM	Centimorgan
CO	Crossover
DNA	Desoxyribonucleic Acid
DOM	*Mus musculus domesticus* subspecies
DSB	(DNA) Double Strand Break
Hba4ps	Hemoglobin alpha, pseudogene 4
HPLC	High Performance Liquid Chromatography
kb	Kilobase Pairs
KH	Almaty-Region in Kasachstan
Mb	Megabase Pairs
MHC	Major Histocompatibility Complex
mL	Milliliter
µL	Microliter
MUS	*Mus musculus musculus* subspecies
NCO	Non Crossover
ng	Nanogramm
PBC	Peptide Binding Cleft
PCR	Polymerase Chain Reaction
SNP	Single Nucleotide Polymorphism
Tcp-1	T-complex protein 1
V	Volt
ZnF	Zinc Finger

Abstract

Genes of the major histocompatibility complex locus that present pathogen peptides to T-Lymphocytes are among the most polymorphic genes in mammals. Within the major histocompatibility complex diversity is under positive selection and new alleles are generated by point mutations and recombination events. In some house mice (*Mus musculus*) however the major histocompatibility complex (called H-2) is part of the t-haplotype, a meiotic driver located on Chromosome 17 carried by 10 to 40 percent of the natural population. Meiotic drivers are selfish chromosomal arrangements defined by the deviation of Mendelian ratio of inheritance, meaning that they are over-proportionally transmitted to offspring. As such purifying selection on meiotic drivers is reduced and deleterious mutations can accumulate although outside of lethal alleles few non-synonymous mutations are observed. Additionally meiotic drivers often show strong linkage disequilibrium which is the result of reduced recombination between the wildtype chromosome and the chromosome carrying the genetic driver often caused by structural variations between chromosomes such as inversions. In this thesis, the physical extent of the t-haplotype inversions is resolved using optical mapping at higher resolution than ever done before. Evidence for a fifth inversion in the t-haplotype of *Mus musculus domesticus* was found. Also the allelic diversity of the MHC class II gene *H2Aa* in t-haplotype individuals of the subspecies *Mus musculus musculus* and *Mus musculus domesticus* was described based on sanger sequenced exons. The degree of diversity found here indicates that recombination between the t-haplotype and the wildtype might play an important role in diversifying this *H2Aa* exon. Lastly, to uncover *de novo* meiotic recombination events within the *H2Aa* gene and the *Prdm9* ZnF Array, which determines the placement of meiotic hotspots, Nanopore Sequencing was implemented. To identify the original template of sequenced amplicons the Primer ID as presented by Jabara et al., 2011 as included in the Primers for amplifying gene regions to be sequenced. However due to low coverage with Primer IDs no definite *de novo* recombination events could be defined, leaving the use of Primer IDs in Nanopore Sequencing up to discussion.

Zusammenfassung

Gene des Haupthistokompatibilitätskomplexes, die T-Lymphozyten Pathogenpeptide präsentieren, gehören zu den polymorphsten Genen bei Säugetieren. Innerhalb des Haupthistokompatibilitätskomplexes unterliegt das Allel-Reichtum einer positiven Selektion, und neue Allele werden durch Punktmutationen und Rekombinationsereignisse erzeugt. Bei einigen Hausmäusen (*Mus musculus*) ist der Haupthistokompatibilitätskomplex (H-2 genannt) jedoch Teil des t-Haplotyps, eines meiotischen Treibers auf Chromosom 17, der in 10 bis 40 % der natürlichen Population vorkommt. Meiotische Treiber sind egoistische Chromosomen Elemente, die sich durch eine Abweichung von den Mendelschen Vererbungsregeln auszeichnen. So wird der t-Haplotyp überproportional häufig an die Nachkommen einer Tägers vererbt. Dadurch wird die negative Selektion auf meiotische Treiber reduziert, und es können sich schädliche Mutationen ansammeln, obwohl abgesehen von letalen Mutationen nur wenige nicht-synonyme Mutationen beobachtet werden. Darüber hinaus weisen meiotische Treiber oft ein starkes Kopplungsungleichgewicht auf, das das Ergebnis einer verringerten Rekombination zwischen dem Wildtyp-Chromosom und dem t-Haplotyp Chromosom ist. Oft wird dieses Kopplungsungleichgewicht verursacht durch strukturelle Variationen zwischen Chromosomen wie Inversionen. In dieser Arbeit wird die physische Ausdehnung der t-Haplotyp-Inversionen mittels optischer Kartierung (“optical Mapping”) mit höherer Auflösung als je zuvor aufgeklärt. Hinweise für eine fünfte Inversion wurden im t-Haplotyp von *Mus musculus domesticus* gefunden. Darüber hinaus wurde die allelische Diversität des MHC-Klasse-II-Gens *H2Aa* in t-Haplotyp-Individuen der Unterarten *Mus musculus musculus* und *Mus musculus domesticus* auf der Grundlage von Sanger-Sequenzierungen der Exons beschrieben. Der hier gefundene Grad der Diversität deutet darauf hin, dass die Rekombination zwischen dem t-Haplotyp und dem Wildtyp eine wichtige Rolle bei der Diversifizierung dieses *H2Aa*-Exons spielen könnte. Um schließlich meiotische *De-novo*-Rekombinationsereignisse innerhalb des *H2Aa*-Gens und des *Prdm9* ZInk Finger Mikrosatelliten aufzudecken, welcher die Platzierung der meiotischen Hotspots bestimmt, wurde die neue Nanopore Sequenzier-Methode eingesetzt. Um die ursprüngliche Vorlage für die sequenzierten Amplikons identifizieren zu können, wurde die von Jabara et al. (2011) beschriebene Primer-ID in die Primer für die Amplifikation der zu sequenzierenden Genregionen aufgenommen. Aufgrund der geringen Abdeckung mit Primer-IDs konnten jedoch keine eindeutigen *De-novo*-Rekombinationsereignisse definiert werden, so dass die Verwendung von Primer-IDs beim Nanopore-Sequenzieren zur Diskussion steht.

1. Introduction

The idea of selflessness is the prerequisite for multicellular organisms as in order for the organism to function all cells and functional elements have to submit their fate to the greater cause within the organism. Our understanding of fundamental mechanisms in biology is often based on this assumption. However there are genetic elements that escape these rules of selflessness in inheritance. These selfish genetic elements challenge our understanding of inheritance and consequent evolution of genes. One such element is the murine t-haplotype, a meiotic driver which locks about 900 genes in a linkage disequilibrium and proposedly isolates them from evolutionary forces caused by chromosome-scale structural variation. The Major Histocompatibility Complex (MHC) is one of the regions encompassed by the t-haplotype. As the classical MHC genes, coding for cell surface proteins that present pathogenic peptides to T-lymphocytes, are subject to special evolutionary forces resulting in exceptional polymorphism, studying them can provide insight on how genes evolve on the t-haplotype.

1.1. History Of The *Mus musculus* Subspecies

The house mouse (*Mus musculus,* L.) is thought to have originated in South Asia, with the highest diversity reported in Iran (Fujiwara et al., 2022). The divergence of the three main subspecies, *Mus musculus musculus* (MUS), *Mus musculus domesticus* (DOM) and *Mus musculus castaneus* (CAST), is currently dated to 130.000 to 500.000 years ago and thought to have happen almost simultaneously (Phifer-Rixey et al., 2020). Still MUS and CAST are considered sister species to DOM within the subspecies (Phifer-Rixey et al., 2020). As the evolution of the house mouse is intertangeld with the human advance and agriculture especially MUS and DOM are thought to have undergone a very recent bottleneck and consequent expansion shaped by human influence (Fujiwara et al., 2022). Currently the MUS subspecies populates central and northern Europe as well as the northeast of Asia, whereas DOM is found in western Europe, along the mediterranean coast of Africa and the Arabian peninsula. The CAST subspecies is found in central and southeast Asia (see Figure S1). The population structure of these subspecies mainly consists of smaller locally isolated groups (demes) with little gene flow to other demes (Linnenbrink et al., 2018).

As the *Mus musculus* subspecies are still in the process of divergence hybrid zones at contact points are formed, most prominently in the hybrid zone between MUS and DOM in central europe. Introgression between subspecies is common and evidence for this is found in large distances to the hybrid zone likely due to influence of human transportation (Ullrich and Tautz, 2020).

1.2. History Of The t-Haplotype

The t-haplotype is a meiotic driver that is naturally found in a fraction of 10 to 25 percent of the house mouse population and located on the proximal third of the chromosome 17 within a region spanning around 40 million base pairs and including around 900 genes (Kelemen et al., 2022). This region is characterized by four inversions relative to the wildtype chromosome 17. As a meiotic driver the t–haplotype is most prominently characterized by a transmission ratio distortion (TRD) in inheritance, as among the offspring of male t-haplotypes about 90 percent also carry the t-haplotype (Lyon, 2003). The TRD is limited to males as the driver impares motility of wildtype

sperm. Multiple distorter loci of the t-haplotype act in trans to overactive a pathway controlling Sperm Motility Kinase (SMOK), which results in abnormal flagellate movement and loss of sperm motility (Herrmann et al., 1999). As this potentially affects both t-haplotype and wildtype sperm, a responder has to act to restore sperm motility in t-haplotypes. Specifically this responder is a gene, comprised of the promoter and exons of a SMOK-member and the 3´-UTR of tripled ribosomal s6 kinase 3 gene, which mediates partial cis resistance to the overactivation by the distorters to a certain degree (Herrmann and Bauer, 2012). Thus the TRD increases the t-haplotypes fitness only if both distorter and responder are present on the same chromosome since the t-haplotype is not innately immune to its own distorter (Lyon, 1986). This need for linkage is thought to be the drive for the emergence of four inversion the t-haplotype carries in comparison to the wildtype.

The primal evolutionary origin of the t-haplotype however is up to debate. Generally the t-haplotype chromosome 17 is thought to have been introgressed into the *Mus* subspecies from a small transient population of an unknown species (Morita et al., 1992; Silver et al., 1987). However there are two possible explanations for the creation of the t-haplotype before this introgression (Hammer et al., 1989). Firstly the TRD associated with the t-haplotype could have arisen in the original population and the inversions followed later as a selective advantage of linking distorter and responder (Charlesworth and Hartl, 1978; Hammer et al., 1989). In the ancestral population the TRD is thought to have been much less pronounced than in the current *Mus musculus* population (Herrmann and Bauer, 2012). In this form already featuring the structural variations the t-haplotype could have been introgressed into the *Mus musculus* subspecies. Secondly Introgression of an at least partially inverted into the *Mus musculus* subspecies could have caused TRD by chance leading to the further rise of inversions locking responders and distorter in a linkage disequilibrium to secure TRD (Hammer et al., 1989; Silver, 1982, p. 198). This possibility is supported by the fact that the second inversion of the t-haplotype is not present in *Mus musculus* wildtypes and their closely related sister species *Mus abbotti* but is found in *Mus spretus* (Hammer et al., 1989). Therefore the ancestor of all three species could have featured an inversion that would have been lost in the lineage to *Mus musculus* and *Mus abbotti* and from species of this ancestral line the t-haplotype could have been introgressed in already inverted form (Hammer et al., 1989). As the evolutionary origin of the t-haplotype remains unclear it is difficult to estimate the age of the t-haplotype both inside and outside the *Mus musculus* species. Additionally complicating this resolution, there are discrepancies between the different inversions of the t-haplotype when it comes to dating their origin indicating that inversions arose time-shifted. For instance the third and fourth inversions show more variability in sequence divergence than the second inversion for example (Dod et al., 2003; Kelemen and Vicoso, 2018). Thus different estimates of the origin of certain inversions paint an contradictory picture (Kelemen and Vicoso, 2018). The origin of the first inversion is dated to 3 Million years ago, while the fourth inversion is thought to be much younger at 1.5 Million years (Hammer and Silver, 1993). Both of these estimations place the origin of the t-haplotype inversions long before the divergence of the *Mus musculus* subspecies around 130.000 to 500.000 years ago (Phifer-Rixey et al., 2020; White et al., 2009). When it comes to loci of the t-haplotype very little sequence divergence is found between different t-haplotypes and subspecies considering the proposed age of their inversions (Hammer and Silver, 1993; Morita et al., 1992). For example there is only one known *Prdm9* allele for the t-haplotype, although *Prdm9* is known to be one of the most rapidly evolving genes in mammals and widely diverse in the *Mus musculus* subspecies (Buard et al., 2014; Kono et al., 2014). Similarly serological analysis of t-haplotype MHC class I alleles disclosed few differences (Artzt et al., 1985).

Despite the fact that the linkage disequilibrium and TRD essentially isolate the t-haplotype from selection, outside of recessive lethal mutations little evidence is found for a high degree of non-synonymous mutations on the t-haplotype (Artzt et al., 1982; Kelemen et al., 2022; Paterniti et al., 1983). The *Tcp1*-Intron which similarly to *Prdm9* is the same for all t-haplotypes and is commonly used for identification of t-haplotypes, indicates that rapid introgression of the t-haplotype might have taken place as recent as 7000 to 9000 years ago and been supported in its spread by human influence in trade and agriculture (Morita et al., 1992). Proposedley only after this spread population specific lethality factors have accumulated explaining the discrepancy between the number of differing lethality factors and the low number of non-synonymous mutations otherwise (Kelemen and Vicoso, 2018; Morita et al., 1992). The t-haplotype therefore paints a confusing contrast between supposed exposition to unhindered accumulation of potentially harmful point mutations and high degree of similarity among t-haplotype genes.

1.3. Structure Of The t-Haplotype

Interest in the structure of the t-haplotype first arose as suppression of recombination between the t-haplotype and wildtype chromosome 17 was observed. First theories that t-haplotype chromatin might intrinsically be unable to undergo recombination were disproven when (Silver and Artzt, 1981) observed recombination between two different t-haplotypes even if the same t-haplotypes rarely recombined with the wildtype chromosome. Following Artzt et al., 1982 found that the murine MHC (H-2) not only was included in this region of recombination suppression but also mapped differently to different regions between the t-haplotype and the wildtype. While they suspected that the recombination suppression within the H-2 might be a general feature of the murine MHC the order of genes on the t-haplotype clearly differed from the wildtype (Artzt et al., 1982). Namely the H-2 was placed between the characteristic t-haplotype marker genes T and tf instead of after tf (Artzt et al., 1982). Finally Shin et al., 1983 found evidence for an inversion including the H-2 region in t-haplotype mice in analysis of restriction fragment length polymorphisms. Artzt, 1984 was able to pin the breakpoint of the inversion to a position between Tcp-1b and cld in proximal direction and TL and Ce-2 in distal direction. A second inversion including T and qk was found just two years later (Herrmann et al., 1986). One reason why inversions as the cause of the recombination suppression were not discovered early is that in electron microscopy first no inversion were visible (Womack and Roderick, 1974). In 1988 Mary Lyon and colleagues could show that the inversion breakpoints were simply in bands of the same intensity and thus hard to discern in electron microscopy. Additionally it became clear that the t-haplotype extended further across chromosome 17 than thought before, laying the foundation for the currently accepted length of about one third of the chromosome or around 40 million base pairs (Kelemen et al., 2022). When crossing *Mus musculus domesticus* wild- and t-haplotypes with their close sister species *Mus abbotti* and *Mus spretus* Hammer and colleagues found two additional inversions within the t-haplotype, termed centromeric and middle by them, with the order of the inversion on the chromosome being centromeric, proximal, middle, distal (Hammer et al., 1989).

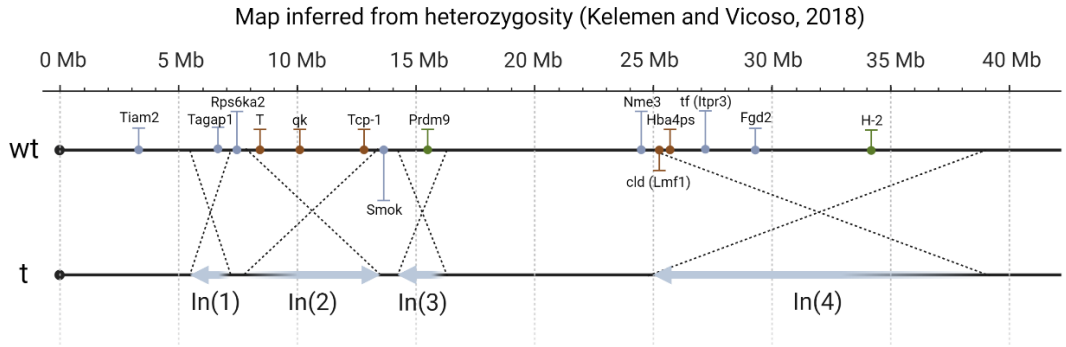

Figure 1: Established map of the proximal third of t-haplotype chromosome 17. Upper chromosome represents the wildtype as a reference for structural variation of the t-haplotype (represented in the lower chromosome). Extent of inversions is based on SNP heterozygosity data as described in (Kelemen and Vicoso, 2018). Positions of markers were derived from the GRCm38/mm10 reference genome for *Mus musculus domesticus* found at the UCSC genome browser (https://genome.ucsc.edu/). Direction of inversions is relative to directions in the ancestor, as the second inversion of the t-haplotype represents the ancestral state shared with *Mus spretus* (Herrmann et al., 1986). Distorter and responder genes of the t-haplotype are marked by gray identifies. Loci historically used for identification of the t-haplotype inversions are emphasized in red, while additional loci important for this research are accentuated in green.

With these new inversions the region of recombination suppression of the t-haplotype was mostly found to be inverted, suggesting that there are no further big inversions to be found. Notably within each inversion loci enhancing the TRD of the t-haplotype were found, meaning that for the maximal TRD all four inversion had to be present (Hammer et al., 1989). Since then efforts have been made to map the t-haplotype chromosome at higher resolution. Kelemen and Vicoso, 2018 took an interesting approach in mapping the t-haplotype using data for Single Nucleotide Polymorphism Heterozygosity along Chromosome 17 of t-haplotype individuals (Harr et al., 2016). This map is shown in Figure 1. However still to this day the exact structure of the murine t-haplotype has not been recovered in this entirety.

1.4. The Murine MHC

The Major Histocompatibility Complex (MHC) is a genomic region in which many genes mostly associated with adaptive immune response are located. The murine MHC is also called H-2 and located within the proximal third of chromosome 17 therefore H-2 is part of the fourth inversion in t-haplotype carriers (Shin et al., 1983). Most prominent members of the MHC are the classical MHC class I and class II genes, which code for proteins that present peptide antigens on the cell surface to be recognized by T lymphocytes (Figure 2). This function in initiating the T lymphocyte mediated immune response is likely the cause of their vast diversity within species as MHC class I and class II genes are regarded as part of the most polymorphic mammalian genes. Class I and class II genes are differentiated based on the cell they are expressed in and what type of T lymphocytes they present to, the origin of the peptides they present, the size of presented peptides as well as their structure (Stuart, 2015). MHC class II proteins present peptides derived from the extracellular room on the surface of professionally antigen presenting cells such as dendritic cells or macrophages. Peptides presented are less rigid in regard to their length with up to 30 amino acid oligomers being presentable. Murine MHC class II genes are *H2Aa*, *H2Ab*, *H2Ea* and *H2Eb*. When expressed, MHC class II proteins form Heterodimers (H2AA with H2AB and H2EA with H2EB) through noncovalent

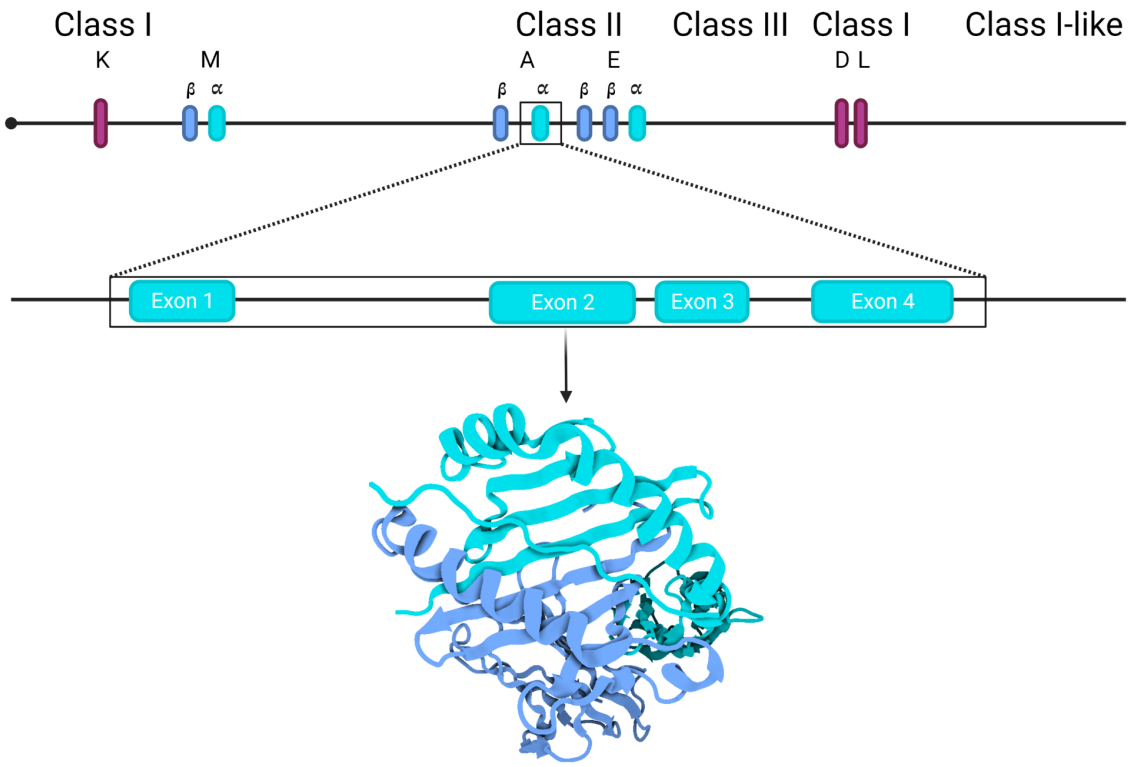

Figure 2: An overview of the murine MHC located on chromosome 17. Most prominent members of the classical MHC class I and class II genes are shown in their relative position on chromosome 17. Class II molecules are heterodimers of two chains coded in two separate genes. Within these genes the second exon corresponds to Peptide Binding Cleft in the protein, which binds the pathogenic peptide. Protein 1ES0 (10.2210/pdb1ES0/pdb) is used here to represent conformation of the PBC of an MHC class II molecule (Corper et al., 2000). This Figure was created on BioRender.com.

interactions (Stuart, 2015). Since MHC molecules are able to bind processed peptides in order to present them to T-lymphocytes, the conformation of the MHC molecules determines which pathogenic peptides and thus which pathogens can be recognized (Liu and Gao, 2011). As cell surface proteins MHC class II molecules are heterodimers formed by two protein chains each consisting of one small intracellular domain, one transmembrane and two extracellular domains. These two chains are called alpha and beta chains and are coded by the correspondingly named genes (Liu and Gao, 2011). Most notable for diversity within the MHC genes is probably the Peptide binding cleft (PBC) to which pathogenic peptides bind. The Peptide binding cleft of MHC class II molecules is formed by the n-terminal domains of both protein chains with each chain contributing 4 beta-strands and one alpha-helix to form the PBC (Liu and Gao, 2011; Roitt and Delves, 2001). Within the PBC the 8 beta-strands are arranged parallel to each other and form the floor of the cleft with the two alpha-helices acting as borders to each side of the floor (Liu and Gao, 2011) (Figure 2). In contrast to MHC class I molecules the MHC class II PBC is open at both ends where the alpha helices meet in the MHC class II molecule. As the interaction of the MHC molecule with the bound peptide is limited to the PBC, this region is the most polymorphic between different alleles, where as the second extracellular domain forms a homolog to the immunoglobulin superfamily and is very conserved among different alleles of a gene. Within the PBC of the MHC two protein-peptide-interactions determine the binding of peptides(Liu and Gao, 2011). Conserved residues within the MHC molecule along the whole PBC form hydrogen bonds with the backbone of the bound peptide while oftentimes

polymorphic residues of the PBC confine specificity of the MHC binding peptides (Liu and Gao, 2011; Madden et al., 1992; Roitt and Delves, 2001).

1.5. Selection For MHC Polymorphism

The vast diversity of MHC alleles is unmatched in mammals although the causing mechanism is up to debate (Shiina et al., 2017). On a molecular level there are mainly two forces that act to create diversity of genes: namely point mutations and recombination events such as crossovers and non-crossovers (Meyer and Mack, 2008). However these forces do not impact the MHC more strongly than the rest of the genome, as mutation and recombination rate are not significantly raised within the MHC in comparison to the remaining genome (Penn and Musolf, 2012). Still positive Darwian selection seems to favor nonsynonymous mutations especially within the Peptide Binding Cleft of MHC molecules, as the ratio of nonsynonymous to synonymous mutations is much higher in the region of the PBC than for the rest of the MHC gene or the genome (Hughes and Nei, 1988; Yeager and Hughes, 1999). Consequently, diversity within the MHC is thought to be the result of natural selection rather than abnormal mutation or recombination rates (Penn and Musolf, 2012). Interestingly MHC allele frequencies are fairly homogen for different Alleles, meaning that the diversity is not created by a few alleles making up only a small part of the genpool but rather by the maintenance of different alleles over longer timespans (Nadeau et al., 1988).

Following the hypothesis of pathogen-mediated overdominant selection, individuals with different MHC alleles or haplotypes show higher fitness, because diversity increases the likelihood of presenting and therefore recognizing a pathogen (Doherty and Zinkernagel, 1975). Although this is a probable explanation it has to be noted that the exceptional diversity of MHC molecules is not found for other loci involved in the reception of pathogens (Penn and Potts, 1999). So while pathogen mediated selection might be an important factor in explaining MHC diversity, additional factors are likely to influence selection too (Penn and Musolf, 2012). Besides pathogen-mediated selection some theories have offered sexual selection as the explaination for the MHC diversity. Evidence that MHC alleles have an influence on an individual's odor which can translate into an influence in mate choice has been presented for the house mouse (Penn and Potts, 1999; Yamazaki and Beauchamp, 2007). Especially in the natural house mouse population inbreeding is problematic as populations exist in small isolated demes that rarely experience gene flow between them (Linnenbrink et al., 2018). Therefore MHC based mate choice can be used to avoid inbreeding and reduce inbreeding depression (Potts and Wakeland, 1993). Additionally ensuring offspring heterozygosity can positively influence fitness of offspring against pathogens (Penn and Potts, 1999). However it is necessary to note that pathogen-mediated and sexual selection do not have to be mutally exclusive in explaining maintainance of MHC diversity (Penn and Musolf, 2012).

Still the population structure of house mice which would lend itself to inbreeding and high degrees of homozygosity is a puzzling factor considering the diversity of murine MHC molecules and the fact that the vast majority of mice is heterozygous at the MHC even in closed demes (Duncan et al., 1979; Linnenbrink et al., 2018). This contrast is not explainable by overdominant selection on its own (Linnenbrink et al., 2018). Linnenbrick et al., 2018 accordingly offered the reservoir model as an alternative explanation for the MHC diversity in wild mice. The reservoir model states that MHC diversity is generated starting from ancestral alleles that are kept within the population over long time spans by balancing selection. Both mutations and recombination events act to generate new

alleles within the demes continuously explaining why even in closed demes high degrees of heterozygosity are kept.

1.6. Meiotic Recombination

In sexually reproducing organisms meiosis is a specialized cell division that results in the division homologous chromosomes of a diploid progenitor cell to form haploid gametes. The major advantage of sexual reproduction is the creation of diversity with each generation which meiosis ensures. Therefore meiosis is a highly regulated process to ensure correct division and generation of diversity between developing gametes. For both of these functions meiotic recombination is a main player. Firstly meiotic recombination regulates the pairing of homologous chromosomes, which is the prerequisite for their subsequent division. Secondly meiotic recombination enables the genetic exchange between homologous chromosomes which creates gametes whose chromosomes differ for the parental ones. During meiotic recombination genetic material can be exchanged in a reciprocal (crossover, CO) and non-reciprocal manner (non crossover, NCO). For correct segregation of homologous chromosomes at least one CO events has to be located on that chromosome pair as it contributes to the right orientation of the paired chromosomes in regard to the meiotic spindle (Arnheim et al., 2007; Baudat et al., 2013). However these CO events are not evenly distributed across the length of the chromosome but rather clustered in recombination hotspots which are regions of 1 to 2 kilobase pairs in length where recombination events are preferentially initiated (Arnheim et al., 2007). Each recombination event starts with a double stranded break (DSB) in the DNA. The mechanism of DSB preparation determines the nature of the recombination event and multiple different pathways for these mechanisms are known (Serrentino and Borde, 2012). However the localisation of all these events is determined by the localisation of the initiating DSB. The placement of these DSB is highly regulated and mainly determined by the protein PRDM9 (Úbeda et al., 2019). PRDM9 binds to certain DNA-motifs with an array of zinc fingers and methylates adjacent histones with a PR/Set-domain which induces a rearrangement of Histones to form Nucleosome depleted regions (NDR) (Baker et al., 2014, p. 9). Later the murine COMPASS-complex is recruited to these NDR additional to an interaction with PRDM9 through its KRAB-domain (Damm and Odenthal-Hesse, 2022). The Proteins of the COMPASS complex initiate the introduction of DSB into the DNA via the endonuclease SPO11 (Tock and Henderson, 2018, p. 11). The zinc finger array of PRDM9 that specifies location of recombination hotspots is subject to rapid evolution, which means that recombination hot spots evolve faster than the genes whose recombination they control (Baudat et al., 2013). However recombination hotspots have also been found to be influenced by the DNA sequences themself. Different murine MHC haplotypes display different placements and intensities, suggesting that they influence the “potential” of a recombination hotspot (Arnheim et al., 2007).

1.7. Meiotic Recombination Within The Murine MHC

Evidence for recombination events within the murine MHC have been found continuously. In fact the presence of recombination hotspots, regions of usually 1 to 2 kb where recombination occurs most preferentially, has been first observed within the murine MHC (Yoshino et al., 1995). Recombination on its own however does not automatically diversify and rather shuffles already diverse regions between homologous chromosomes to gain new combinations. In the example of the MHC recombination is actually proposed to lead to a homogenisation of introns (where additionally most

recombination hotspots are located) as most diversity is found within the exons (Cereb et al., 1997; Yeager and Hughes, 1999). One known recombination hotspot has been defined to an intronic region of the *H2Eb* gene (Yauk et al., 2003). Additionally evidence has been presented that recombination shaped the diversity of the PBC (She et al., 1991).

1.8. Meiotic Recombination Within The t-Haplotype Inversions

Recombination events are thought to be rare between the t-haplotype and wildtype due to the inversion of the t-haplotype. Generally while inversions suppress recombination, formation of CO is not impossible, evident as there are rare recombination events between wildtype and t-haplotype chromosomes observed (Silver and Artzt, 1981). On one hand there are partial t-haplotypes where one or several inversions of the t-haplotype are replaced by their wildtype equivalent. Partial t-haplotypes can be created by CO in the inverted regions between the t-haplotype inversions, by CO within inversion or by double CO events (Wallace and Erhart, 2008). On the other hand there are "mosaic" t-haplotypes, where within the t-haplotype inversion fragments of wild type alleles are found, which differ for different t-haplotypes (Wallace and Erhart, 2008). As this phenomenon can not be explained by the mutation rate in the t-haplotype (with respect to its recent introgression) this seems to be the result of recombination events, mainly gene conversions (Wallace and Erhart, 2008). While this has been found for genes within the middle of the fourth and biggest inversion, no evidence has been presented for recombination towards the ends of this inversion, namely where the H-2 complex is located. In meiosis inversions prevent the homologous pairing of chromosomes and lead to the formation of unsynapsed inversion loops (Torgasheva et al., 2013). Torgasheva et al., 2013 found that in female meiosis of house mouse inversion loops of an inversion of the first chromosome can be fully resolved only if a CO is formed in the middle of the inversion. This results in preferential recombination between uninverted and inverted regions within this middle. Because of this, inversion breakpoints are thought to show the highest recombination suppression of the whole inversion (Villoutreix et al., 2021).

2. Objective

The aim of this project is to identify the effects the proposed recombination suppression of the t-haplotype has on the evolution of the MHC and if there actually still might be recombination events happening within the t-haplotype MHC. Firstly the structural variation present in the t-haplotype is to be analyzed in order to determine the relative position of relevant loci within the t-haplotype chromosome 17. The main thing to observe would be the extent of the t-haplotype inversion and compare them to previous approximated maps at the resolution optical mapping offers. Additionally any differences between the t-haplotypes of *Mus musculus* subspecies DOM and MUS are to be described. Therefore optical maps of mice of these different *Mus musculus* subspecies were analyzed.

Secondly the natural diversity of *H2Aa* alleles with a focus on the PBC in t-haplotypes is to be characterized. For some genes there is only a singular t-haplotype associated allele. Defining if diversity in the MHC which is so strongly favored by selection still can be still found in t-haplotypes could offer thus evidence to resolve the nature of selection on the t-haplotype. For this purpose the second exon of MHC class II gene *H2Aa* was sanger sequenced and phased using downstream nanopore sequencing data, combining both methods to gain accuracy from sanger sequencing and phased base calling from nanopore sequencing.

Lastly from nanopore sequencing data potential recombinants along the whole of MHC class II gene *H2Aa* and additionally of the ZnF Finger array of *Prdm9* are to be discovered. This analysis might lend insight on how recombination suppression of the t-haplotype affects the MHC and *Prdm9*. Based on the limited serological diversity of MHC class II genes and the singular *Prdm9* allele of the t-haplotype, recombinants between haplotypes are expected to be rare. For this purpose long-read sequencing of the *H2Aa* gene as well as the Zinfinger Array of *Prdm9* is to be performed for somatic and potentially recombinant sperm DNA using Primer IDs in amplification to be able to infer origin of each sequenced read and reduce the error of nanopore sequencing.

3. Materials and Methods

3.1. Origin Of Studied Mice

Mice in this study were treated following the local laws, namely the German Animal Welfare Law (Tierschutzgesetz) and the Experimental Animals Ordinance (Versuchstierordnung). However the mice studied here were captured for another experiment and the DNA that had already been extracted was simply reused for this study (Table 1).

Table 1: List of mice used for this study. For each mice the subspecies and the catching location (see Supplementary Figure S1) as welll as the sex is given. The mice not assigned a number were used only in Optical Mapping

No.	Mouse ID	Subspecies	Strain	catching location	sex
WM	50026607	MUS	KH	Almaty, Kasachstan	male
TD1	50027236	DOM	AH	Ahvaz, Iran	male
TD2	50028955	DOM	AH	Ahvaz, Iran	male
TM3	50036273	MUS	KH	Almaty, Kasachstan	male
TM4	50036275	MUS	KH	Almaty, Kasachstan	male
TD5	50036599	DOM	CB	Cologne-Bonn, Germany	male
Ttp4	50099943	unknown	T/tp4	inbred strain	male
	50049461	MUS	KH	Almaty, Kasachstan	male
	50080530	DOM	CB	Cologne-Bonn, Germany	female
	50096581	unknown	Tp4	inbred strain	male
	50101580	DOM	CB	Cologne-Bonn, Germany	male
	50101764	DOM	CB	Cologne-Bonn, Germany	male
	50102342	DOM	AH	Almaty, Kasachstan	male

3.2. Extraction And Preparation Of DNA

DNA had previously been extracted from ear by a standard Phenol-Chloroform extraction and from sperm following (Cole and Jasin, 2011). Concentration of extracted DNA was determined using a Nanodrop Spectrophotometer (Thermo Fisher Scientific). All DNA dilutions were made in HPLC-grade water. Both DNA stocks and dilutions were stored at -20°C and thawed for use.

3.3. Preparation Of Primers

Primers were bought from Integrated DNA technologies in lyophilized form. Primer stocks were set up upon arrival by adding HPLC-grade water following the manufacturer's instructions to reach an end concentration of 100µM. Working primers (10µM) were diluted from the stock using HPLC-grade water. Both working primers and primer stocks were stored at -20°C and thawed at RT for use.

3.4. Gel Electrophoresis

Loading Dye (30% (v/v) glycerol with enough bromophenol to add blue color in 5X TBE-Buffer) was added to the cycled PCR product in 1:5 ratio. For gel electrophoresis gels were casted with Top Vision Low Melting Point Agarose (Thermo Fisher Scientific) at 0.8% (w/v) or 1.0% (w/v), TAE-Buffer (diluted from 50X TAE (ROTH)) and 0.02µL SyberSafe per 1mL of TAE. Probes were loaded into the gel leaving an empty pocket between each sample to avoid cross-contamination. As a size standard Quick-Load® Purple 1kb Plus DNA ladder was added once for each row of pockets.Gel electrophoresis was run in cooled TAE-Buffer at 120V for 30 to 60 minutes, depending on the size of the gel and needed resolution using a BIO RAD Laboratories Power Pack™ Basic Power supply. Gel was visualized and photographed in a Molecular Imager® Gel Doc XR+ Trans-Illuminator (BIO RAD laboratories) using the Image Lab imaging system (BIO RAD Laboratories). Bands matching the expected size were excised using a scalpel and transferred to individual Safe-Lock Tubes (Eppendorf). Excised bands were stored at 4°C until further use. On a heating block (Thermomixer comfort, Eppendorf) tubes containing excised bands were heated at 70°C for 10 minutes to melt the gel. Following the tubes were cooled on the same block to reach 42°C. Once temperature was stable at 42°C 2µL (2U) Agarase (Thermo Fisher Scientific) was added to each tube for every 100mg of a 1% Low Melting Point Agarose gel. Tubes were left to incubate at 42°C and 300 rpm for 30 minutes. Treated bands were stored at 4°C until further use.

3.5. Sanger-Sequencing

4 µL of treated bands were used in cycle-sequencing with Big Dye Terminator (version 3.1 Thermo Fisher Scientific). For PCR reactions of 10µL 1.75µL of Big Dye Buffer, 0.5µL of BigDye and Primers each and 3.25µL HPLC-grade water were added to each well of a 100µL reaction plate (Applied Biosystems MicroAMp® Fast 96-Well Reaction Plate) followed by 4µL of the treated bands. Cycling was performed in a Veriti PCR machine (Applied Biosystems), starting with an initial denaturation step of 96°C for 1 minute. Denaturation step at 95°C for 10 seconds, annealing step at 55°C for 15 seconds and elongation step at 60C for 4 minutes were run for 40 cycles. Plate was cooled at 12°C until removed from the cycler. The PCR product was purified using the BigDye XTerminator™ purification kit (Thermo Fisher Scientific). 50µL of Mixture of SAM-Solution and XTerminator™, as described by manufacturer, were added to each well. Plate was shaken at 2000 rpm for 30 minutes before being centrifuged at 1000g for 2 minutes. Finally the plate was submitted for sanger sequencing in a 3500 Series Genetic Analyzer (Thermo Fisher Scientific).

3.6. t-Haplotype Genotyping

Presence of t-haplotype chromosome 17 was determined via Alternative Fragment Length Polymorphism. Microsatellites of loci Tcp-1 and Hba4ps were amplified using the Qiagen Multiplex Kit and about 40 ng DNA as input according to the manufacturer's guide. After the amplification in

PCR, the product was diluted with 100 µL HPLC grade water and left to incubate at room temperature for 30 Minutes. After incubation 10µL of 1:100 ROX 500 to HiDi mixture was added to 1µL of diluted PCR product in a plate. Mixture was denatured in a Thermocycler at 90°C for 2 minutes followed by an incubation at 20°C for 5 minutes. The plate was removed from the thermocycler and immediately transferred onto ice before submission for fragment analysis in the genetic analyzer. Raw data was analyzed in Geneious Prime using the Microsatellite Plugin.

3.7. *Prdm9* Genotyping

Prdm9 genotyping was carried out using established primers (Buard et al., 2014; Kono et al., 2014). PCR reactions were carried out in 15µL reactions each in a 100µL reaction plate (Applied Biosystems MicroAMp® Fast 96-Well Reaction Plate). As a Buffer for PCR 1x AJJ (modified from Jeffreys et al., 1988 and described in table 1) and 12.5mM Tris-Base (Trizma-Base, Sigma-Aldrich) were used. Primers were added at 0.5µM. Two DNA Polymerases were used to combine the amplification speed of 0.025 U/µL Taq DNA Polymerase Recombinant (Thermo Fisher Scientific) and the 3' → 5' exonuclease proof-reading ability of 0.0033 U/µL Cloned Pfu DNA Polymerase AD (Agilent Technologies). 0.75 µL of DNA dilution (20 ng/µL) were added to supply 15 ng of DNA per PCR Reaction. Cycling was carried out using a Veriti PCR machine (Applied BIosystems) and started with an initial denaturation step at 96°C. Three step cycling was performed 30 times, consisting of an denaturation step at 95°C for 15 seconds, an annealing step at 55°C for Kono Primers and 56°C for Buars Primers for 20 seconds and an elongation step at 70°C for 2 minutes. After three step cycling a final elongation step at 70°C was added for 5 minutes. The plate was cooled at 12°C until removed from the cycler. Cycled PCR product was stored at 4°C until further use. The PCR product was subjected to electrophoresis and prepared for sanger sequencing as described before. After Sanger-Sequencing raw reads of Genetic Analyzer were assembled and analyzed in Geneious Prime (de-novo assembly with default settings). *Prdm9* motif prediction was done at http://zf.princeton.edu/logoMain.php (Persikov et al., 2009; Persikov and Singh, 2014).

3.8. Genotyping Of The Second *H2Aa* Exon

PCR amplifying MHC exons was carried out similarly to *Prdm9* PCRs with only a few differences. Firstly Annealing temperature was 57°C for *H2Aa* primers. Secondly cycling was performed 40 times instead of thirty. Again PCR product was prepared for sanger sequencing after gel electrophoresis. Raw reads were assembled in Geneious Prime and trimmed to start and end with the respective primer sequences. Reads were phased by hand using Nanopore sequencing data as a guide. Phased reads were aligned using ClustalW with default settings in Geneious Prime. Phylogenetic tree was built within Geneious Prime using the Neighbour-joining method (HKY substitution model) and bootstrapped 100 times to create a consensus tree. Variability of sites within the exon were calculated using the Shannon entropy equation (Reche and Reinherz, 2003; Shannon, 1949) as followingly:

$$V = \; - \sum_{i=1}^{20} P_i \, log_2(P_i)$$

Where V is the Variability assigned to each site and Pi is the portion of amino acids i at the site.

3.9. Primer Design For Nanopore Sequencing

Primers of Primary DNA consisted of four elements (Figure 3). Firstly a sequence specific for the target gene was used to amplify the wanted gene. Secondly a Primer ID of eight random nucleotides followed upstream of the sequence specific primers. Thirdly a sequence to which the universal

primers M13 could bind was placed upstream of the Primer ID to be used as a template in secondary PCR which was run with M13 primers. Fourthly a driver consisting of four Guanosine-Nucleotides was placed in between the sequence specific primer and the Primer ID. The function of this driver was to evade a bias in the Primer ID for stronger nucleotides (such as Guanosine and Cytosine). Primer ID can be used to infer the original template of a PCR amplicon if PCR was performed one a probe with DNA molecules of different nature or sequence (Jabara et al., 2011). For Next generation sequencing where output is higher but also errors accumulate, this can be used to generate a consensus of amplicons originating from the same template (Zhou et al., 2015). By generating this consensus, errors in sequencing (which should be random and therefore differ between sequenced amplicons) are minimized, leaving a proposedly error free consensus sequence.

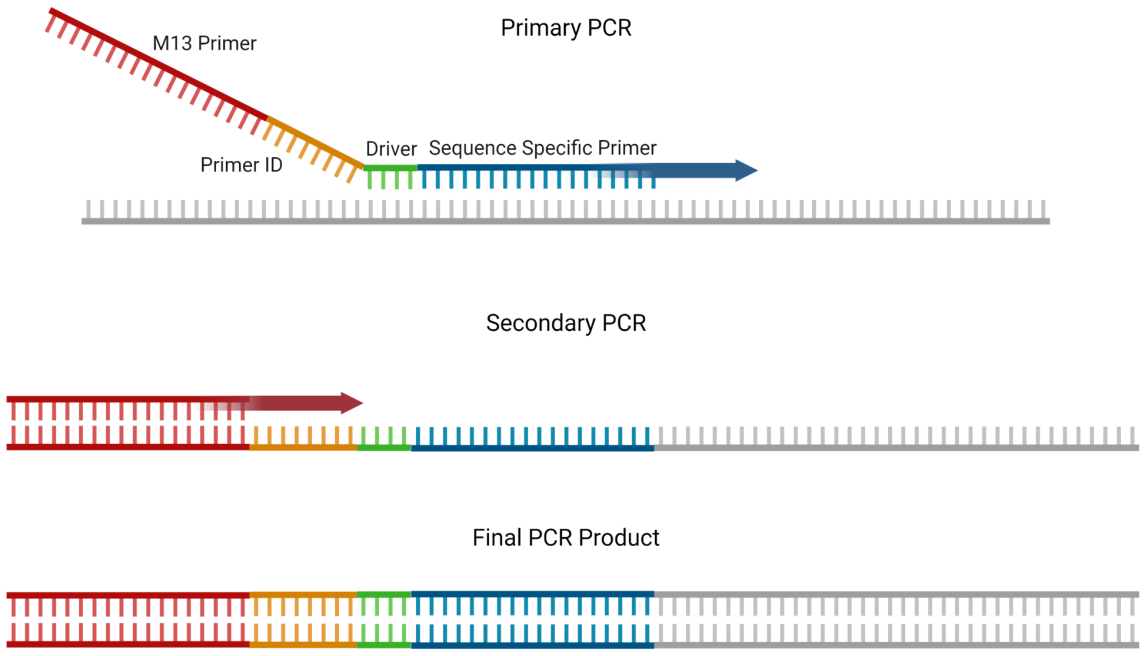

Figure 3: Overview of primer design. Primers of primary PCR were designed to attach Primer ID as well as template for M13 Primer to the amplicon generated by a Sequence specific primer. A driver was added to separate sequence specific primer and Primer ID to minimize nucleotide bias in Primer ID. Secondary PCR utilized the attached M13 Primer as a template for amplification using M13 primers, meaning that for each amplicon Primer ID attached in Primary PCR was included. The final PCR Product should therefore include the M13 Primer followed by the Primer ID. FIgure was created on BioRender.com.

3.10. Nanopore Long-read Sequencing

MHC gene *H2Aa* was amplified in primary PCR using Primers with random Tag. PCR reactions were carried out in 20µL Reactions of 1x AJJ-Buffer and 12.5 mM Tris-Base. Primers were added at 0.5 µM each and a combination of Taq and Pfu Polymerase was used as described before. To each well 300 ng DNA were added. Cycling was carried out with the equipment described before starting with denaturing at 96°C for 1 and a half minutes. Cycles consisting of 20 seconds at 95°C, 30 seconds at 59°C and eight minutes at 65°C were performed 25 times followed by a final elongation at 67°C for 5 Minutes. Plate was cooled on the cycler at 12°C until further use. PCR Product was purified using Exonuclease I (Thermo Fisher) and Shrimp Alkaline Phosphatase to remove primer dimers and leftover nucleotides. For each sample 3µL of a Master mix (consisting of 0.12µL Exonuclease 1, 0.45µL Shrimp Alkaline Phosphatase and 2.43µL HPLC grade water) was added to 5µL of the primary PCR product. Mixture was left to incubate at 37°C for 20 Minutes followed by incubation at 80°C for 20 minutes in a Veriti Thermocycler. Purified Primary PCR product was used as a template for secondary PCR with M13 Primer, which was carried out in 15µL reactions with 1.5µL of primary PCR Product being added. Secondary PCR was run similarly to primary PCR the only exceptions being the number of cycles as 30 instead of 25 and the annealing temperature at 58°C. Secondary PCR product was run in gel electrophoresis where bands were cut out and treated as described before. Treated Bands were purified using AMPure XP Reagent (Beckmann Coulter Life Science). Beads were vortexed and added to treated bands at 0.6x of the volume of the bands. Tubes were incubated on a Hula Mixer for 5 minutes at room temperature. After Incubation tubes were placed on a magnetic rack to pellet the beads and remove supernatant. Pelleted beads were washed with 200µL of 70% freshly prepared ethanol 2 times, discarding the supernatant after each wash. Briefly, tubes were left to dry for about 30 seconds, before adding 22µL HPLC grade water to each tube. Tubes were removed from the magnetic rack and left to incubate for 2 minutes at RT before being placed on the rack again to pellet the beads. Supernatant containing the DNA was transferred into a Quali Low Retention tube (Kisker Biotech) and quantified in a Qubit 3 Fluorometer (Thermo Fisher Scientific) following the manufacturer's instructions. All probes were diluted to match the sample with the lowest concentration of DNA. Samples were prepared for MinION1KB (Oxford Nanopore Technologies) sequencing according to the 1D Native barcoding genomic DNA (with EXP-NBD104, EXP-NBD114, and SQK-LSK109, all Kits from Oxford Nanopore Technologies) provided by the oxford nanopore community (version from August 14th, 2019) expect that volumes were halved for end preparation and native barcode ligation. Equal parts in volume of each probe were used for preparation for MinION, in a way that the sum of all samples reached the maximum input of 1µg of DNA. Prepared library was loaded onto a FLO-MIN106 (R9.4.1) flow cell (Oxford Nanopore Technologies) as described in protocol and the sequencing run was started with direct base calling enabled.

For *H2Eb* and *Prdm9* genes workflow was kept the same with following exceptions. Only 20ng DNA was added into primary PCR. Annealing temperature was 59°C (*H2Eb*)/ 54°C (*Prdm9*) and in Primary PCR only 20 cycles were completed. Additionally elongation times in both primary and secondary PCR were 12 minutes and two minutes respectively. The Primary PCR product was used as a template in four separate secondary PCR reactions, which were pooled after Cycling and directly Purified with AMPure XP Reagent as described before.

3.11. Analysis Of Nanopore Sequencing Data

Base-called data of nanopore sequencing was demultiplexed by barcode using the EPI2ME Desktop Agent (Oxford Nanopore Technologies) and split into seperate Folders by barcode. Minimum quality score for data was 7 for first run and 10 for second run. After Barcoding reads were assembled to a reference consisting of the target gene extended by the primary Primer to include the Primer ID. The Minimap2 plugin in Geneious Prime (Dotmatics) was used for mapping. If parental Haplotypes were unknown, the first three SNPs were manually searched for with the criteria being an approximate 50:50 ratio excluding reads with errors or deletions. The most common combinations of the first three SNPs were defined as the Parental Haplotypes and extracted from the alignment into separate files. Complete sequence of each Haplotype was generated in the form of a consensus sequence of each file. Followingly sequences where the last two SNPs matched with the other haplotype were extracted as potential recombinants. For these sequences each switch between Haplotypes was defined as the first clear SNP of the opposing haplotype, after which there is no more than one SNP of the original Haplotype among the next two SNPs. This means that sites that show deletions or the two bases not represented in the haplotypes can not be used to define the position of change. For each sequence the position of the first SNP of the other Haplotype was noted for further analysis. For breakpoint distribution the centimorgan per Megabasepairs value was calculated as:

$$\frac{cM}{Mb} = \left(\frac{n_{potential\ recombinants}}{n_{total\ sequence\ number}} \div \frac{SNP_2}{SNP_1}\right) \times 1000000$$

with $n_{potential\ recombinants}$ describing the number of sequences where the change of between Haplotypes could be first defined at SNP_2.

3.12. Optical Mapping

Optical mapping employs the use of light microscopy for detecting large structural variation in the genome via fluorescent labeling of DNA at specific sequence or enzyme-reaction motifs. In contrast to both short and long read sequencing this ideally provides data contingent across a whole genome, enabling the identification of large structural variants and resolution of complex and/or highly repetitive genomic regions (Yuan et al., 2020). The Bionano Saphyr system forms the current gold standard of optical mapping, where the use of an Direct Labeling Enzyme (DLE-1) provides labeling of the DNA without fragmenting it as in restriction enzyme digestion (Neely et al., 2010). Paired with updated optical tools this enables an resolution of up to 500 bp while keeping the contiguenty of the optical mapping methodology (Yuan et al., 2020). To generate an optical map first a reference for the desired target is digested in silico to firstly generate a pseudo reference map, which can be used for mapping later and secondly to ensure sufficient density of labeling sites in the target. For the target HMW DNA has to be extracted in labeled under conditions minimizing breakage before being transferred onto chips specialized for optical detection of the DNA labels within a single molecule. After imaging via the optical mapping system the optical data has to be transposed into data of physical distance between labels. This data can be used for de novo assembly which is challenging as sources for the introduction of errors are various (e.g. missing labels in the DNA due to underlabeling via the enzyme, breakage of DNA in preparation, noise in imaging). However after this first assembly the data can be mapped to the map of the in silico digested reference to identify structural variants.

3.13. Statistics

Statistical tests were all performed on the VassarStats website (http://www.vassarstats.net/). Regressions for visual analysis of cumulative mutation frequency and breakpoint distribution of Nanopore Sequencing data was performed in GraphPad Prism 9.4.1. (GraphPad Software, LLC).
Tajima's D for analysis of selection on the *H2Aa* exon was completed by hand. Here d was defined as the difference of A and M with A being the sum of the number of polymorphisms between each of the sequences. M being defined as:

$$M = \frac{n_{polymorphic\ sites}}{n_{individuals}}$$

4. Results

4.1. t-Haplotype Genotyping

Presence of the t-haplotype was determined by Alternative Fragment Length Polymorphism (AFLP) of the marker genes *Hba4ps* and *Tcp1* (Hammer and Silver, 1993; Morita et al., 1993) (Table 1). Shorter fragments of both markers correspond to the wildtype chromosome while longer fragments comprise the t-haplotype chromosome. Only one wildtype mouse was included in this analysis (WM) while all other mice presented to be t-haplotype carriers. T-haplotype individuals of the different subspecies produced slightly different fragment lengths for the t-haplotype associated fragment of *Tcp1*. In MUS mice the fragment was determined to be slightly longer at 610, while in DOM the length was 608.

Table 2: t-haplotype makers identified by Alternative Fragment Length Polymorphism (AFLP). The marker *Hba4ps* is located within the fourth inversion of the t-haplotype, while the marker *Tcp1* is located within the second inversion. Shorter fragments are associated with the wildtype chromosome while longer ones characterize the t-haplotype.

mouse ID	Individual	*Hba4ps*		*Tcp1*	
		wt	t	wt	t
50026607	WM	198	No peaks	412	No peaks
50027236	TD1	198	214	412	610
50028955	TD2	198	214	412	610
50036273	TM3	198	214	412	608
50036275	TM4	198	214	412	608
50036599	TD5	198	214	412	610
50099943	Ttp4	198	214	412	608

4.2. *Prdm9* Genotyping

Prdm9 is thought to be essential in determining the localisation of recombination events (Úbeda et al., 2019). Based on this zinc finger array which mediates the binding specificity to the DNSdifferent *Prdm9* Alleles can be differentiated. Most crucial for the Zinc Finger composition are the amino acids that interact with the DNA, here at the positions -1, 3 and 6 relative to the beginning of the alpha-helix (Persikov et al., 2009). The number of known *Prdm9* alleles for *Mus musculus* is high and the majority of wild mice are heterozygous for *Prdm9* (Buard et al., 2014; Kono et al., 2014). As *Prdm9* is located within the inversion of the t-haplotype the singular associated allele carriers the same Zinc Finger Array for all t-haplotype carriers (mmt1) which paints a striking contrast to the natural diversity of *Prdm9* alleles (Kono et al., 2014). Here also in all t-haplotype the mmt1 is present, while the non-haplotype alleles differ (Table 2).

Table 3: *Prdm9* alleles for examined mice by their Zinc Finger Array. For each allele the composition of the Zincfingers by the three most variable amino acids at position -1, 3 and 6 relative to the beginning of the alpha-helix within the ZnF is given. Other amino acids differing are indicated in brackets in their relative position to the main three variable amino acids. Hashtag indicates the 13 bp deletion within the first ZnF found in the t-haplotype allele.

No	Zinc Finger array (first allele)	Zinc Finger array (second allele)
WM	DNE QDK QDK QDK ANQ ANQ QDK ANQ QNK ENQ QNK ANQ (W)QNQ (W)QDQ	DNE QDK QVQ QDK ANQ ESK ANQ QNK ANQ (W)QNK (W)QDQ
TD1	DNE Q(N)HQ QDK VVK Q(N)HQ AVQ Q(N)HQ QDK QDK ANQ QNK (W)QDQ	DNE# AS(V)Q ASQ Q(N)HK (I)ANQ TDK Q(N)HQ (I)ANQ TDK Q(N)HQ QDQ
TD2	DNE ANK QDK VNQ AVK VVQ ASK VVQ QDK QNK A(N)NQ (W)QDQ	DNE# AS(V)Q ASQ Q(N)HK (I)ANQ TDK Q(N)HQ (I)ANQ TDK Q(N)HQ QDQ
TM3	DNE QDK QDK ASK VVQ VNQ VNQ QDK ASK QNQ VSK VVQ QDK QNQ	DNE# AS(V)Q ASQ Q(N)HK (I)ANQ TDK Q(N)HQ (I)ANQ TDK Q(N)HQ QDQ
TM4	DNE QDK QVQ QDK ANQ ESK ANQ QNK ANQ (W)QNK (W)QDQ	DNE# AS(V)Q ASQ Q(N)HK (I)ANQ TDK Q(N)HQ (I)ANQ TDK Q(N)HQ QDQ
TD5	DNE Q(N)HQ QDK QDK QVK QVK AVQ ANQ AVQ AVQ QNK (W)QDQ	DNE# AS(V)Q ASQ Q(N)HK (I)ANQ TDK Q(N)HQ (I)ANQ TDK Q(N)HQ QDQ

4.3. Inversion Maps Of t-Haplotypes

Optical mapping was used to determine the position and extent of the inversions characterizing the t-haplotype. However calling these inversions in the presented data was problematic because the optical mapping generally produced smaller contigs within the region of t-haplotype inversions, the proximal third of chromosome 17, than for the rest of the chromosomes. Few contigs actually showed inversion in their entirety and especially regions directly adjacent to an inversion, where the inversion breakpoints are located, mapped very poorly to the reference making the exact determination of inversion start and end impossible. However the composition of the contigs when compared to the reference could be used to smaller regions where inversions of the t-haplotype faced each other or an uninverted region (see Supplementary FIgure S2). From the conformation and orientation of these contigs the extent of the t-haplotype inversion could at least be confined to a smaller possible region. Inferring Structural variation in this way, general maps for t-haplotypes of different *Mus musculus* subspecies could be made (Figure 4). Examined were one map of a MUS mouse, one map of a Ttp4 mouse and five maps of DOM mice of different populations (Cologne-Bonn, Germany and Ahvaz, Iran), where a consensus of all maps was generated. For the MUS t-haplotype, four distinct inversions were found. The placement of the fourth (distal) inversion in MUS t-haplotype generally matched the estimation Kelemen and Vicoso, 2018 made based on SNP heterozygosity although the extent is seemingly about 5 Million bp bigger in the optical mapping towards the proximal end of the inversion.

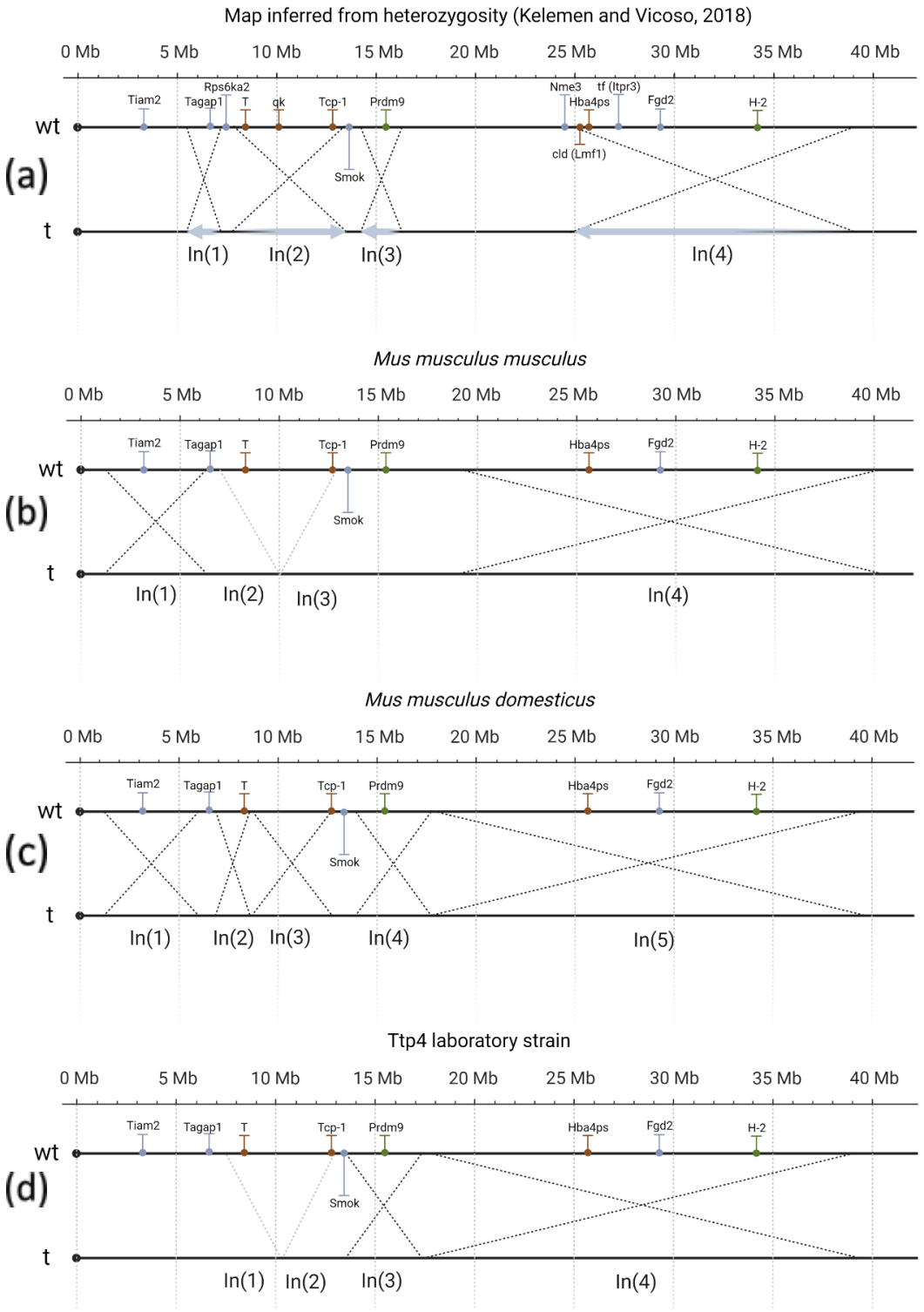

Figure 4: Position of t-haplotype inversions in *Mus musculus* subspecies MUS and DOM, as well as in laboratory strain Ttp4. All positions are relative to UCSC Genome Browser Reference GRCm38/mm10. The position of distorter and responder loci is marked in gray, loci used as markers in determining the t-haplotype or its inversions are displayed in red. The positions of the H-2 complex and *Prdm9* are shown in green. Extent of inversions were inferred from optical mapping data and mapped on the lower chromosome corresponding to the t-haplotype and are indicated by black dotted lines. Single gray dotted lines indicate inversion where only one breakpoint was visible and thus the extent of the inversion could not be determined. (a) Map of the t-haplotype inversions as determined by heterozygosity (Kelemen and Vicoso, 2018). Map of t-haplotype inversions in (b) MUS subspecies, (c) DOM subspecies and (d) Ttp4 laboratory strain.

Interestingly it seems that within the region of the second inversion of the heterozygosity map two inversions were found in the optical map, although their respective extent could not be inferred (see Supplementary Figure S2). Therefore mainly the first inversion of the optical map is placed at an anomalous position centromeric of the heterozygosity map´s first inversion. This has the consequence that in the *Mus musculus* t-haplotype *Prdm9* is not part of any inversion within the optical map. On the other hand this placement of the first inversion would probably include a recently discovered distorter of the t-haplotype *Tiam2* (Lindholm et al., 2019). Within DOM mice no relevant differences were observed based on the populations. Hence observations of all maps were combined to generate a consensus map. This consensus of DOM t-haplotypes showed five distinct inversion. Strikingly the first three and the last inversion were placed very similarly to the four inversions of the MUS respectively. Only the fourth inversion, containing *Prdm9*, of the DOM subspecies was not present in the sister subspecies. Although bigger in extent this inversion roughly corresponded to the proposed third inversion of Kelemen and Vicoso, 2018, meaning that here *Prdm9* was included in the inversion. Lastly the t-haplotype of the Ttp4 mouse matched the DOM t-haplotype in the four inversion that were found although the outer breakpoints of the first and second inversion relative to the reference of the Ttp4 strain could not be inferred as it had been the case for the MUS t-haplotype. Only the first inversion present in the DOM t-haplotype was missing in the Ttp4 t-haplotype. This similarity between the Ttp4 strain and the DOM subspecies also persisted outside the t-haplotype inversions and along the rest of the genome. As the reference the t-haplotype individuals were mapped to was based on the laboratory stain C57BL/6 the MUSs map featured numerous unmatched labels. The Ttp4 optical map however presented few unmatched labels and similar coverage to the DOM map.

4.4. Diversity Of The Second Exon Of *H2Aa* In t-Haplotypes

The second exon of MHC class II genes corresponds to the peptide-binding cleft (PBC) of the translated molecule, where most diversity is expected to be found between alleles of a gene (Cereb et al., 1997; Yeager and Hughes, 1999). To understand how the t-haplotype influences this diversity, the second exon of the H2-Aa of four t-haplotype individuals (TD1, TD2, TM3 and Ttp4) as well as one wildtype individual (WM) was analyzed.

4.4.1. Differences Between Alleles In SNPs

Phased alleles of the second *H2Aa* exon were compared for similarity in SNPs (Figure 5). Interestingly unlike for *Prdm9* no singular allele shared among t-haplotypes was identified. The number of differences in SNPs in the second of *H2Aa* varied mainly in a range from 10 to 20 SNPs with only three comparisons landing numbers of differences outside this range. One of these exceptions was the comparison of Ttp4 Allele A and TD2 Alle B with a total number of 22 differences. The comparisons with the fewest number of differences were WM allele B with TM3 Allele A and TD1 Allele A with TD2 allele B respectively.

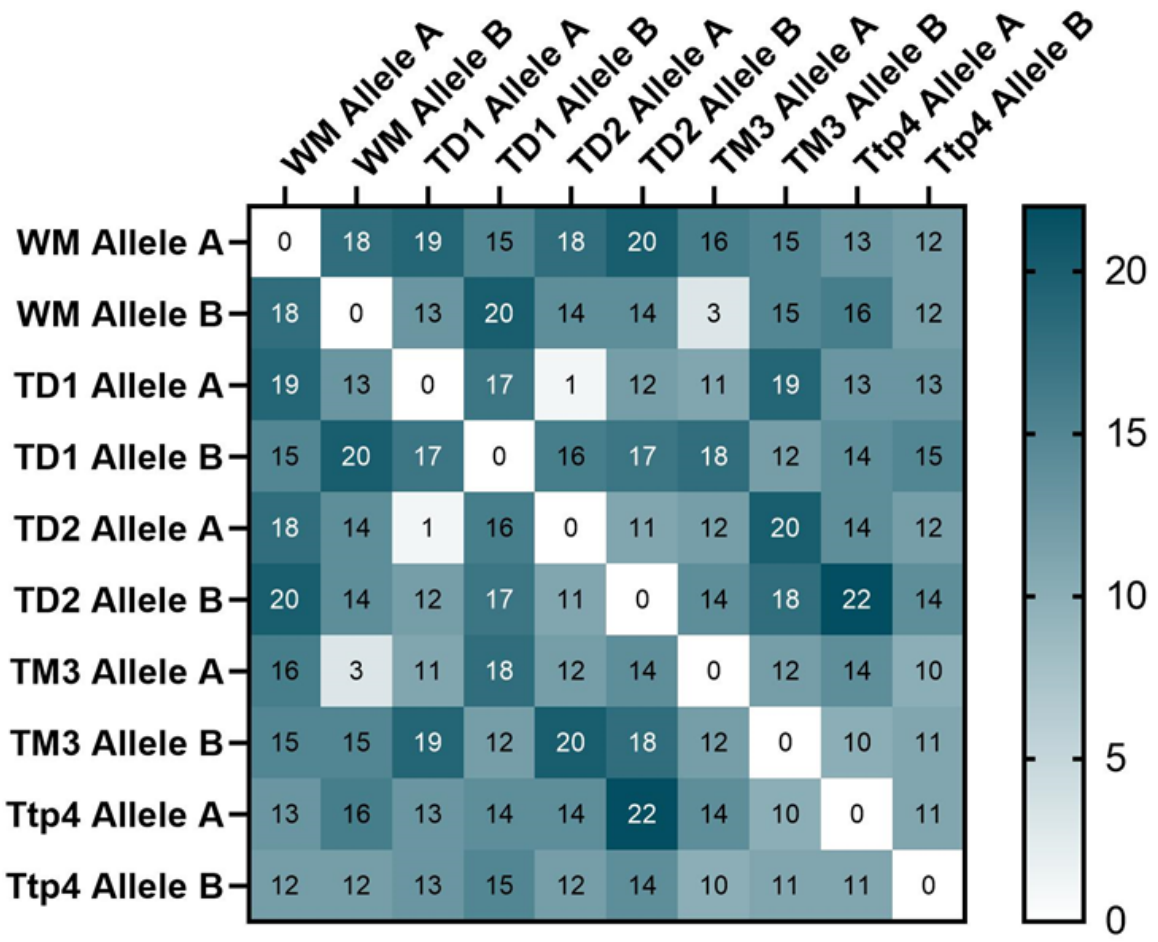

Figure 5: Heatmap of differences in Single Nucleotide Polymorphisms in the second exon of MHC H2-Aa gene. Exact number of differences between SNPs is given for each square of comparison, with lighter coloring of the square indicating less differences. For each heterozygous individual parental alleles are defined as Allele A and Allele B at random as it is not known definetively which allele belongs to the t-haplotype.

Within these pairs both mice originated in the same with the MUS mice located in KH and the DOM mice coming from AH. Between t-haplotype individuals no allele displayed no differences or just fewer differences than 10 for all t-haplotypes. The allele B of TM3 can be inferred to correspond to the t-haplotype as the allele A shows high similarity to one allele of the WM mice. However this allele did not demonstrate high similarity to any other allele of a t-haplotype with the minimum number of differences being 10 to the Ttp4 strain.

4.4.2. Distance Between *H2Aa* Alleles

When looking at phylogenetic distance of the MHC alleles found here there was no allele that all t-haplotype individuals shared and that could therefore be identified as the t-haplotype allele (Figure 5). In fact distances did not disclose relationships between alleles in a way that could be used to infer the t-haplotype alleles among the alleles observed. This is shown in the tree built based on the MHC alleles (Figure 6). Tree of genetic distance of the second exon of *H2Aa* gene was built using the

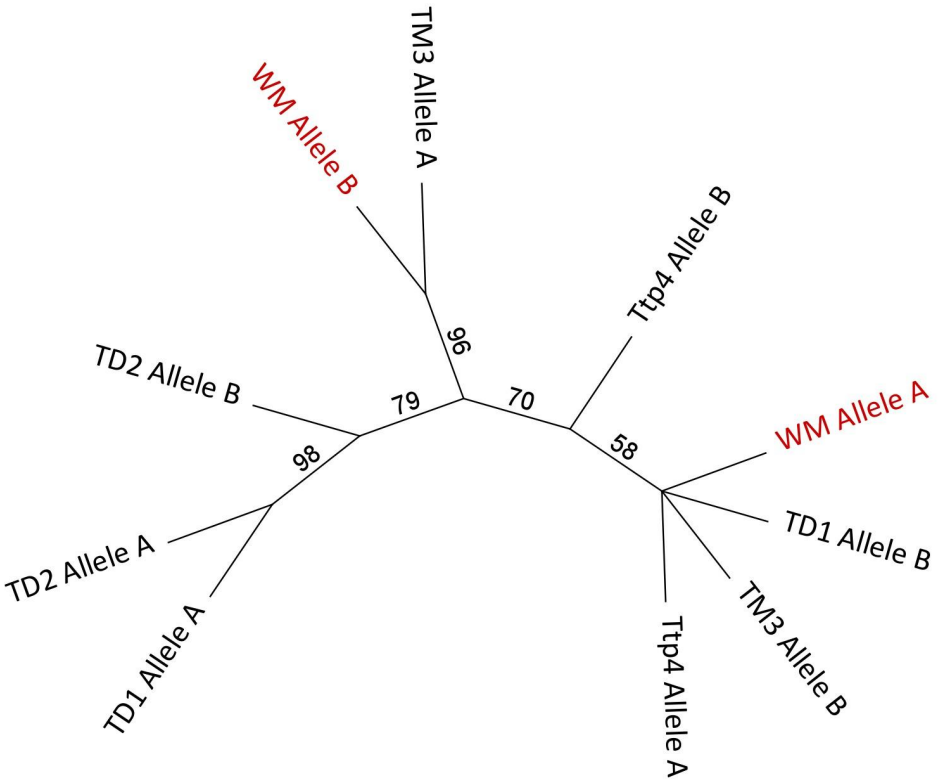

Figure 6: Tree of genetic distance between Alleles of the second H2-Aa exon (Neighbor-joining, HKY-Substitution-Model). Tree was built in geneious prime using default settings for Jukes and Cantor substitution model and Neighbor-joining Building method. Branch values indicate bootstrap support of 100 replicates. Alleles of the non-t-haplotype carrier are marked in red.

Neighbor-joining-method and HKY-Substitution model. As expected alleles with high similarity as shown in the distance matrix were paired. Notably however the proposed t-haplotype allele of TM3 (Allele B) shares a node with two other t-haplotype alleles but also one wildtype allele. While this subtree could not be resolved, it is interesting to note that only three of the four t-haplotypes are placed adjacent to each other here. In fact no node of the tree resolves the t-haplotypes in a way, where only they are grouped together. The tree does not disclose proximity between any set of alleles to be used to identify the t-haplotype allele. So not only are there several different MHC alleles for t-haplotype carriers (in contrast to the singular allele found for *Prdm9*) but within these alleles differences are great enough to make the identification of t-haplotype alleles based on phylogenetic distance indiscernible.

4.4.3. Nonsynonymous And Synonymous Mutations

The ratio of nonsynonymous mutations to synonymous mutations (d_n/d_s) can disclose the nature of selection affecting a gene. A d_n/d_s of values lower than one indicates that purifying selection shapes a gene, while a d_n/d_s higher than one indicates the opposite. The highest d_n/d_s ratio was found in the wildtype mice (WM) while the lowest d_n/d_s was present in the mouse of the inbred laboratory Ttp4 strain, which was a t-haplotype carrier (Figure 7). Even for this mouse however nonsynonymous mutations were on average more frequent than nonsynonymous mutations although one allele of

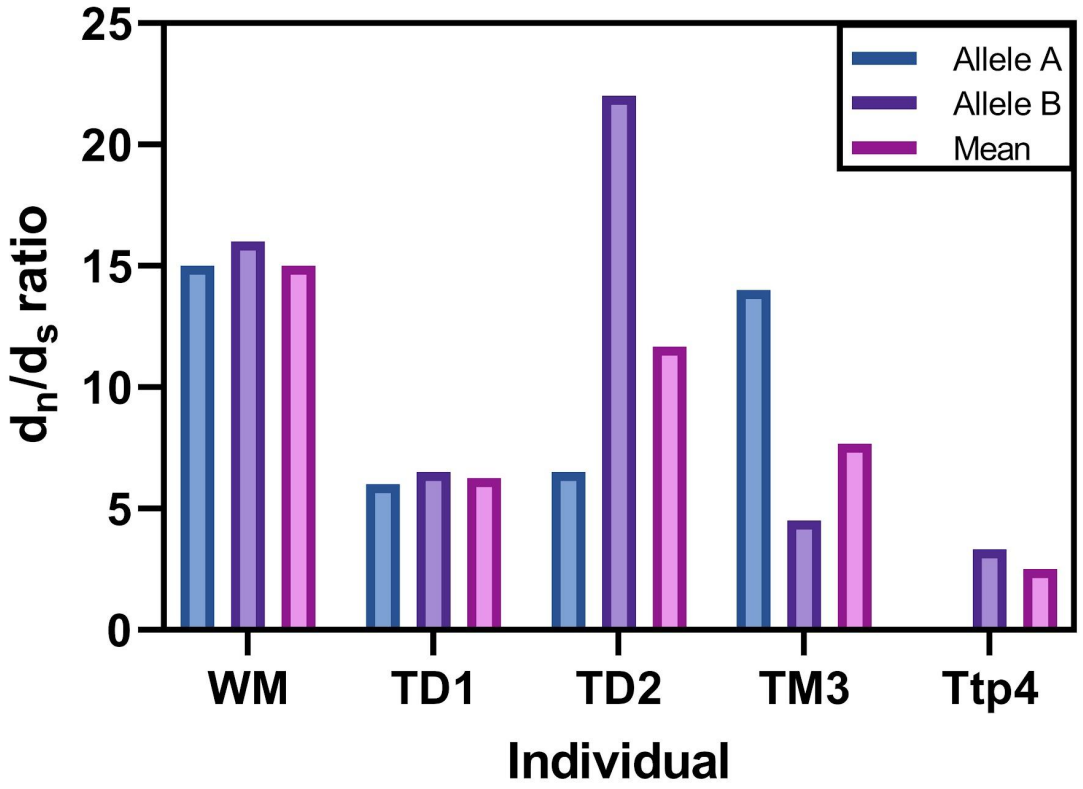

Figure 7: Ratio of nonsynonymous to synonymous mutations in the second exon of H2-Aa. For heterozygous t-haplotypes parental alleles are named allele A and allele B independent of their association to the t-haplotype. Displayed is the d_n/d_s ratio for the respective parental alleles as well as the average of these two alleles.

this mouse did not display any non-synonymous but only synonymous mutations. The d_n/d_s ratio of the wild t-haplotype carriers varied slightly. When comparing the d_n/d_s values for each of the individual alleles the wildtype WM and t-haplotype TD1 offered similar values for both alleles, while in the rest of individuals one allele had distinctively more nonsynonymous mutations than the other leading to a bigger difference in d_n/d_s ratio. The inferred t-haplotype allele B of TM3 (see 4.4.2.) showed the lower d_n/d_s ratio in comparison to the other allele indicating that the alleles of the other individuals with lower d_n/d_s ratio might correspond to the t-haplotype allele.

Tajima's D which allows assertions about the nature of selection for the three wild caught t-haplotype mice (TD1, TD2 and TM3) was 13.45 indicating positive selection on the second exon of H2Aa.

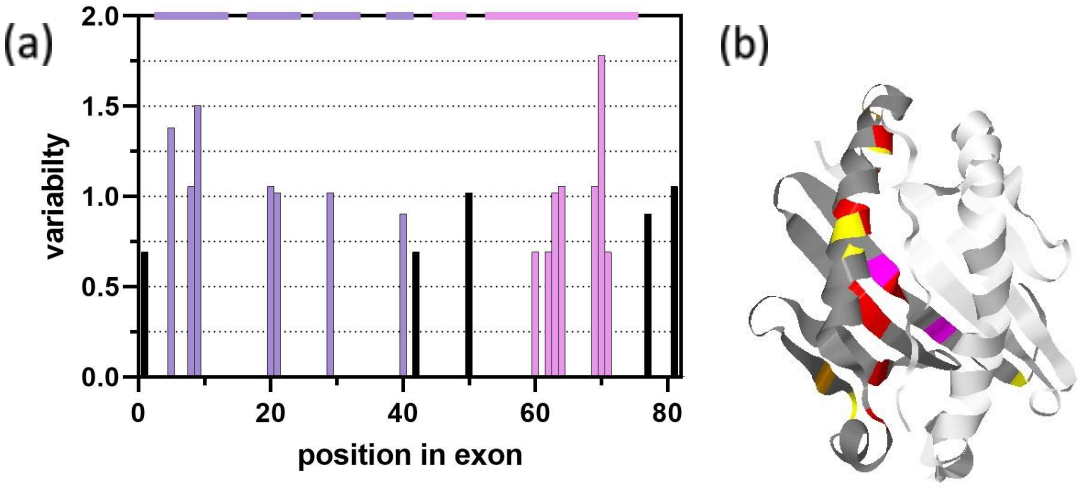

Figure 8: Amino acid variation within the PBC for five t-haplotype mice. Protein translations were generated in Geneious using the standard genetic code and annotated from protein 1ES0 (10.2210/pdb1ES0/pdb) retrieved from (Corper et al., 2000). (a) Variation calculated using the Shannon entropy equation is mapped against the respective position of the Amino acid in the PBC based on the corresponding region of the second exon. POsition 0 within the exon represents the C-terminal end. Based on the number of sequences analyzed the variation can take values in range from 0 to 3.32 with 3.32 representing the maximal variation (different amino acids within each sequence). Proposed secondary structures are marked with beta-strands highlighted in purple and alpha-helices highlighted in pink. (b) Within three dimensional conformation of the PBC of MHC class II molecule amino acids are marked by variability in the alpha chain of H2-Aa represented here by protein 1ES0 (10.2210/pdb1ES0/pdb) (Corper et al., 2000). Different colors indicate variability scores (V<0.8 shown as yellow, V between 0.8 and 1.0 shown as orange, V between 1.0 and 1.2 shown as red and V>1.2 shown as pink).

4.4.4. Localisation Of Variation Within The Peptide Binding Cleft In t-Haplotype Individuals

Variation within the PBC was distributed across the whole region corresponding to the second exon although variable residues were more tensely distributed towards the N-terminal end (corresponding to position 82 in the exon) than in the middle and C-terminal end (Figure 8). In total 19 of 82 amino acid residues differed between different proposed translations of the alleles. Whereas in the latter regions variable sites are mostly surrounded by non-variable sites, within the N-terminal end variable sites are directly adjacent to each other. The highest variation value (1.76) for a single site also is found on the N-terminal end. Within the translated protein the amino acids with the highest scores of variation (>1.2) two are located in the floor of the PBC formed by the four beta-stands while one is part of the alpha helix. The same number of variable sites was found for the beta-strands and alpha-helices.

4.5. Quality And Output Of Nanopore Long-Read Sequencing

The first Nanopore sequencing run for the *H2Aa* gene produced few reads (132,193 after Barcoding with the Epi2ME Desktop Agent, Oxford Nanopore Technologies) probably due to mistakes that were made in preparation of the samples. Subsequently the coverage with Primer IDs was low, with only 7.0 % of reads sharing a Primer ID (either in Forward or Reverse Primer) as seen in Figure 9. To raise recovery of DNA for the second run only a tenth of the DNA input of the first run into the Primary

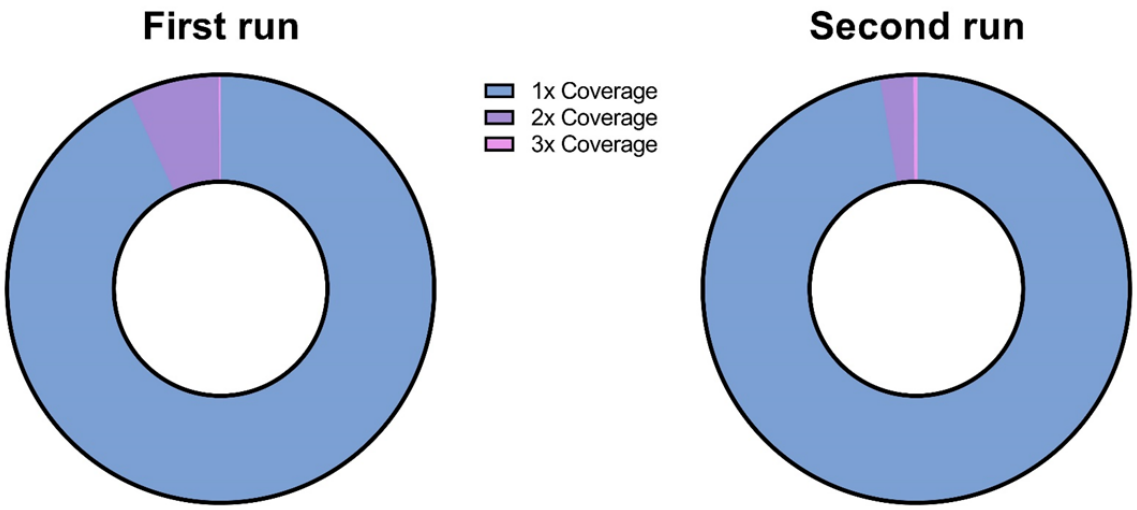

Figure 9: Coverage of sequenced reads by Primer IDs. First Nanopore Sequencing Run was used for analysis of the *H2Aa* gene (6kb). In the second Run the ZnF array of *Prdm9* (1kb) as well as the *H2Eb* gene (10kb) were sequenced. Unique IDs are defined as 1x Coverage, while ID found at least twice are classified as 2x Coverage. 3x Coverage refers to Primer IDs found thrice.

PCR was used as the AJJ PCR buffer is optimized for lower DNA concentrations. Additionally, the primary PCR product was separated into four secondary PCR reactions which were pooled before AMPure Purification. While in the second run for the *Prdm9* ZnF array and *H2Eb* gene the number of sequenced reads could be increased drastically (estimation of 10.23 Million reads in the MinKNOW software, Oxford Nanopore Technologies) coverage of Primer IDs could not be improved (Figure 9). In fact the percentage of duplicate Primer IDs decrease in the second run (2.5% as opposed to 6.79%) While for the second run only 3.232.000 reads (after Barcoding with the Epi2ME Desktop Agent, Oxford Nanopore Technologies) were analyzed because the number of templates in primary PCR had been decreased this loss of coverage is surprising. The overall quality of the reads the second run produced could be improved as the nanopore quality score increased from 8.7 to 9.44.

As the goal is to uncover CO events, sequences that display a change between Haplotypes determined by their respective SNPs are filtered for. As it is unclear whether these sequences represent actual COs in the following the change or switch between Haplotypes as denoted by their SNPs is termed as such or as mutation rather than as CO. Sequences displaying this switch are called potential recombinants or CO candidates.

4.6. Determination Of Recombination Events Within *H2Aa*

To uncover possible CO events in the H2-Aa gene of t-haplotypes, six different mice were chosen to be analyzed, among them 5 t-haplotypes. However only the wildtype mice (WM) produced heterozygous data.

Across the *H2Aa* gene 206 SNPs were found to differ between the haplotypes of the wildtype MUS individual (WM). These SNPs were distributed following the normal distribution (Shapiro-Wilk-Test, p is not significant) across the gene, indicating that there is no prevalence of SNPs for any region within the gene. In somatic DNA haplotype A was significantly overrepresented in the number of sequences in the MinION data (two-tailed approximation to binomial via normal distribution, $p<0.005$).

However this bias was not found for sperm DNA, where both haplotypes were represented equally. For both sperm and somatic DNA however the number of potential recombinants found for each haplotype differed significantly (Chi-Square-Test of Association with Yates-Correction, somatic DNA: $p<0.0001$; Sperm DNA: $p<0.05$), with Haplotype A being more abundant for both types of DNA (see Supplementary Table S2). Surprisingly the total number of potential recombinants did not differ significantly between somatic and sperm DNA (Chi-Square-Test of Association with Yates-Correction, $p>0.1$). Additionally their number did not differ by the factor 100 as expected following (Yauk et al., 2003, p. 2), where generally somatic crossover rate in H2Eb gene was present at factor 10-5 in comparison of meiotic crossover rate at 10-3. In fact for *H2Aa* the frequency of potential recombinant molecules was astonishingly high, at 17% of analyzed reads in somatic and 19% of reads in sperm DNA being classified as showing a switch between haplotypes.

For *H2Aa* 409 CO candidates were found in 2363 analyzed sequences of somatic DNA while in 2882 sequences of sperm 559 displayed potential recombination events. Figure 10 shows cumulative mutation frequency for these CO candidates, indicating at which SNPs change between Haplotypes could be determined, as well as transmission ratio at these SNPs. In sperm DNA Haplotype A was favored in transmission across the whole gene, resulting in a transmission distortion of 59.41% in favor of Haplotype A at its highest at position 3267 and on average of 55.99%. This type of transmission distortion is not as pronounced in somatic DNA, where the transmission ratio was 51.6% on average. Interestingly in somatic DNA transmission ratio is distorted mainly in the region of the last three exons, whereas in sperm transmission ratio is constantly raised and only lowers towards the ends of the exon. Based on the analysis of the entire interval of *H2Aa* three smaller regions were picked based on visual identification of where switches between Haplotypes were especially numerous. The Fine-Scale resolution of two these potential recombination hotspots are presented in Figure 11 and 12.

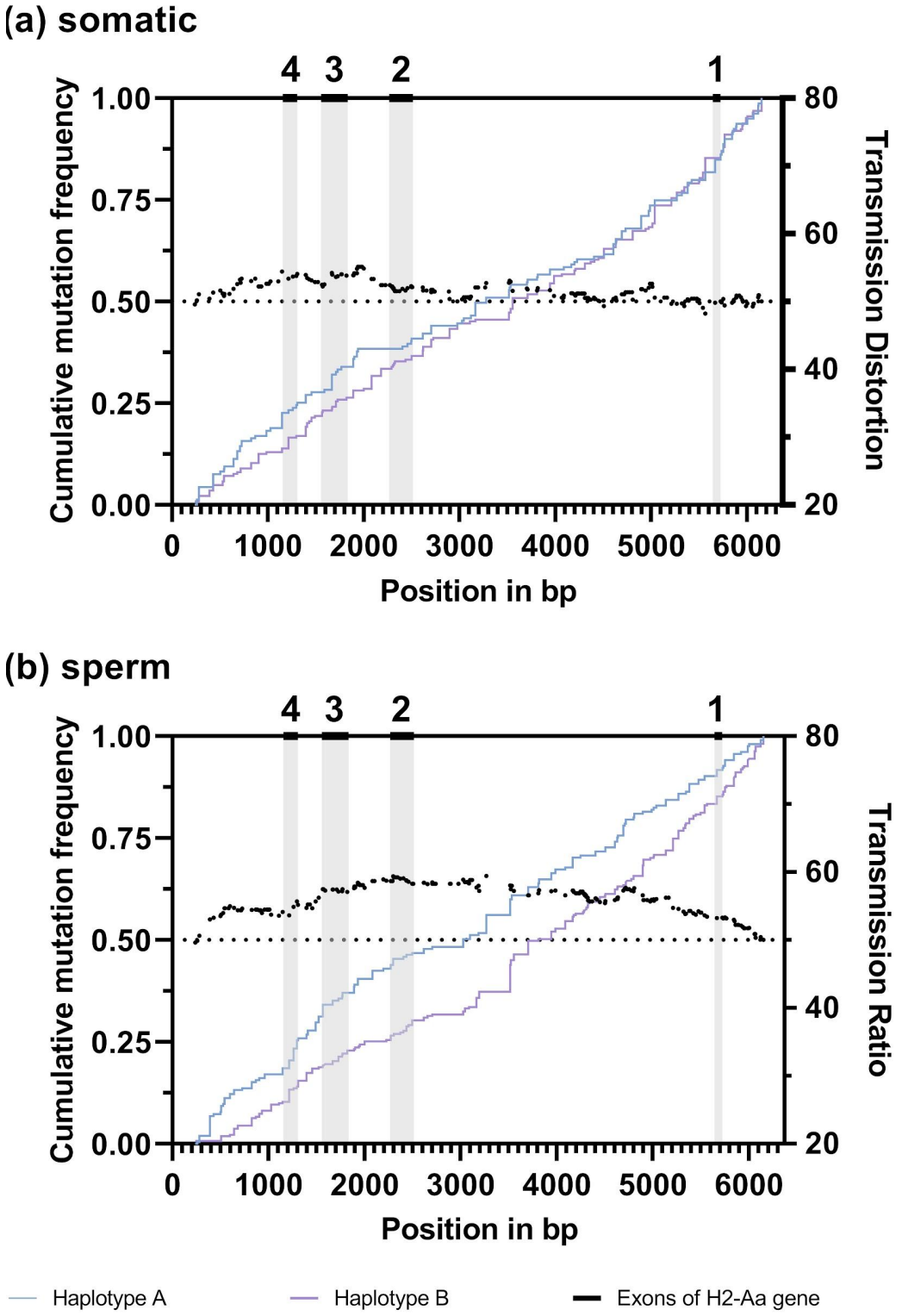

Figure 10: Cumulative mutation frequency and transmission ratio of haplotypes in somatic (a) and sperm (b) DNA across the H2-Aa gene of wildtype MUS mouse WM. Switches from Haplotype A to Haplotype B are cumulated by the blue line, while the purple line shows switches in reciprocal direction, relative to the position of the first switched SNP within the amplicon of the *H2Aa* gene. Transmission ratio at each Single Nucleotide Polymorphism is displayed as a black dot, with values >50 disclosing distortion in favor of Haplotype A and values <50 indicate distortion in favor of Haplotype B. The four exons of the gene are represented by the black bars as well as their corresponding gray shadows on the upper horizontal axis and numbered.

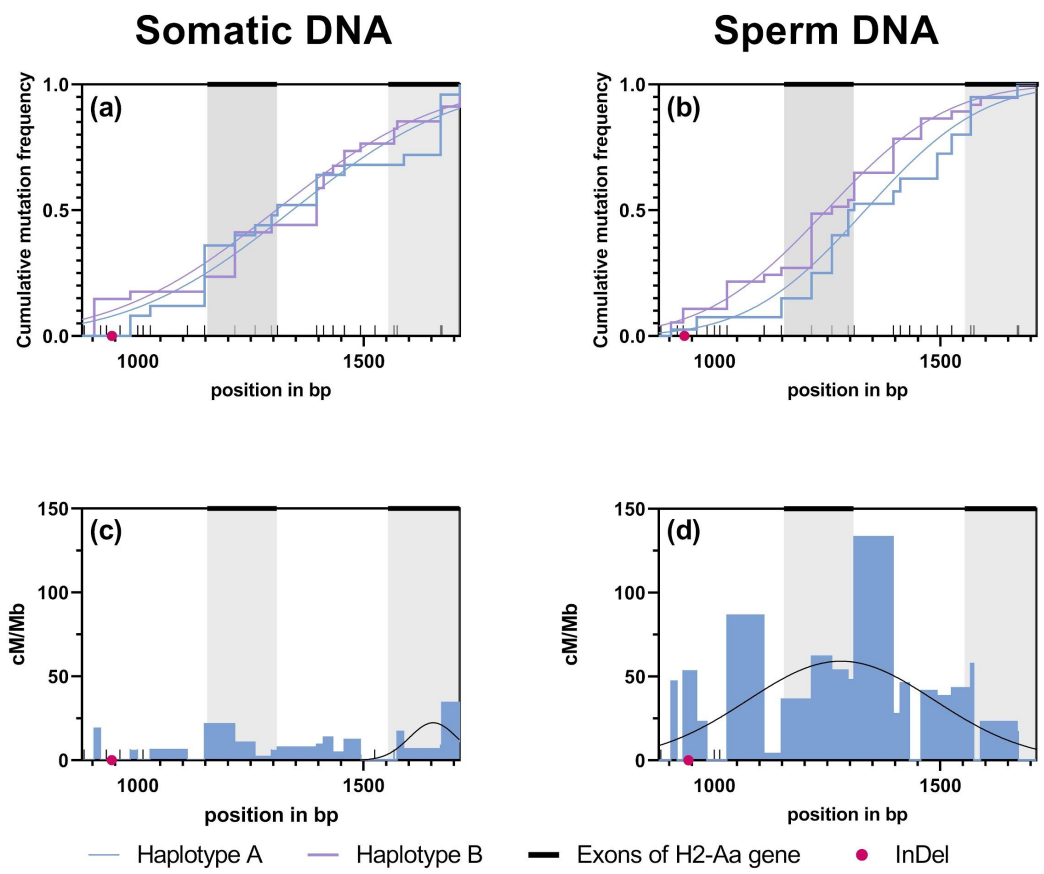

Figure 11: Fine-Scale resolution of first region of raised recombination in the *H2-Aa* gene of MUS wildtype individual WM for somatic (left) and sperm (right) DNA. Color of the lines in (a) and (b) indicate direction of change between haplotypes, with blue representing change from Haplotype A to Haplotype B and purple representing the reciprocal direction. Position of exons within the interval of the hotspot, are indicated as black bars on the upper horizontal axis and mark the interval in gray. InDel positions are shown as pink dots. All SNPs used for analysis are presented as additional ticks on the horizontal axis. All Curves are fitted via cumulative Gaussian regression. (a) and (b) exhibit cumulative mutation frequency. (c) and (d) map breakpoints of switch between Haplotypes cumulative for Haplotype A and B.

For both potential hotspots the comparison of somatic and sperm DNA is shown for both the cumulative mutation frequency in both directions and the distribution of the haplotype switch position for both Haplotypes combined. These regions comprised 59 and 66 somatic as well 77 and 104 CO events in sperm respectively. Notably the first potential hotspot overlaps with the third and fourth Exon of the *H2Aa* gene, with the switch between Haplotypes most common within the exons for somatic DNA but intronic for sperm DNA. Additionally the number of switches in haplotype towards the edge of the proposed hotspot found here in somatic DNA is puzzling. This also distorts the fit of the curve describing the breakpoint distribution to the far right away from the expected position at the inflection point of the cumulative mutation frequency. In sperm for the first region the curves of reciprocal haplotype switches are lightly staggered. Generally even though sperm and somatic DNA did not differ significantly in the number of potential recombinants, the difference in number, with less recombinants being found in somatic DNA is visible in the breakpoint distribution for both regions.

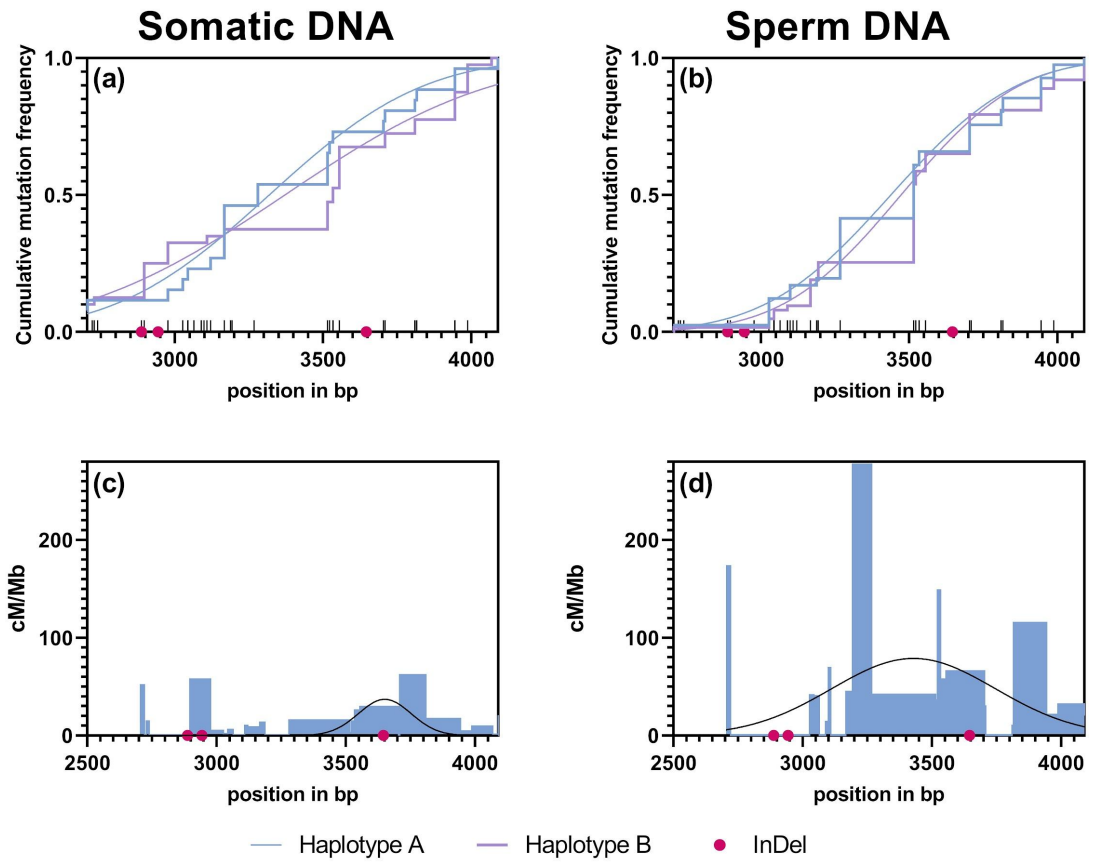

Figure 12: Fine-Scale resolution of the second region of raised recombination in the H2-Aa gene of MUS wild type individual for somatic (left) and sperm (right) DNA. Color of the lines in (a) and (b) indicate direction of change between haplotypes as described in Figure 11. InDel positions are displayed as pink dots. All SNPs used for analysis are shown as additional ticks on the horizontal axis. All Curves are fitted via cumulative Gaussian regression. (a) and (b) present cumulative mutation frequency. (c) and (d) map breakpoints of switch between Haplotypes cumulative for Haplotype A and B.

Curiously for the second potential recombination hotspot in somatic DNA the vertex of the breakpoint distribution curve is located in very close proximity to an 7 bp InDel. The corresponding vertex in sperm DNA is shifted slightly towards the beginning of the amplicon, indicating that the center of the proposed recombination hotspot differs slightly between sperm and somatic DNA. Unlike the first region this interval is located within the second intron, which is also the largest one of the *H2Aa* gene.

4.7. Determination Of Recombination Events in *Prdm9*

Possible recombination events were evaluated for three different mice,among them one wildtype mouse (WM) and two t-haplotype mice (TM4, TD5).

The number of SNP differing between *Prdm9* Zinc Finger Array haplotypes was much lower for the wildtype than for the t-haplotype. WHile the wildtype haplotypes were differentiated by 9 SNPs and one 252 bp long InDel, t-haplotypes TM4 and TD5 showed 33 and 36 SNPs respectively. TD5 also displayed an 168 bp InDel additionally to the deletion in the t-haplotype mmt1 allele within the first Zincfinger of the array. Consequently support for CO events was very low in the wildtype mouse so these results are to be analyzed with caution. Interestingly there seemed to be a bias in somatic DNA in favor of the shorter haplotype as here in WM and TD5 the longer Haplotype was overproportional represented (two-tailed binomial test, $p < 1x10^{-6}$). This bias however was not present in sperm where

surprisingly the longer Haplotype was in fact more common than the shorter Haplotype (see Supplementary Table S2). Only TM4 displayed significant differences in the number of recombinants dependent on direction of change between haplotypes in both sperm and somatic DNA, while also being the only individuals where no major InDels separated the Haplotypes (Chi-Square-Test of Association with Yates-Correction, $p<0.0001$). Generally again there was no significant difference in the number of recombinant molecules between sperm and somatic DNA for all individuals. Again however the frequency of potential recombinants was quite high compared to 4.4% total recombination (including NCOs) at the A3 hotspot (Cole et al., 2010), ranging from 0.75% (WM, somatic DNA) to 6.34% (TM4, sperm DNA). For total number of recombinants and analyzed reads of both *H2Aa* and *Prdm9* see Supplementary Table S2.

The *Prdm9* Zinc Finger array is a minisatellite of 84 bp repeats with three aminoacids at position -1, 3 and 6 in relation to the first amino acid of the alpha-helix displaying most of the differences between alleles as they confine DNA binding specificity (Persikov et al., 2009). Depending on the position of breakpoints during recombination either new ZnFs (CO in between the bases corresponding to these three amino acids) or just shuffling of existing ZnFs would result. The ratio of recombinants within ZnFs to recombinants between ZnFs was between 1.4 and 2.5 for all t-haplotype individuals. While in TM3 the ratio was greater for sperm DNA TD5 showed the opposite relationship. Therefore there seems to be no consistent trend in bias depending on the nature of the DNA. Both t-haplotypes additionally displayeded Transmission Distortion in favor of the non-t-haplotype allele in the second half of the ZnF array for somatic DNA (Figure 13). The sperm DNA of TM4 notably displayed transmission distortion in favor of both haplotypes, with distortion towards the t-haplotype Allele found in the second half of the array and the reciprocal distortion found in the first half. Meanwhile the sperm DNA of TD5 mainly displayed distortion in the first half of the array towards the t-haplotype.

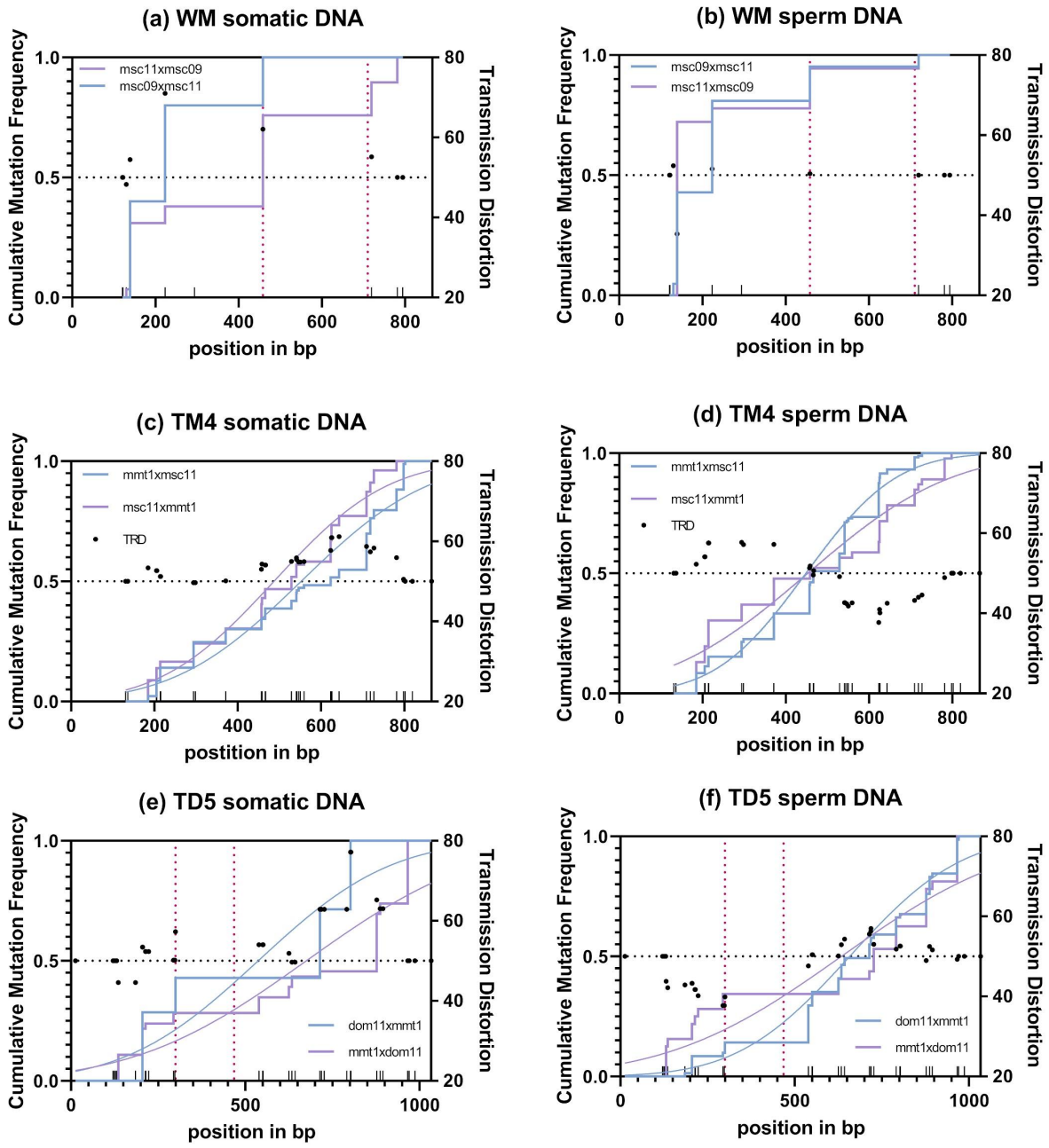

Figure 13: Cumulative Crossover frequency and Transmission Distortion of somatic (right) and sperm (left) crossovers in the minisatellite ZnF array of *Prdm9* for Individuals WM, TM4 and TD6. Different colors of line indicate the direction of crossover between the two alleles as disclosed in the according legend. Transmission ratio is disclosed by black dots and is shown for the longer allele (subsequently the non-t-haplotype allele in (c) to (f)). Dotted line indicates no distortion in transmission for one haplotype (50:50 ratio in transmission). Pink dotted lines mark the extent of InDels.

5. Discussion

5.1. Structural Variation Of The t-Haplotype

Although the presence of inversions within the t-haplotype has been discovered almost forty years ago, their actual extent has mostly been studied using marker loci for Restriction Fragment Length Polymorphism analysis leading to generally correct but imprecise results (Shin et al., 1983, p. 199).While the molecular mechanism creating the TRD associated with male t-haplotype carriers has been uncovered and numerous loci acting as distorters have been identified the knowledge about the structure of the t-haplotype has almost remained stagnant. Most recent analysis of Kelemen and Vicoso, 2018 used heterozygosity of t-haplotype carriers as an approximation for localisation of the four inversion. In this analysis optical mapping was used to identify the inversions of the t-haplotype at an accuracy never achieved before. Notably, although four main inversions currently postulated to be part of the t-haplotype could be found for all subspecies, differences in the position and presence of an additional inversions differentiated the t-haplotypes of the subspecies MUS and DOM. Firstly the MUS t-haplotype lacked the traditional third inversion of the t-haplotype, which is where *Prdm9* is located. This is astonishing to note as the lack of inversion would enable recombination between the wildtype and the t-haplotype. However, t-haplotype carriers have been found to exclusively carry the mmt1 allele of *Prdm9* (Kono et al., 2014). This lack of inversion could be an artifact of the singular individual examined as only one MUS mouse presented sufficient quality in the optical map to analyze. Additionally because the reference used for optical mapping was of a DOM mouse, subspecies dependent differences might have caused the contig showing the inversion to be filtered from the final map. In the DOM subspecies too, few individuals presented evidence for all inversions as a result of fragmentation of contigs in the proximal third of chromosome 17. Therefore additional MUS mice will have to be analyzed to confirm the lack of presence in the region of the third inversion as determined by Kelemen and Vicoso, 2018.

Both subspecies showed evidence for an additional inversion in the region of the second traditionally defined inversion. The most promising evidence for this inversion was the fact that one contig mapped a region in the beginning of the second inversion to the end of the second inversion in the same directional orientation, indicating that two inversions were facing each other here (see Supplementary Figure S2). However this also means that the extent of this additional inversion is unclear, as only its beginning can be identified clearly. While individuals across subspecies showed this motif, only one individual (50101580) of DOM offered a motif that could be used to determine the extent of the two inversions taking up the region of the previously defined second inversion starting just before T and ending before SMOK2a. Within this individual the proximal inversion is the smaller one, spanning about 2 million bp, while the rest of the historical second inversion is taken up by the 4 million bp third inversion. Additionally the first inversion seems to be much larger in size than proposed before, taking the recently discovered distorter Tiam2 into account (Lindholm et al., 2019). The observation of the two inversions in place of the second inversion in Kelemen and Vicoso, 2018 adds to an interesting consideration. While inversions are thought to suppress recombination, this suppression is not actually homogen along the whole inversion. As also found for gene conversions in the t-haplotype the middle of an inversion is more vulnerable for a recombination event than its distal sites (Wallace and Erhart, 2008). The breakpoints of inversions play a notable role, as they add suppression of recombination outside of the inversion but in proximity to the breakpoint (Villoutreix et al., 2021). Thus possible advantageous mutations at the breakpoint itself add into the equation of explaining why an inversion can evolve (Villoutreix et al., 2021). If the

linkage disequilibrium of the inversion is advantageous one larger inversion would suffice to gain advantage. However if breakpoint mutations or the especially high recombination suppression of the breakpoints are the main benefactor, multiple but smaller inversion should arise (Villoutreix et al., 2021). Evidently the t-haplotype consists of multiple inversions, each of which carry distorters needed for expression of the maximal TRD (Hammer et al., 1989). Based on the optical maps described here many distorters are placed in proximity to the proposed inversion breakpoints. Also taking the influence of the inversion breakpoints into account could explain the placement of some distorters as well as the most important responder element (SMOK2a) outside of inverted regions but close to the respective inversion breakpoints.
Also striking to find was the similarity between the optical maps of DOM t-haplotype with the mice of the laboratory Ttp4 strain. As the subspecies of this strain has been undetermined, this finding provides evidence to classify the Ttp4 strain.

5.2. Variation In The Second *H2Aa* Exon Of t-Haplotype Individuals

The t-haplotype MHC alleles can be expected to high similarity between different t-haplotypes since they proposedly emerged from the same ancestral introgression of the t-haplotype into the population and have since been detached from natural selection (Morita et al., 1992). Artzt et al., 1985 found that serologically MHC alleles of the t-haplotype were very similar. However this analysis shows that the MHC alleles of the t-haplotype are not completely homogenous in the way *Prdm9* alleles for example are.

The fact that some loci on the t-haplotype (like *Prdm9* and *Tcp-1*) display no differences between t-haplotypes and non-synonymous mutation rate seems to be lower than expected along the t-haplotype seems puzzling considering the vast number of lethal mutations distinguishing the t-haplotype (Kelemen et al., 2022). The second exon of MHC class II gene Aa expected to be very variable in wildtype mice had different alleles for each t-haplotype that was examined here. As so far no phased sequence of the whole t-haplotype exists, unfortunately t-haplotype alleles of examined mice could not be determined definitively. Furthermore no homogene or even more closely related t-haplotype allele could be identified, indicating further distance between t-haplotype MHC alleles than previously thought. In fact the inferred t-haplotype allele of TM3 did not show much more similarity to one of the other t-haplotypes alleles than it did to the wildtype mouse. The differences in the sequence between different t-haplotypes is especially puzzling considering the proposed recent introgression of the t-haplotype into the *Mus musculus* subspecies. considering an mutation rate of $2.2x10^{-9}$ mutations per site and year (Wallace and Erhart, 2008) and an dating of the introgression to 7000 to 9000 years ago (Morita et al., 1992) only $4.5x10^{-3}$ to $5.8x10^{-3}$ sites should show difference within the amplified 294 bp sequence of the *H2Aa* exon. Even the estimation of introgression of the t-haplotype over a million years ago (Hammer and Silver, 1993) would only lead to about 0.5 mutations within the exon. The vast diversity in sequences found here therefore can not be explained by point mutation alone. Even if the introgression of the t-haplotypes separately into each subspecies at multiple events (which contradicts the uniformity of *Tcp-1* and *Prdm9*) it is curious that the Ttp4 strain, which based on the optical mapping likely is also part of this subspecies, does not show any similarity to the shared allele of the DOM mice. As the wildtype and t-haplotype MUS mice also share higher similarity for one allele, the similar allele of DOM might very well be a population specific rather than t-haplotype allele. Thus it is likely that other forces have acted to diversify the *H2Aa* exon like recombination that could explain why exons are highly diversified while

other t-haplotype genes are not. Wildtype MHC alleles would be exchanged against the t-haplotype alleles and transmitted preferentially on the t-haplotype. One recombination event can result in large differences between the recombined and an unrecombined t-haplotype, explaining the vast diversity found for *H2Aa* exon equivalent to the PBC as described here similar to the mosaic t-haplotypes described by (Wallace and Erhart, 2008).
Adding to this is concentration of variation between different alleles on few amino acid residues, rather than occurring randomly distributed across the whole of the exon. As in the PBC of MHC class II molecules certain residues are conserved to interact with the peptide backbone of the bound antigen, any degenerated t-haplotype alleles with mutations in one of these conserved amino acid residues would be renewed by recombination between t-haplotype and wildtype. In fact recombination has been thought to regenerate some lethality factors accumulated on the t-haplotype like this (Kelemen and Vicoso, 2018).

5.3. The Implementation Of Nanopore Sequencing For Detection Of Recombination

The long-read-sequencing method of Nanopore sequencing enables a never-seen-before length and through-put in sequencing. To be able to use this sequencing method for characterization of recombination events is advantageous, as larger spans of regions can be analyzed. In the experiments conducted here the general fit of nanopore sequencing data for recombination analysis can be investigated. The first Nanopore Sequencing run produced fewer reads than expected, probably due to a mistake in washing with Ethanol of a low percentage during Ampure purification which led to premature release of DNA from the Beads under aqueous conditions. So the coverage of the same molecule was sparse as only a fraction of Primer IDs appeared at least twice. This complicated the analysis of the data, as the high error rate of the MinION Nanopore sequencing could not be compensated. However even after lowering the number of template molecules in PCR and pooling multiple identical reactions in secondary PCR the coverage of each unique Primer ID could not be deepened. In fact the coverage in the second run was slightly lower, although it has to be mentioned that only a third of the dataset could be analyzed. The high error rate of the MinION data probably also impedes identification of Primer IDs. Even based on the higher quality threshold (Q>10) used in the second run, the error rate in sequencing was at most 21.32 % (estimated using calculation of (Delahaye and Nicolas, 2021)). The Probability of a Primer ID with at least one error was therefore be calculated as:

$$P = 1 - (1 - E)^n$$

where E is the error rate derived from the quality score and n is the number of nucleotides in the Primer ID (Kebschull and Zador, 2015). Consequently an error rate of 21.32% P equals 85.31% meaning that in fact only one in six Primer IDs do not contain an error. As the high error rate within the Primer ID prevents the creation of consensus sequences for the samples, classifying NCO events is also ambiguous. The mean track length of NCO events in mice is 200 bp (Odenthal-Hesse et al., 2014), meaning that only few SNPs are going to be included. At the murine A3 Hotspot for example many gene conversions were identified by a single SNP change (Cole et al., 2010). This makes Nanopore sequencing with its large error rate unfitting for the analysis of NCO events if Primer IDs cannot be implemented to generate statistically error free consensus sequences. Additionally, because of the high proportion of deletion sequencing errors (Delahaye and Nicolas, 2021) the mapping of complex repetitive regions was challenging and mostly these could not be used for identification of haplotypes or recombinants. Also for the *Prdm9* amplicons of somatic DNA there seemed to be a bias against the longer sequences, eliminating these almost completely from the

resulting data. While this bias was only present for somatic DNA it is problematic. Potentially this bias might be caused in the PCR of the target genes. Primers might have prevalence for one of the haplotypes if a SNP is included in the Primer sequence, leading to reduced amplification of the other. This could also explain the lack of heterozygous data for H2Eb amplicons of the second nanopore sequencing run, where both in wildtype and t-haplotype mice only one haplotype could be recovered. Additionally the Nanopore sequencing could create an artificial bias for the shorter haplotype, as the shorter Haplotype is sequenced faster and thus more sequences can accumulate. Lastly only a third of the sequences could be analyzed here, so the bias for the shorter Haplotype might be a sampling error.

The last major issue in analyzing the nanopore sequencing data was the lack of appropriate tools for analyzing the data. As the application of nanopore sequencing currently is focused on metagenomic or genomic scale in analysis, working with data generated by PCR, so a much smaller length of sequences at a much higher coverage, is not possible without specialized programs. Even among these none seemed to be the perfect fit for the goals of this study. Additionally these programs require deep bioinformatical knowledge and often larger computing powers therefore being currently inaccessible for the general community. Still the nanopore sequencing offers a length of reads, which here is only restricted by the ability of the PCR, which sets it apart from any other sequencing technique. Especially the MHC with its numerous SNP could be a valuable target to be analyzed for recombination events. Thus, once proper bioinformatic tools for analysis of this kind are established and made accessible, the method presented here holds a promising future for changing the way recombination can be characterized.

5.4. Validity of Detected Recombination Events

Curious about the potential CO events observed is the high frequency and the lack of significance in number between sperm and somatic DNA. At the hotspot A3, considered an intensely active hotspot, frequency of recombination events is 4.4% which includes NCO at a tenfold rate of COs (Cole et al., 2010). Especially at the *H2Aa* Gene the potential CO frequency of 17% to 19% seems suspiciously high. A known error of PCR is the phenomenon of template switching, where especially in later cycles of the amplification, incompletely extended amplicons can act as Primers to create chimeric molecules (Odelberg et al., 1995). Chimeric molecules between parental haplotypes can thus be misinterpreted as recombinant molecules. Given that the other explanation for these unusually high and similar for sperm and somatic DNA recombination frequencies would be somatic recombination this type of PCR error seems more likely to be the cause. Using the Primer IDs chimeric molecules should have been somewhat visible, as IDS were added both to the forward and reverse primer. However since Primer ID coverage was low chimeric molecules could not be identified. Consequently if either somatic or meiotic recombinants are actually present they are disguised by the high number of total recombinants unable to be identified. Although the creation of chimeric molecules is thought to occur randomly across the length of the amplicon (Kebschull and Zador, 2015) there are observable differences in the positioning of where the switch between Haplotypes occurs. If solely caused by template switching this might represent the origin of switching in cycling as earlier switches are amplified in following cycles. Crucial for causing template switching would be the number of cycles in Primary PCR. Thus as in the second nanopore sequencing run cycle number in the primary PCR was reduced a difference in number of potential recombinations should be observed. While this is in fact the case and *Prdm9* shows much less CO candidates than *H2Aa*, this could also be caused by the length of the amplicons. The *Prdm9* ZnF array at around 1 kb length is

much shorter than the *H2Aa* gene at 6 kb therefore naturally less recombination events would occur. Thus based on this data the actual degree of template switching error can not be determined.
Still some of the data displays characteristics of recombination hotspots. For example the breakpoint distribution as shown in Figure 11 (b) (*H2Aa*, sperm DNA) displays similarities to the reciprocal crossover asymmetry as described by Cole et al., 2010. Plus there are differences that can be observed between sperm and somatic DNA although they were treated and analyzed under the exact same conditions. Thus either the error in this analysis is not random or there is some actual biological difference between the types of DNA hidden under the noise of the error. Either way it is unclear if the observed recombinants are true recombinants of artificially induced, evaluation is problematic.

5.5. Recombination Within The *H2Aa* Gene

As the number of recombinants did not differ significantly between sperm and somatic DNA, there is no evidence for *de novo* meiotic recombination. The bias for Haplotype A in somatic DNA might be explained by the nature of determining the Haplotype. As the coverage of Primer IDs was not sufficient to build consensus sequences, errors of the sequencing could not be corrected for analysis. For determination of the Haplotype the first three SNPs had to be unambiguous. For the Haplotype A these three SNPs were three Thymine-Nucleotides, while for Haplotype the motif of the three SNPs was CAC. It is known that transitions between Adenine and Guanine are much more common than any other substitution error (Delahaye and Nicolas, 2021). Therefore more sequences for the Haplotype B were probably discarded due to an error in the Adenin of the second SNP. The fact that this is not the case for sperm DNA might be due to the fact that less sequences were analyzed for this type of DNA (although more sequences of each Haplotype could be recovered), reducing the number of sequences excluded by this error.
Firstly the Transmission Ratio of mutation across the whole *H2Aa* gene is distorted in favor of Haplotype A. In recombination Transmission distortion is the result of biased DSB formation, meaning that as one Haplotype is preferentially cleaved, the other Haplotype is preferentially used as a template in repair of the DSB. Under an evolutionary perspective a distortion in Transmission ratio is notable as it could lead to the fixation of a SNP or a Haplotype. The magnitude of transmission distortion is not only influenced by the recombination frequency at a hotspot and the degree of biased recombination initiation but also by the number of CO and NCO (Cole et al., 2010). Cole and colleagues, 2010 found that at the A3 hotpot in mice NCOs contribute almost twice as much as COs to transmission distortion. As the NCO events are missing in this analysis, the picture of transmission distortion is incomplete. Still at least for potential CO events a Transmission distortion can be defined in sperm DNA. Why this distortion is only present in sperm is questionable. Although as discussed before the legitimacy of CO events here is questionable, this difference could represent the difference between initiation in somatic and meiotic recombination. In meiotic recombination DSB are formed by SPO11 in NDR determined by PRDM9 Histone Methylation. It is known that Haplotypes influence recombination and thus formation of DSB might also be influenced by the Haplotype. In somatic recombination this initiation differs. In Hypermutation of antibodies for example a deaminase induces recombination (Di Noia and Neuberger, 2007). Thus it could be that Transmission Distortion is only present for one type of recombination.
Lastly there is the matter of multiple potential recombination hotspots in close proximity. Usually between hotspots mechanisms interfering with the formation of another hotspot closeby are present (Otto and Payseur, 2019). If the potential COs observed here are actually recombinations, the

lack of CO interference would strengthen the hypothesis of the reservoir model offered by Linnenbrick et al., 2018.

5.6. Recombination Within The *Prdm9* Zinc Finger Array

As for the *H2Aa* gene, there is no significant difference in the number of recombinants between somatic and sperm DNA. Thus no de novo meiotic recombination can be defined for the ZnF Array of *Prdm9*. Within *Prdm9* the potential recombinants seem to lightly favor recombination within a ZnF. This is notable, as this would have an impact on the rapid evolution of the *Prdm9* ZnF array. However, considering the error discussed in 5.4. this could just be true for this error and not actual recombinants.
Notably, there was an extreme bias for the shorter Haplotype in two samples (WM and TD5). Since shorter reads are translocated through the Nanopore in a shorter amount of time, more reads can be sequenced, leading to an innate bias. However this was not the case for the samples of different DNA types so this cannot be the sole cause. Another explanation would be an error in sampling of the Nanopore Sequencing, as only the third of the reads that were basecalled first could be analyzed.

6. Conclusion

Here a map of the t-haplotype inversions is presented at a resolution not given before. While the exact inversion breakpoints of the t-haplotype are still not totally clear, based on the restriction of their possible location given here breakpoint PCR primers can be designed. Ideally these would provide higher resolution of the inversion breakpoints and could eventually be used to determine t-haplotypes by all inversion (to also classify partial t-haplotypes). Especially if partial t-haplotypes in nature exist would provide interesting prospects for the consequent evolution of t-haplotypes and could help in further defining the introgression event(s) and or recombination history of the t-haplotype. Additionally a potential fifth inversion of the t-haplotype was found for *Mus musculus domesticus* individuals. The presence of this inversion may also be confirmed in breakpoint PCR.

Diversity of the second Exon of the *H2Aa* gene described here indicates more complex evolution of genes on the t-haplotype than thought before. The diversity found here could be explained by recombination between the wildtype and t-haplotype additionally to point mutations. Under this premise for existence of genes where no variation is found between different t-haplotypes strong negative selection would have to occur. Especially for *Prdm9*, with regards to its role in recombination and its connection to hybrid sterility, this can be an important piece of information.

While there was no evidence for *de novo* meiotic recombination in t-haplotypes, still the method presented here for analyzing recombination has great potential for future studies, once proper bioinformatical tools are available. For now, while nanopore sequencing achieves an incredible length and resolution of potential recombination hotspots, the analysis is tedious and error prone due to the nature of nanopore sequencing and its high error rate.

7. References

Arnheim, N., Calabrese, P., Tiemann-Boege, I., 2007. Mammalian meiotic recombination hot spots. Annu. Rev. Genet. 41, 369–399. https://doi.org/10.1146/annurev.genet.41.110306.130301

Artzt, K., 1984. Gene mapping within the T/t complex of the mouse. III: t-lethal genes are arranged in three clusters on chromosome 17. Cell 39, 565–572. https://doi.org/10.1016/0092-8674(84)90463-X

Artzt, K., Shin, H.S., Bennett, D., 1982. Gene mapping within the T/t complex of the mouse. II. Anomalous position of the H-2 complex in t haplotypes. Cell 28, 471–476. https://doi.org/10.1016/0092-8674(82)90201-x

Artzt, K., Shin, H.S., Bennett, D., Dimeo-Talento, A., 1985. Analysis of major histocompatibility complex haplotypes of t-chromosomes reveals that the majority of diversity is generated by recombination. J. Exp. Med. 162, 93–104. https://doi.org/10.1084/jem.162.1.93

Baker, C., Kajita, S., Walker, M., Petkov, P., Paigen, K., 2014. PRDM9 binding organizes hotspot nucleosomes and limits Holliday junction migration. Genome Res. 24. https://doi.org/10.1101/gr.170167.113

Baudat, F., Imai, Y., de Massy, B., 2013. Meiotic recombination in mammals: localization and regulation. Nat. Rev. Genet. 14, 794–806. https://doi.org/10.1038/nrg3573

Buard, J., Rivals, E., Segonzac, D.D. de, Garres, C., Caminade, P., Massy, B. de, Boursot, P., 2014. Diversity of Prdm9 Zinc Finger Array in Wild Mice Unravels New Facets of the Evolutionary Turnover of this Coding Minisatellite. PLOS ONE 9, e85021. https://doi.org/10.1371/journal.pone.0085021

Cereb, N., Hughes, A.L., Yang, S.Y., 1997. Locus-specific conservation of the HLA class I introns by intra-locus homogenization. Immunogenetics 47, 30–36. https://doi.org/10.1007/s002510050323

Charlesworth, B., Hartl, D.L., 1978. POPULATION DYNAMICS OF THE SEGREGATION DISTORTER POLYMORPHISM OF DROSOPHILA MELANOGASTER. Genetics 89, 171–192. https://doi.org/10.1093/genetics/89.1.171

Cole, F., Jasin, M., 2011. Isolation of meiotic recombinants from mouse sperm. Methods Mol. Biol. Clifton NJ 745, 251–282. https://doi.org/10.1007/978-1-61779-129-1_15

Cole, F., Keeney, S., Jasin, M., 2010. Comprehensive, Fine-Scale Dissection of Homologous Recombination Outcomes at a Hot Spot in Mouse Meiosis. Mol. Cell 39, 700–710. https://doi.org/10.1016/j.molcel.2010.08.017

Corper, A.L., Stratmann, T., Apostolopoulos, V., Scott, C.A., Garcia, K.C., Kang, A.S., Wilson, I.A., Teyton, L., 2000. A Structural Framework for Deciphering the Link Between I-Ag7 and Autoimmune Diabetes. Science 288, 505–511. https://doi.org/10.1126/science.288.5465.505

Damm, E., Odenthal-Hesse, L., 2022. Orchestrating recombination initiation in mice and men, in: Current Topics in Developmental Biology. https://doi.org/10.1016/bs.ctdb.2022.05.001

Delahaye, C., Nicolas, J., 2021. Sequencing DNA with nanopores: Troubles and biases. PloS One 16, e0257521. https://doi.org/10.1371/journal.pone.0257521

Di Noia, J.M., Neuberger, M.S., 2007. Molecular Mechanisms of Antibody Somatic Hypermutation. Annu. Rev. Biochem. 76, 1–22. https://doi.org/10.1146/annurev.biochem.76.061705.090740

Dod, B., Litel, C., Makoundou, P., Orth, A., Boursot, P., 2003. Identification and characterization of t haplotypes in wild mice populations using molecular markers. Genet. Res. https://doi.org/10.1017/S0016672303006116

Doherty, P.C., Zinkernagel, R.M., 1975. Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. Nature 256, 50–52. https://doi.org/10.1038/256050a0

Duncan, W.R., Wakeland, E.K., Klein, J., 1979. Heterozygosity of H–2 loci in wild mice. Nature 281, 603–605. https://doi.org/10.1038/281603a0

Fujiwara, K., Kawai, Y., Takada, T., Shiroishi, T., Saitou, N., Suzuki, H., Osada, N., 2022. Insights into

Mus musculus Population Structure across Eurasia Revealed by Whole-Genome Analysis. Genome Biol. Evol. 14, evac068. https://doi.org/10.1093/gbe/evac068
Hammer, M.F., Schimenti, J., Silver, L.M., 1989. Evolution of mouse chromosome 17 and the origin of inversions associated with t haplotypes. Proc. Natl. Acad. Sci. U. S. A. 86, 3261–3265. https://doi.org/10.1073/pnas.86.9.3261
Hammer, M.F., Silver, L.M., 1993. Phylogenetic analysis of the alpha-globin pseudogene-4 (Hba-ps4) locus in the house mouse species complex reveals a stepwise evolution of t haplotypes. Mol. Biol. Evol. 10, 971–1001. https://doi.org/10.1093/oxfordjournals.molbev.a040051
Harr, B., Karakoc, E., Neme, R., Teschke, M., Pfeifle, C., Pezer, Ž., Babiker, H., Linnenbrink, M., Montero, I., Scavetta, R., Abai, M.R., Molins, M.P., Schlegel, M., Ulrich, R.G., Altmüller, J., Franitza, M., Büntge, A., Künzel, S., Tautz, D., 2016. Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus spretus. Sci. Data 3, 160075. https://doi.org/10.1038/sdata.2016.75
Herrmann, B., Bućan, M., Mains, P.E., Frischauf, A.M., Silver, L.M., Lehrach, H., 1986. Genetic analysis of the proximal portion of the mouse t complex: evidence for a second inversion within t haplotypes. Cell 44, 469–476. https://doi.org/10.1016/0092-8674(86)90468-x
Herrmann, B.G., Bauer, H., 2012. The mouse t-haplotype:: a selfish chromosome – genetics, molecular mechanism, and evolution, in: Piálek, J., Macholán, M., Munclinger, P., Baird, S.J.E. (Eds.), Evolution of the House Mouse, Cambridge Studies in Morphology and Molecules: New Paradigms in Evolutionary Bio. Cambridge University Press, Cambridge, pp. 297–314. https://doi.org/10.1017/CBO9781139044547.014
Herrmann, B.G., Koschorz, B., Wertz, K., McLaughlin, K.J., Kispert, A., 1999. A protein kinase encoded by the t complex responder gene causes non-mendelian inheritance. Nature 402, 141–146. https://doi.org/10.1038/45970
Hughes, A.L., Nei, M., 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335, 167–170. https://doi.org/10.1038/335167a0
Jabara, C.B., Jones, C.D., Roach, J., Anderson, J.A., Swanstrom, R., 2011. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proc. Natl. Acad. Sci. 108, 20166–20171. https://doi.org/10.1073/pnas.1110064108
Jeffreys, A.J., Wilson, V., Neumann, R., Keyte, J., 1988. Amplification of human minisatellites by the polymerase chain reaction: towards DNA fingerprinting of single cells. Nucleic Acids Res. 16, 10953–10971. https://doi.org/10.1093/nar/16.23.10953
Kebschull, J.M., Zador, A.M., 2015. Sources of PCR-induced distortions in high-throughput sequencing data sets. Nucleic Acids Res. 43, e143. https://doi.org/10.1093/nar/gkv717
Kelemen, R.K., Elkrewi, M., Lindholm, A.K., Vicoso, B., 2022. Novel patterns of expression and recruitment of new genes on the t-haplotype, a mouse selfish chromosome. Proc. Biol. Sci. 289, 20211985. https://doi.org/10.1098/rspb.2021.1985
Kelemen, R.K., Vicoso, B., 2018. Complex History and Differentiation Patterns of the t-Haplotype, a Mouse Meiotic Driver. Genetics 208, 365–375. https://doi.org/10.1534/genetics.117.300513
Kono, H., Tamura, M., Osada, N., Suzuki, H., Abe, K., Moriwaki, K., Ohta, K., Shiroishi, T., 2014. Prdm9 Polymorphism Unveils Mouse Evolutionary Tracks. DNA Res. 21, 315–326. https://doi.org/10.1093/dnares/dst059
Lindholm, A., Sutter, A., Künzel, S., Tautz, D., Rehrauer, H., 2019. Effects of a male meiotic driver on male and female transcriptomes in the house mouse. Proc. Biol. Sci. 286, 20191927. https://doi.org/10.1098/rspb.2019.1927
Linnenbrink, M., Teschke, M., Montero, I., Vallier, M., Tautz, D., 2018. Meta-populational demes constitute a reservoir for large MHC allele diversity in wild house mice (Mus musculus). Front. Zool. 15, 15. https://doi.org/10.1186/s12983-018-0266-9
Liu, J., Gao, G.F., 2011. Major Histocompatibility Complex: Interaction with Peptides, in: ELS. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470015902.a0000922.pub2

Lyon, M.F., 2003. Transmission Ratio Distortion in Mice. Annu. Rev. Genet. 37, 393–408. https://doi.org/10.1146/annurev.genet.37.110801.143030
Lyon, M.F., 1986. Male sterility of the mouse t-complex is due to homozygosity of the distorter genes. Cell 44, 357–363. https://doi.org/10.1016/0092-8674(86)90770-1
Lyon, M.F., Zenthon, J., Evans, E.P., Burtenshaw, M.D., Willison, K.R., 1988. Extent of the mouse t complex and its inversions shown by in situ hybridization. Immunogenetics 27, 375–382. https://doi.org/10.1007/BF00395134
Madden, D.R., Gorga, J.C., Strominger, J.L., Wiley, D.C., 1992. The three-dimensional structure of HLA-B27 at 2.1 A resolution suggests a general mechanism for tight peptide binding to MHC. Cell 70, 1035–1048. https://doi.org/10.1016/0092-8674(92)90252-8
Meyer, D., Mack, S.J., 2008. Major Histocompatibility Complex (MHC) Genes: Polymorphism, in: John Wiley & Sons, Ltd (Ed.), ELS. Wiley. https://doi.org/10.1002/9780470015902.a0006133.pub2
Morita, T., Kubota, H., Murata, K., Nozaki, M., Delarbre, C., Willison, K., Satta, Y., Sakaizumi, M., Takahata, N., Gachelin, G., 1992. Evolution of the mouse t haplotype: recent and worldwide introgression to Mus musculus. Proc. Natl. Acad. Sci. 89, 6851–6855. https://doi.org/10.1073/pnas.89.15.6851
Morita, T., Murata, K., Sakaizumi, M., Kubota, H., Delarbre, C., Gachelin, G., Willison, K., Matsushiro, A., 1993. Mouse t haplotype-specific double insertion of B2 repetitive sequences in the Tcp-1 intron 7. Mamm. Genome 4, 58–59. https://doi.org/10.1007/BF00364666
Nadeau, J.H., Britton-Davidian, J., Bonhomme, F., Thaler, L., 1988. H-2 polymorphisms are more uniformly distributed than allozyme polymorphisms in natural populations of house mice. Genetics 118, 131–140. https://doi.org/10.1093/genetics/118.1.131
Neely, R.K., Dedecker, P., Hotta, J., Urbanavičiūtė, G., Klimašauskas, S., Hofkens, J., 2010. DNA fluorocode: A single molecule, optical map of DNA with nanometre resolution. Chem. Sci. 1, 453–460. https://doi.org/10.1039/C0SC00277A
Odelberg, S.J., Weiss, R.B., Hata, A., White, R., 1995. Template-switching during DNA synthesis by Thermus aquaticus DNA polymerase I. Nucleic Acids Res. 23, 2049–2057. https://doi.org/10.1093/nar/23.11.2049
Odenthal-Hesse, L., Berg, I.L., Veselis, A., Jeffreys, A.J., May, C.A., 2014. Transmission distortion affecting human noncrossover but not crossover recombination: a hidden source of meiotic drive. PLoS Genet. 10, e1004106. https://doi.org/10.1371/journal.pgen.1004106
Otto, S.P., Payseur, B.A., 2019. Crossover Interference: Shedding Light on the Evolution of Recombination. Annu. Rev. Genet. 53, 19–44. https://doi.org/10.1146/annurev-genet-040119-093957
Paterniti, J.R., Brown, W.V., Ginsberg, H.N., Artzt, K., 1983. Combined lipase deficiency (cld): a lethal mutation on chromosome 17 of the mouse. Science 221, 167–169. https://doi.org/10.1126/science.6857276
Penn, D.J., Musolf, K., 2012. The evolution of MHC diversity in house mice, in: Piálek, J., Macholán, M., Munclinger, P., Baird, S.J.E. (Eds.), Evolution of the House Mouse, Cambridge Studies in Morphology and Molecules: New Paradigms in Evolutionary Bio. Cambridge University Press, Cambridge, pp. 221–252. https://doi.org/10.1017/CBO9781139044547.011
Penn, D.J., Potts, W.K., 1999. The Evolution of Mating Preferences and Major Histocompatibility Complex Genes. Am. Nat. 153, 145–164. https://doi.org/10.1086/303166
Persikov, A.V., Osada, R., Singh, M., 2009. Predicting DNA recognition by Cys2His2 zinc finger proteins. Bioinforma. Oxf. Engl. 25, 22–29. https://doi.org/10.1093/bioinformatics/btn580
Persikov, A.V., Singh, M., 2014. De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. Nucleic Acids Res. 42, 97–108. https://doi.org/10.1093/nar/gkt890
Phifer-Rixey, M., Harr, B., Hey, J., 2020. Further resolution of the house mouse (Mus musculus) phylogeny by integration over isolation-with-migration histories. BMC Evol. Biol. 20, 120. https://doi.org/10.1186/s12862-020-01666-9

Potts, W.K., Wakeland, E.K., 1993. Evolution of MHC genetic diversity: a tale of incest, pestilence and sexual preference. Trends Genet. TIG 9, 408–412. https://doi.org/10.1016/0168-9525(93)90103-o
Reche, P.A., Reinherz, E.L., 2003. Sequence Variability Analysis of Human Class I and Class II MHC Molecules: Functional and Structural Correlates of Amino Acid Polymorphisms. J. Mol. Biol. 331, 623–641. https://doi.org/10.1016/S0022-2836(03)00750-2
Roitt, I.M., Delves, 2001. Essential Immunology. Blackwell, London.
Serrentino, M.-E., Borde, V., 2012. The spatial regulation of meiotic recombination hotspots: are all DSB hotspots crossover hotspots? Exp. Cell Res. 318, 1347–1352. https://doi.org/10.1016/j.yexcr.2012.03.025
Shannon, C.E., 1949. The mathematical theory of communication. University of Illinois Press, Urbana.
She, J.X., Boehme, S.A., Wang, T.W., Bonhomme, F., Wakeland, E.K., 1991. Amplification of major histocompatibility complex class II gene diversity by intraexonic recombination. Proc. Natl. Acad. Sci. U. S. A. 88, 453–457.
Shiina, T., Blancher, A., Inoko, H., Kulski, J.K., 2017. Comparative genomics of the human, macaque and mouse major histocompatibility complex. Immunology 150, 127–138. https://doi.org/10.1111/imm.12624
Shin, H.-S., Flaherty, L., Artzt, K., Bennett, D., Ravetch, J., 1983. Inversion in the H–2 complex of t-haplotypes in mice. Nature 306, 380–383. https://doi.org/10.1038/306380a0
Silver, L.M., 1982. Genomic analysis of the H-2 complex region associated with mouse t haplotypes. Cell 29, 961–968.
Silver, L.M., Artzt, K., 1981. Recombination suppression of mouse t-haplotypes due to chromatin mismatching. Nature 290, 68–70. https://doi.org/10.1038/290068a0
Silver, L.M., Hammer, M., Fox, H., Garrels, J., Bucan, M., Herrmann, B., Frischauf, A.M., Lehrach, H., Winking, H., Figueroa, F., 1987. Molecular evidence for the rapid propagation of mouse t haplotypes from a single, recent, ancestral chromosome. Mol. Biol. Evol. 4, 473–482. https://doi.org/10.1093/oxfordjournals.molbev.a040457
Stuart, P.M., 2015. Major Histocompatibility Complex (MHC): Mouse, in: ELS. John Wiley & Sons, Ltd, pp. 1–7. https://doi.org/10.1002/9780470015902.a0000921.pub4
Tock, A.J., Henderson, I.R., 2018. Hotspots for Initiation of Meiotic Recombination. Front. Genet. 9, 521. https://doi.org/10.3389/fgene.2018.00521
Torgasheva, A.A., Rubtsov, N.B., Borodin, P.M., 2013. Recombination and synaptic adjustment in oocytes of mice heterozygous for a large paracentric inversion. Chromosome Res. 21, 37–48. https://doi.org/10.1007/s10577-012-9336-6
Úbeda, F., Russell, T.W., Jansen, V.A.A., 2019. PRDM9 and the evolution of recombination hotspots. Theor. Popul. Biol. 126, 19–32. https://doi.org/10.1016/j.tpb.2018.12.005
Ullrich, K.K., Tautz, D., 2020. Population Genomics of the House Mouse and the Brown Rat, in: Dutheil, J.Y. (Ed.), Statistical Population Genomics, Methods in Molecular Biology. Springer US, New York, NY, pp. 435–452. https://doi.org/10.1007/978-1-0716-0199-0_18
Villoutreix, R., Ayala, D., Joron, M., Gompert, Z., Feder, J.L., Nosil, P., 2021. Inversion breakpoints and the evolution of supergenes. Mol. Ecol. 30, 2738–2755. https://doi.org/10.1111/mec.15907
Wallace, L.T., Erhart, M.A., 2008. Recombination within mouse t haplotypes has replaced significant segments of t-specific DNA. Mamm. Genome Off. J. Int. Mamm. Genome Soc. 19, 263–271. https://doi.org/10.1007/s00335-008-9103-3
White, M.A., Ané, C., Dewey, C.N., Larget, B.R., Payseur, B.A., 2009. Fine-Scale Phylogenetic Discordance across the House Mouse Genome. PLoS Genet. 5, e1000729. https://doi.org/10.1371/journal.pgen.1000729
Womack, J.E., Roderick, T.H., 1974. T-alleles in the mouse are probably not inversion. J Hered U. S. 65.
Yamazaki, K., Beauchamp, G.K., 2007. Genetic basis for MHC-dependent mate choice. Adv. Genet. 59, 129–145. https://doi.org/10.1016/S0065-2660(07)59005-X
Yauk, C.L., Bois, P.R.J., Jeffreys, A.J., 2003. High-resolution sperm typing of meiotic recombination in

the mouse MHC Ebeta gene. EMBO J. 22, 1389–1397. https://doi.org/10.1093/emboj/cdg136
Yeager, M., Hughes, A.L., 1999. Evolution of the mammalian MHC: natural selection, recombination, and convergent evolution. Immunol. Rev. 167, 45–58. https://doi.org/10.1111/j.1600-065x.1999.tb01381.x
Yoshino, M., Sagai, T., Lindahl, K.F., Toyoda, Y., Moriwaki, K., Shiroishi, T., 1995. Allele-dependent recombination frequency: homology requirement in meiotic recombination at the hot spot in the mouse major histocompatibility complex. Genomics 27, 298–305. https://doi.org/10.1006/geno.1995.1046
Yuan, Y., Chung, C.Y.-L., Chan, T.-F., 2020. Advances in optical mapping for genomic research. Comput. Struct. Biotechnol. J. 18, 2051–2062. https://doi.org/10.1016/j.csbj.2020.07.018
Zhou, S., Jones, C., Mieczkowski, P., Swanstrom, R., 2015. Primer ID Validates Template Sampling Depth and Greatly Reduces the Error Rate of Next-Generation Sequencing of HIV-1 Genomic RNA Populations. J. Virol. 89, 8540–8555. https://doi.org/10.1128/JVI.00522-15

8. Acknowledgements

DNA extraction and labeling for optical mapping was performed by Sven Künzel and Cornelia Burghardt. Kristian K. Ullrich generated de novo Maps and Assemblies of data. Mapping data to mm10 reference was done by Linda Odenthal-Hesse and Kristian K. Ullrich.
DNA extractions of somatic and sperm DNA for sanger and Nanopore sequencing were performed by Nicole Thomsen.

9. Supplementary Material

9.1. Map of *Mus musculus* subspecies

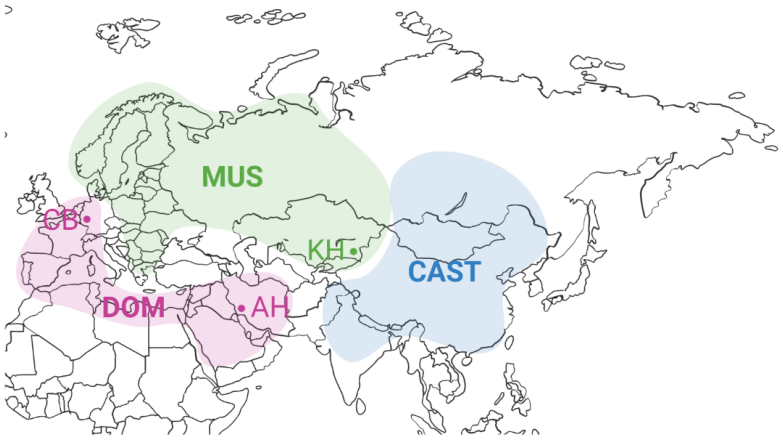

Figure S1: Map of dispersal of *Mus musculus* Subspecies MUS, DOM and CAST in Eurasia and Northern Africa. Respective dispersal of subspecies is shown in green for MUS, pink for DOM and blue for CAST. Populations of sampled mice are marked by dots in the respective color of their subspecies. Marked are Ahvaz (AH) in Iran (DOM), Cologne-Bonn (CB) in Germany (DOM) and Almaty (KH) in Kasachstan (MUS).

9.2. Inferring inversions in optical maps

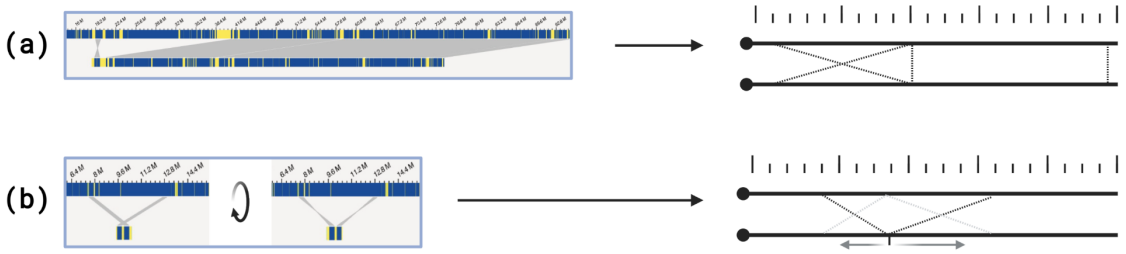

Figure S2: Example of optical mapping data and the respective conclusions. (a) Breakpoint between an inversion and an un-inverted region. By the position of the small inversion the approximate whole range of the inversion can be inferred and breakpoints are confined to a small region. (b) Breakpoint of two inversions facing each other. To show the correct alignment of the contig its direction has to be inverted to show the inversion. THe point where the inversions meet cannot be located, therefore the relative extent of the two respective inversions can not be determined, only the range of both inverted regions is obvious.

As the optical mapping data was fragmented, inversions were not visible in their whole extent. Thus inversions were inferred mainly from two types of scenarios. In the first scenario shown in case (a) of Figure S2 only a small end of the inversion, next to an uninverted region, is visible. Here one end of the inversion was defined as the marker closest to the uninverted region on the contig. The second end is inferred to be the last unmatched marker in the reference before the uninverted region begins. Thus from this motif the approximate extent of the inversion can be inferred. The second motif as seen as in case (b) of Figure S2 represents to inversions facing each other. There the last markers facing each other of each inversion are defined as the “outer” ends of the inversions projected onto the reference. However the inner end of the inversions facing each other can not be defined based on this contig. Thus the extent of these two inversions can not be inferred. This is represented by the gray rather than black markings in Figure S2 and Figure 4.

9.3. Primers

Table S1: Primers used in this study. Gray marked Primers were used for long-read sequencing. ZFA Primers were used for *Prdm9* Genotyping and *H2Aa* Primers amplified the second exon of *H2Aa*.

Name of Primer	Primer Sequence
L_H2Aa_R_Fs	GTAAAACGACGGCCAGTNNNNNNNNGGGGGGGGGACACAGTCTATCATGA
L_H2Aa_R_Rs	CAGGAAACAGCTATGACNNNNNNNNGGGGAAAGAACCTCACCAACCGGG
L_H2Eb_R_Fs	GTAAAACGACGGCCAGTNNNNNNNNGGGGGCCCTGTCTCTTATTATCTTAGC
L_H2Eb_R_Rs	CAGGAAACAGCTATGACNNNNNNNNGGGGAGTGTCTCTGGAGGGATGTG
L_ZFA_R_F	GTAAAACGACGGCCAGTNNNNNNNNGGGGGAAAGTAAGAGAACTGTGGAAGAGGTCAGAA
L_ZFA_R_R	CAGGAAACAGCTATGACNNNNNNNNGGGGGAGATGTGGTTTTATTGCTGTTGGCTTTCTC
M13F	GTAAAACGACGGCCAGT
M13R	CAGGAAACAGCTATGAC
ZFA_F	GAAAGTAAGAGAACTGTGGAAGAGGTCAGAA
ZFA_R	GAGATGTGGTTTTATTGCTGTTGGCTTTCTC
H2Aa_Ex2_F62	ATCCACCTAAATCCATCAGCA
H2Aa_Ex2_R62B	AGGGAGGAGCGGGGTGTGAGC

9.4. Sequence Numbers

Table S2: Number of analyzed sequences. For each individual and gene the number of analyzed sequences for somatic and sperm DNA is given. Haplotypes are differentiated by length into shorter Haplotype (SHT) and longer Haplotype (LHT). Additionally the number of potential recombinants for both directions is given.

		total number of sequences				potential recombinants			
		somatic DNA		sperm DNA		somatic DNA		sperm DNA	
Individual	Gene	SHT	LHT	SHT	LHT	SHT x LHT	LHT x SHT	SHT x LHT	LHT x SHT
WM	*Prdm9*	4382	130	1838	1807	29	5	18	21
TM4	*Prdm9*	1943	5282	1546	2351	93	79	177	46
TD6	*Prdm9*	1709	232	1653	2396	46	7	32	71
WM	*H2Aa*	1261	1093	1482	1420	175	235	243	310

9.5. H2Aa second Exon sequences

>WM Assembly consensus sequence A (reversed)
GGGAGGAGCGGGGTGTGAGCACGCACCATTGGTAGCTGGGGTGGAATTTGACTCCTTAATCAAGATTTCCAAGTTGTATTT
TCCTGTAGCTATGTCTTGCAGTCCACCTTGGGGGTCAAAGCTTGCCAATTGGCCAAACTCAGGAAGCATCCAGACAGTCTCC
TTCTTATCCAAGTCCACATAGAACCACTCATCACCATCAAATTCAAATGTGTACTGGCCAATGTCTCCAGGAGACTGATATACA
CTTATACCATAGAAGCCTACGTGGTCTGCTGATGGATTTAGGTGGAT
>WM Assembly consensus sequence B (reversed)
GGGAGGAGCGGGGTGTGAGCACGCACCATTGGTAGCTGGGGTGAAATTTGACCTCTTAGTCCACTCTCCCAAGTTGTGTTT
TCCTGTAGCTATGTTTTGCAGTCCACCTTGGGGGTCAAAGCTTGCCAATTGGCCAAACTCAGGAAGCATCCAGATAGTCTCCT
TCTTATCCAAGTCCACATAGAACCGCTCATCACCATCAAATTCATGTGTGTACTGGCCAATGTCTCCAGGAGACTGATATACAG
TTGTACCATAGACGCCTACGTGGTCTGCTGATGGATTTAGGTGGAT
>TD1 Assembly consensus sequence A (reversed)
GGGAGGAGCGGGGTGTGAGCACGTACCATTGGCAGCTGGGGTGAAATTTGACCTCTTAGTCAAGACTTCCAAGTTGTGTTT
TACTACAGCTATGTTTTGCAGTCCACCTTGGGGGTCAAAGCTTGTCAATTGGCCAAACTCAGGAAGCATCCAGACAGTCTCC
TTCTTATCCAAGTCCACATAGAACCACTCATCACCATCAAATTCATGTGTGAACTGGCCAATGTCTCCAGGAGACTGATAAACA
GTTGTACCATAGACGCCTACGTGGTCTGCTGATGGATTTAGGTGGAT
>TD1 Assembly consensus sequence B (reversed)
GGGAGGAGCGGGGTGTGAGCACGTACCATTGGCAGCTGGGGTGAAATTTGACCTCTTAGTCAAGATTTCCAAGTTGTGTTT
TACTGTAGCTATGTCTTGCAGTCCACCTTGGGGGTCAAAGCTTGCCAATTGGCCAAACTCAGGAAGCCTCCAGATAGTCTCCT
TCTTATCCAAGTCCACATAGAACAACTCATCACCATCAAATTCAAATGTGAACTGGCCAATGTCTCCAGGAGACTGATATACAA
CTATACCATAGGAGCCTACATGGTCTGCTGATGGATTTAGGTGGAT
>TD2 Assembly consensus sequence A (reversed)
GGGAGGAGCGGGGTGTGAGCACGTACCATTGGCAGCTGGGGTGAAATTTGACCTCTTAGTCAAGATTTCCAAGTTGTGTTT
TACTACAGCTATGTTTTGCAGTCCACCTTGGGGGTCAAAGCTTGTCAATTGGCCAAACTCAGGAAGCATCCAGACAGTCTCC
TTCTTATCCAAGTCCACATAGAACCACTCATCACCATCAAATTCATGTGTGAACTGGCCAATGTCTCCAGGAGACTGATAAACA
GTTGTACCATAGACGCCTACGTGGTCTGCTGATGGATTTAGGTGGAT
>TD2 Assembly consensus sequence B (reversed)
GGGAGGAGCGGGGTGTGAGCACGTACCATTGGCAGCTGGGGTGAAATTTGACCTCTTAGTCCAGCTTTCCAAGTTGTGTTT
TCCTGTTACTATGTTTTGCAGTCCACCTTGGGGGTCAAAGCTTGTCAATTGGCCAAACTCAGGAAGCATCCAGACAGTCTCCT
TCTTATCCAAGTCCACATAGAACCACTCATCACCATCAAATTCATGTGTGAACTGGCCAATGTCTCCAGGAGACTGATATACAA
CTGTACCTTAGACGCCTACGTGGCCTGCTGATGCATTTATGTGTAT
>TM3 Assembly consensus sequence A (reversed)
GGGAGGAGCGGGGTGTGAGCACGCACCATTGGTAGCTGGGGTGAAATTTGACCTCTTAGTCAATTCTCCCAAGTTGTGTTT
TCCTGTAGCTATGTTTTGCAGTCCACCTTGGGGGTCAAAGCTTGCCAATTGGCCAAACTCAGGAAGCATCCAGATAGTCTCCT
TCTTATCCAAGTCCACATAGAACCACTCATCACCATCAAATTCATGTGTGTACTGGCCAATGTCTCCAGGAGACTGATATACAG
TTGTACCATAGACGCCTACGTGGTCTGCTGATGGATTTAGGTGGAT
>TM3 Assembly consensus sequence B (reversed)
GGGAGGAGCGGGGTGTGAGCACGCACCATTGGTAGCTGGGGTGGAATTTGACCTCTTAGTCAATTCTCCCAAGTTGTGTTT
TCCTGTAGCTATGTTTTGCAGTCCACCTTGGGGGTCAAAGCTTGTCAATTGGCCAAACTCAGGAAGCATCCAGACAGTCTCC
TTCTTATCCAAGTCCACATAGAACAACTCATCACCATCAAATTCAAATGTGTACTGGCCAATGTCTCCAGGAGACTGATATACA
ACTATACCATAGGAGCCTACATGGTCTGCTGATGGATTTAGGTGGAT
>Ttp4 Assembly consensus sequence A (reversed)
GGGAGGAGCGGGGTGTGAGCACGTACCATTGGTAGCTGGGGTGGAATTTGACCTCTTAGTCAAGACTCCCAAGTTGTGTTT
TACTACAGCTATGTTTTGCAGTCCACCTTGGGGGTCAAAGCTTGCCAATTGGCCAAACTCAGGAAGCATCCAGACAGTCTCC
TTCTTATCCAAGTCCACATAGAACAACTCATCACCATCAAATTCAAATGTGTACTGGCCAATGTCTCCAGGAGACTGATATACA
CTTATACCATAGGTGCCTACGTGGTCTGCTGATGGATTTAGGTGGAT
>Ttp4 Assembly consensus sequence B (reversed)
GGGAGGAGCGGGGTGTGAGCACGCACCATTGGTAGCTGGGGTGAAATTTGACCTCTTAGTCAAGATTCCCAAGTTGTGTTT
TCCTGTAGCTATGTTTTGTAGTCCACCTTGGGGGTCAAAGCTTGTCAATTGGCCAAACTCAGGAAGCATCCAGACAGTCTCCT

TCTTATCCAAGTCCACATAGAACMACTCATCACCATCAAATTCATGTGTGTACTGGCCAATGTCTCCAGGAGACTGATATACAC
TTATACCATAGGAGCCTACGTGGTCTGCTGATGGATTTAGGTGGAT

9.6. Parental Haplotypes of Nanopore-Sequencing

9.6.1. *H2Aa* (WM)

>Parental Haplotype A
GGGGGGACACAGTCTATCATGATGGGGTAGACAANGGGAACATGAGGTGTTTGATCACACTGCATCCACAGTCAGGATGCA
GAGAGAGGTAAAGTCTGGCTCACTTTCTTTTTGGCCAACACACAGGACACTGCTATCTATATTCAGGTTGGGTCTTGCTACTT
CACATAAATATTTCTGGAAATGCTACAGAGGTGCACAAGAGGTATGTCACCTTGATGAGAATAAAATGTTTTGAGAACAATTT
TGACAATTGATAACATTTTAAAAACCTTTATCCCTTTTGAGATTTAAAGTTCATTTCATAGTAGTTCCTTCNTTTTTTTTCTTCTC
TCTAATTGTTTAGACAGAGGCTTCCTATAAAGCTCCAGCTTGCATTGTGTAGACCAGGTTAGTCTTAAATTTGTGATGACTTTC
CTGCCTTGTCCCTGGAGTGCTGGGATTACTGGTGTGTACCTTGTGCCTTGGTATTTCCTTATAGCAAGTGTACACTGAGGCAG
ATTTCACACCCAACCTCCCTGTACCTGACTGGAGACAGATGAGAGTGGTCANGGCTGGGCTGGGAACAAARAGAAGAAAG
TTCAACTTTGACTCAGATACACAGCATTCTGGGCTGCATTTGGTTGCTCCTGTGACACTGGACATGGGATTACATGAGTCTGG
CTCCAGTAAGCTCCAAAATGGAAGACARGRAAAACCAAACCCAACCCAGGAGACAACATATAAGAAGAGATTTTATTGGTTT
TGTGGATCAGGGTCTCTGAGAGAGTCTGCAGTAGATGTGACTTGGAAGATGCCAGAAAACAAAATTCAGATCAGGGAAGA
ATTCCAAGGGTGTGTGAGCTGTGAGAGGGATAGCCTCMGGRSCAARGGAAAAGCAAGTTGGGGGTTACTTGAAGAAGAG
AAAGAACACCATGAATTGGTCCAGGCCANNCACCCAGCACACCACTTCTTCCCTGAAGAGGGACACACGCCTGCAACAAA
ACAGACAGGAGACAGATAGGTTCCAGAGACTACAGCAAAGTTCCAACCACGTTTTCATGTCTCCCCTTCCTTTATGTCCCCCT
GTGTGGTAGCTTTGGCGCCCACATATACAAACCCTTACCTTCCTTTCCAGGGTGAGACTCATAAAGGCCCTGGGTGTCTGGA
GGTGCCACCTGATCGCAGGCCTTGAATGATGAAGATGGTTCCCACCACGATGCCCACAAGGCCCACAGACAACCCCAGGGC
ACAGACCACAGTCTCTGTCAGCTCTGACATNGGGGCTGGGATCTCAGGTTCTGAGAAAGAGAAGTTTTGGAGTCAGAAGG
GAGTATGTTCATGTATGTGTAAGATAGGGCCAGGGTGAGGGGGCAAGTTAATCTACATACACTTTGTGGGTTGGGCTTATGG
AATGACATGAAATTTAGCCTTAGGAGCTCTTTAAGAGTAGAAAGTGAGTTTGGTCACTAGAATCCAGAAGACATGCTGTTCT
AGAGATGACAATGAGTTAGAAAGTGCAGAAGATGGCAGAGCTCGTACGCACCCCAGTGTTTCAGAATTGGCTCGTCCAGGC
CCCAGTGTTCCACCTTGCAGTCATAAATGTCATCGTCAGAAGGGATGAAGGTGAGATAAGACAGCTTGTGGAAGGAATGGT
CACGGTTAACAARGAAGCTGGTCTCATAAACGCCGTCTGTGACTGACTTGCTATTTCTGAGCCATGTGATGTTGATCACAGGA
GGGAAGATGTTGTCCACAAAGCAGATAAGGGTGTTGGGCTGACCCAGCAGCACAGGGGACTTRGGGAACACAGTTGCTT
GAGGAGCCTCTGTAAAGGAAATGGTCTGGACATGAGTGGCTGATGGAGTTTCTCTCCCTGAAGTTCTGCATTGCCTGGAGA
GGGAGCCCTCTTGCCAAGACTTGCACACATGATGGTCACAGAGCCCAATTGTCCCCTCTGCTTGGACCTAATGTGCACCACA
TCCACTTTCTCCCTGTGCCCTTCTCTCTAGGGTCTTAGGTAGGATGTCTGTGGGGCCTCTTCAAAGATTCCAGGCATGCACTG
TGGGAACAGCAAGTTGTATATTACATGAAATGTATGATTTGTAAGGAGAGGGAAGGTGAGTATTCAGTGATGAGGAGATCTG
GGTGCTCCTTGGAAAGGAGAGGATTTTACAGAACCATTTATACACAAGTCATATCCACCCACCACCACCCACCCCAAGGGAA
GGAAGGGAGGAGCRGGGTGTGAGCACGCACCATTGGTAGCTGGGGTGRAATTTGACCTCTTAGTCCACTCTCCCAAGTTGT
GTTTTCCTGTAGCTATGTTTTGCAGTCCACCTTNGGGGTCAAAGCTTGCCAATTGGCCAAACTCAGGAAGCATCCAGATAGT
CTCCTTCTTATCCAAGTCCACATAGAACCGCTCATCACCATCAAATTCATRTGTGTACTGGCCAATGTCTCCAGGAGACTGATAT
ACAGTTGTACCATAGACGCCTACGTGGTCGGCTGATGGATTTAGGTGGATAGTGGAAAAATGAAAAACAGGAAGAAAAAAT
AATGTTATTAAACAAACTGATAGAAAATGCTTTTACAAATACCACGAATTTCATTTTTTGGGGTTTGTTTCTACCACTCATTCCT
CCCTCCCACTTGGTATGGATTACAATCCAGTCTTTCCCTCTGTGACACGTGCTACATAACTGCGGAATGCCATGTTGATTTATGC
CATCCTGACTGTCTTGTTCCATGCCTGGTACACAGACAGCTCAANCYAAGTAGCTTACAGCTAAACAGAGAGAGAGAGAGA
GAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGCTTTCTTTTCAAAGATTCAGTCAGC
ATATTTCCTTCTCCTATTTATTAGATTTTGTATCTCATTGATCTTCTACTAATTAATTACATATTACATTTTANNNATTTTCCATACAA
GGAGACTTTCTCATTTATTTTAACAAATGGAAAAGAATATAGACCCTTAGATAAGAGAACAGTGACATCATGTCAGTGTGAGG
AGACTGTTCTGTCTAAAAGTTCCTGTATTAGCTGGTGACCCACAGTGTGGCCTTCTGGTTTCTTCTGGAATAAAAGAAATAGA
GGCTTTATTGGACTTCATGCATCTGTCAACCTCACCTCTTACTTTTTTTCTAAATCAACACTGGTGTTCACTTGCAGATTCATGG
TGCATTAAAGCACTTTCCTCCGACAGTGAGCATGTGCGAGCTGCCACGGGGAATGGCTGCCAGCGCAGACAACTGAACCCA
AAGGAAGCCAAGTAAACAACTATCGTCAGGTAAGCATCAGCAAGGCACAGATTCTAGTATTTAAGTTTTGTATCTGTTTTCCT
TCATTTGTTGAGTTCCAGCGAATGTCAGAGGTAGTGCTGTTCAAGGAGCCAGGCCTGCCTCTGCAGTCACCACCCTAGTGAA
TGTGAGGAGAGCAGGGTCAGTCAACAGACAGAGTGCATCTAAGGAGCTGTGAGTGGCGGTGGGCAAGTTACAGCCTGGC

GTTCTGCACAGTGACTGTCAAACAAAACTAACTACCCTATTCTCTCATGGCCCTGTGGATGGCAGGAGTCTCCTTTTCTCACA
GAAATGAAGGCTGTTGTTGTTGGGGCTGTTGNTTTTCAGTTTGCTGAAACCTCCTCTACTCTGGAGGTTTCTGCTCAGTTTTA
ATTAATAGGCTTCCAACTGCACTATGANCCCCAGCCTCTCTGTCTATCTATCCCTGTCTGTCTGTCTATTCACATTTGACTACTCC
TTTGGCCTGTCCTTTCCCAGTGTGGCCACCTTGGTTGGTATTATCAGCCAATAACATGCCACACCTCTCCATCAAAGCTCCCTT
ATTCATTATCAATAAACCTTCTCGACTTCTTTTCCCACTGTGGATGTACACATCGTTCACTTCTGCAGGTGCTTGTCCTGGGCCT
CAGGCGTTGCTGAGGACTGCACGCTGCCCCCTGCTGGTACAGTTTTGCCCTCAGCTGGGTCCACTCTCCACACAATGGCCTT
TCTNTCCTGACCTTGATCTGTTTTCCTCACTNCTCACTCCCTGGGGGCAGGCAGCAGGTTTTATATACTTCAGCACTGACTTGT
TAGCTGAGCAAATGCGTGACAGAGCGAGTATTGAGTGGTGATTGCTAATTTTGAGGCCCGTGGAGACTTTGGGGATGATGG
GAAGGCTGCAGAACCGTGTGAAAGTCACTGTCAGGGATCACCTGGGCTGCATTGGGACCGGTATGTCCCATGCACTTAGCT
TTTGGGCTCTTCCTAGATGAAAATATTGAAGGTCATATTCTATATATAAGAAAATGTAATAACACAGGCTAAATTAAAGTCAAA
GCATTTTCTTAGATCTTGAAACTTTCACTTTTTTCTAAATTAAAAAGAAGTAGGGATATTTTCATGTAGCCCTTCAGTTATATGT
TCCTCAGAACTGTGCCCACTGTGCCCGCTCATACCAGATTAAGGTGCCCTAGTGTGACACTGAAGAAATTGGGGACAAAGAA
GCCACCACCCCCACCCACCCTCCATGAGGGACAGGTGAGACTTCCATTTCCTGCCAAAGCTGGAAAGAGAGCAACCACGCT
TTGCAAGGACGCTCAACCCCTTTCACTTCCTGTTTCAGACATTTGACTTTTCAATTGCTATTTTACACAGTTCCTATTCGCGTCT
GTAAAACAGCATGGAGTTAAGGAGAGATCTGACAGAGGACATGGAATGTCAGCAGAGTGTGGAGGGAGGTGGGCAGCTC
AGGCAGAAGGCCACAGGCTCCTGTAGCTTGTGCTTGCCTGGGTTGTTACAGAACTTCATGGGGTATAAATGGGCTATAGAAC
AGGCAGCTCACATTCCTTACTTTGTTACAAAAATGTTCAATGGTATCTATAATGGAAAGAAAGAGGAAGATGGCATTTAAAGC
ATTTATTTATGAATGGATACATTCTTGAGACATGATCTCACTAGGGGATCAGGCTATCCTTGAACTCACAATCCTCCTGCTTGG
GCACCTTAGCACCGTAGTTACTGCCAAGCACTGCCAGGCCCACCCAGCGTTGCCTTCACTATTTTTAAAGCACATCTACATATT
TATTTGACTTGTCACTTGTGAGTTGACACGTGAGGAAATCCAAAACCACTAAGATCGTCRGGTAAAATAGTGACCTCCTTCTC
TCACTTCCCACCAGGGCCAGCTCCTAGCTTGAAAAAGGTGTGTGGAAACCACCACACAAGGGAAGCATTTGCTGTATATTTG
TGAACATGATTCACGGAAAGGTTTTCCTCAGAGTACACTACAGGTACAAAGGCTTCAGAAACTCACAGAAATTTTAGAGTCA
TAGTAATTATGGAAATTAGATTTTCAGTAATGAAAATCATTTATTCGTGGGAAACACTCCCAATTCCCAATTTCTTCCTTCCCTC
ACTGGGCCCACTAGTAGAATCATGGTTTGGTGGATTGTGAGCTGACCAACCAGAGCCTCCTGAAGAAGGGATCACACAGTT
GTTAGACCTCTTCATATTGAGAATTTGTATCTCCCTCTTCTGTGGGTTTCCCATTGCCCGGCTCCAGGCATCCCTCAGCTGCAC
AACTCACCCTCAATGTCGTCTTCACCTCCACAGAGGCTGAGCATGGTGGTCAGGGCGAGGACCCCCAGAATCAGAGCTCTG
CTGCGCGGCATCCTAGTCTCTGGGAGGTCTCTGGAGCTGCTCTCCTCAGTCTTCAAGAGATTGTGAGGATCCAGCCAAGGA
GATCACACACCCAAACCAAACTCGCCAGAACCAATCAGAAAATTCTCTTGTGATGACGTCACAGTTGCCAGCTGCAGACTTT
GTATTAGGAGGCCTGGAAGAGGCTANGGGAACTCCCTGAACCTGCTGCAGGCTCTACTCAGAGTGGACCCGAGAGAGTGC
CTGGGAAGAAGTGGGTTCTAAGTCAGACTCTTTCTCTTGTGCTAATGTGCACAGATGGCAAGAGAGGANGGGATTTCTGAA
CATGGACAGGCGAACTCACCCACAATAATATTGTTCTTCATAACTGTCTGGGCTGCACTGTTCAGAAAGCCTCCCCCCCCNGA
ATTTATTTCTTTCCTGGTGCCCCTCCAACTTCGACACCCAGTTACTTTCTCTTGCTTGCATGCATCATGAGTTAGCTGGTTTTTC
TTTATTTCTTCCCATAGTGTGTGCAGAGTACTCCAAATTGTGAGATGACCCACCTACCCAAGGTGTCCCCACAGACTGTGGTC
AGTCAATGGCTTTGCTCTTTCTGCTTGTGTAAACCCGGTTGGTGAGGTTCTTTCCCC
>Parental Haplotype B
GGGGGGGGACACAGTCTATCATGATGGGGTAGACAANGGGAACATGAGGTGTTTGATCACACTGCATCCACAGTCAGGAT
GCAGAGAGAGGTAAAGTCTGGCTCACTTTCTTTTTGGCCAACACACGGGACACTGCTACCTATATTCAGGTTGGGTCTTGCT
ACATCACATAAATATTTCTGGAAATGCCACAGAGGTGCACAAGAGGTATATCACCTTGATGAGAATAAAAAGTTTTGAGAACA
ATTTTGACATTTGATAACATTTTAAAAACCTTTATCCCTTTTGAGATTTAAAGTTCATTTCATAGTAGTTCCTTCNTTTTTTTTCT
TCTCTCTAATTGTTTAGACAGAGGCTTCCTATAAGGCTCCAGCTTGCATTGTGTAGACCAGGTTAGTCTTAAACTTGTGATGAC
TTTCCTGCCTTGTCCCTGGAGTGCTGGGATTACTGGTGTGTACCTTGTGCCTTGGTACAGCATGTAAACTGAGGCAGATTTCA
CACCCCACCTCCCTGTATCTGATTGGAGACAGATGAGAGTGGTCAGGGCTGGGAACAAAANGAAGAAAGTTCAACCTTGA
CTCAGATACACAAGCATTCTGGGCTGCATTTGGTTGTTCCTGTGACACTGGACATGGGATTACATGAGTCTGGCTTCAGTAAG
CTCCAAAATGGAAGACAGGAAAAARAAACCAAACCCAGTAGACAACATATAAGAAGAGATTTTATTGGTTTTTGTGGATCAG
GGTCTCTGAGAGAGTCTGCAGTAGATGTGACTTGGAAGATGCCAGAAAACAAAATTCAAATCAGGGAAGAATTCCAAGGG
TGTGTGAGCTGTGAGAGGGACAGCCTCAGGGCCAAGGGAAAAGCAAGTTGGGGGCCACTTGAAGAAGAAAAAGAACAC
CATAAATTGGTCCAGAAAACACACTGTGCCAGGTCACCCAGCACACCACTTCTTCCATGAAGAGGGACACACACCTGCAAC
AAAAAGACAGGAGACAGATGGGTTCCAGAGACTACAGCAAAGTTCCAACCACGTTTTCATGTCTCCCCTTCCTTTATGTCCC
CCTGTGTGGTAGCTTTGGYKCCCACATATACAAACCCTTACCTTCCTTTCCAGGGTGTGACTCATAAAGGCCCTGGGTGTCTG

GAGGTGCCACCTGATCGCAGGCCTTGAATGATGAAGATGGTGCCCACCACGATGCCCACAAGGCCCACAGACAACCCCAGG
GCACACACCACAGTCTCTGTCAGCTCTGACATNGGGGCTGGAATCTCAGGTTCTAAGAAAGAGAAGTTTTGGAGTCAGAAG
GGAGTATGTTCATGTATGTGTAAGATAGGGCCAGGGTGAGGGGGCAAGTTAATCTACATAGACTTTGYGGGTTGGACTTATG
GAATGACATGAAATCTAGCCTTAGGAGCTCTTTAAGAGTGGAAAGTGAGTTTGGTCACTAGAATCCAGAAGACATATTGTTC
TAGAGATGACAATGAGTTAGAAAGTACAGAAGATGGCAGAGCTCGTACGCACCCCAGTGTTTCAGAACCGGCTCCTCCAGG
CCCCAGTGCTCCACCTTGCAGTCATAAATGTCATCGTCAGAAGGGATGAAGGTGAGATAAGACAGCTTGTGGAAGGAATGG
TCACGGTTGACGAGGAAGCTGGTCTCATAAACGCCGTCTGTGACTGACTTACTATTTCTGAGCCATGTAATGTTGATCACAGG
AGGGAAGATGTTGTCCACAAAGCATATGAGGGTGTTGGGCTGACCCAGCAGCACAGGGGACTTGGGGAACACAGTCGCTT
GAGGAGCCTCTGTAAAGGAAATGGTCTGGACATGAGTGGCTGATGGAGTTTCTCTCCCTGAAGTTCTGTACTGCCTGGTGA
GGGAGCCCTCTTGCCAAGACTTGTATACATGATGGCCACAGAGCCTGATTGTCCCCTCTGCTTGGAGCTAATGTGCACTGCAT
CCACTTTCTCCCTGCGCCCTTCTCTCTAGGGTCTTAGGTAGGATGTCTGTGGGGCCTCTTCAAAGATTCCAGGCATGCACTGT
GGGAACAGCGAGCTGCATATTACATGAAATGTGTGATTTGTAANGGAGAGGGANNNNNGGAGGTGAGTATTCAGTGATGA
GGAGATCTGGGTGCTCCTTGGAAAGGAGAGGATTTTAAGAACCTTTTATACACAAGTCATATCCACCCACCACCAGCCTCCN
ANCCAAGGGAAGGAAGGGAGGAGCGGGGTGTGAGCACGTACCATTGGTAGCTGGGGTGGAATTTGAGKYCTTAATCAAG
ATTTCCAAGTTGTATTTTCCTGTAGCTATGTCTTGCAGTCCACCTTNGGGGTCAAAGCTTCTCAATTGGCCAAACTCAGGAAG
CATCCAGACAGTCTCCTTCTTATCCAAGTCCACATAGAACAACTCATCACCATCAAATTCAAATGTGTACTGGCCAATGTCTCC
AGGAGACTGATATACACTTATACCATAGAAGCCTACGTGGTCGGCTGATGGATTTAGGTGGATAGTGGAAAAATGAAAAACA
GGAAGAAAAAATAATGTTATTAAACAAACTGATAGAAAATGCTTTTACAAATACTACGAACTTCATTTTTTGGGGTTTGTTTCT
ACCACTCATTCYTCCCTCCCACTTGGTATGGATTACAATCCAGTCTTTCCCTCTGTGATATGTGCTACATAACTGTGGAATGTCAT
GTTGATTTCTGCCATCCTGACTGTCTTGTTCCATGCCTGGTACACAGACAGCTCAAGCYAAGTAGCTTACAGCTAAACAGAGA
NNNNNNRSRRAGAGAGAGAGAGAGAGAGAGAGAGAGCTTTCTTTTCAAAGATTCAGTCAGCATATTTCCTTCTCTTATTAG
ACTTTGTATCTCATTGATCTTCTACTAATTAATTACATATTACATTTTACTTTCCTTCCTAAAAGGGTGCATTTTCCATACCAGGAG
ACTTTCTCATTTATTTTAACAAATGGAAAAGAATATAGACCCTTCTATAAGAGAACAGTGATATCATGTCAAAGTGTGAGGAAA
CTGTTCTGTCTAAAAGTTCCTGCATTAGCTGGGGACCCACAGGGTGGCCTTCTGATTTCTTCTGGAATAAAAGAAATAGAGG
CTTTATTGGACTTCATGCTTCTGTCAACCTCACCTCTTTTTTTTCNCTAAATCAACACTGGTGTTCACTTGCAGATTCATGGTGC
ATTAAAGCACTTTCCTCCGACAGTGAGCATGTGTGAGCTGCCACGGGGAATGGCTGCCAGCGCAGACAACTGAACCCAAAG
GAAGCCAAGTAAACAACTATCGTCAGGTAAGCATCAGCAAGGCACAGATTCTAGTATTTAAGTTTTGTATCTGTTTTCCTTCAT
TTGTTGAGTTCCAGCGAATGTCAGAGGTAGTGCTGTTCAAGGAGCCAGGCCTGCCTCTGCAGTCACCACCCTAGTGAATGTG
AGGAGAGCAGGGTCAGTCAACAGACAGAGTGCATGTAAGGAACTGTGAGTGGCAGTGGGCAAGTTACAGCCTGGTGTTCT
GCACAGTGACTGTCAAACAAAACTAACTACCCTATTCTCTCATGGCCCTGTGGATGGCAGGAGTCTCCTTTTCTCACAGAAAT
GAAGNNNNNNTGCTGTTGTTGTTGGGGCTGTTGTTTTTCAGTTTGCTGAAACCTCCTCTAGTCTGAAGGTTTCTGCTCAGTT
TTAATTAATAGGCTTCCAACTGCACTATGACCCCCAGCCTCTCTGTCTATCTATCCCTGTCTGTCTGTCTATTCACATTTGACTGC
TCCTCTGGCCTGTCCTTTCCCAGTGTGGCCACCTTGGTTGGTATTATCAGCCAATAACATGCCACACCTCTCCATCAAAGCTCC
CTTATTCATTATCAATAAACCTTCTCGACTTCTTTTCCCACTGTGGATGCACACATCGTTCACTTCTGCAGGTGCTTGTCCTGGG
CCTCAGGTGTTGCTGAGGACTGCACGCTGCCCCCTGCTGGTACAGTTTTGCCCTCAGCTGGGTCCACTCTCCACACAATGGC
CTTTCTCCTTCTGACCTTGATCTGTTTCCCTCACTAGTCTACNCTCNNNGGGGCAGGCAGCAGGTTTTATATACTTCAGCACT
GACTTGTTAGCTGAGCAAATGCTTTACAGAGTGAGTATTGAGTGGTGATTGCTAGTTTTGAAGCCTGTGGAGACTTTGGGAA
TGACTGAMAGGTGACAGAACCATGTGAAAGTCACTGTCAGCGATGACCTGGGCTGCGTTGGGACTGGTGTGYCCCACGCA
TTTAGCTTTKGGGCCCTTTCTTGATGAAAATATTGAAATTCATATTCTATATATAAGAAAATATAATACAGGCTAAATTAAAGTCA
AAGCATTTTCCTAGACCTTGAAACTTACANTTTTTTCTAAATTAAAAAGAAGTRGGAATATTTTCATGTAGCCCTTCAGTTATAT
GTTCCTCAGAACTGTGCCCACTGACACCACATTAAGGTGCCCTAGTGTGACACTGAAGAAAGAAATTGGGGACAAAGAAGC
CACCACCCCCACCCACCCTCCATGAGGGACAGGTGAGACTTCCATCTCCTGCCAAAGCTGGAAAGAGAGCAGCMACACTTT
GCAAGGACGCTCAACCCCTTTCACTTCCTGTTTCAGACATTTGACTTTTCAACTGCTATTTTACACAGCTCCTATTCACGTCTG
TAAAACAGCGTGGAGTTAAGGAGAGATCTGAGAGAGGACATGGAATGTCGGCAGAGTGTGGAGGGAGGTGGGCAGCAC
AGGCGGAAGGCCACAGGCTCCTGTAGCTTGTGCTTGCCTGGGTTGTTACAGAACTTCATGGGGTATAAATGGGCTATAGAAC
AGGCAGCTCACGTTCCTCACTTAGTTACAAAAATGTTCAATGGTATCTATAATGGAAACAAAGAGGAAGATGGCATTTAAAGC
ATTTACTAATGAATGGATACGTTCTTGAGACATGATCTCACTACAGGATCAGGCTATCCTTGAACTCACGATCCTCCTGCTTGG
GCACCTTAGCACCGTAGTTACTGCCAAGCACTGCCAGGCCCACCCAGCGTTGCCTTCACTATTTTTAAAGCACATCTACATATT
TGTTTGACTTGTCACTTGTGAGTTGACACATGAGGAAATCCAAAACCACTAAGATCGTCGGGTAAAATAGTGACCYCCTTCTC

TCACTTCCCACCAGGGCCAGCTCCTAGCTTGAAAAAGGTGTGTGGAAACCGCCACACAAGGGAAGCATTTGCTGTATATTTG
TGAACATGATTCACGGAAATGTTTTCCTCAGAGTACCCTACAGGCACAAAGGCTTCAGAAACTCACAGAAATTTTAGAGTAA
TAGTAATTACTGAAATTAGATTTTCAGTAATGAAAATCATTCATTCGTGAGAAACACTCCNANNTCCCAATTTCTYCCTTCCCTC
ACTGGGCTCACTAGTAGAATCATGGTTTGTTGGATCGTGAGCTGTGTTCACCAACCAGAGCCTCCTGAAGAAAGGATCACAC
AGTTGTTAGACCTCTTAATATTGAGAATTTGTATCTCCCTCTTCTGTGGGTTTCCCATTGCCCGGCTCCAGGCATCCCTCAGCT
GCACAACTCACCCTCAATGTCGTCTTCACCTCCGCAGAGGCTGAGCATGGTGGTCAGGGCGAGGACCCCCAGAATCAGAGC
TCTGCTGCACGGCATCCTGGTCTCTGGGAGGTCTCTGCAGCTGCTCTCCTGAGTCTTCAAGAGATTGTGAGGATCCAGCCAA
GGAGATCACACACCCAAACCAAACTCGCCAGAACCAATCAGAAAATTCCCTTGTGGTGACGTCACCGTTGCCAGCTGCAGA
CTTTGTATTACGAGTCCTGGAAGAGGCTANGGGAACTCCCTGGACCTGCTGCAGGCTCTGCTCAGAGTGGACCCAAGAGAA
TGCCTGGGAAGAAGNGGGTTCTAAGTCAGACCCTTTCTCTTGTGCTTCTGTGCACAGATGGCAAGAGAGGANGGGATTTCT
GAACAGCTTCATGGACAGGCGAGCTCATCCACGATAATATTGTTCTTCATAACTGTCTGGGCTGCACTGTTCAGAAAGTNSCC
CCTGAATTTATTTCTTTCCTGGTGCCCCTCTAACTTCGACACCCAGTTACTTTCTCCGGCTTGCATGCATCATGAGTTAGCTNG
GKTTTTCTTTATTTCTTCCCATAGTGTGTGCAGAGTACTCCAAATTGTGAGATGACCCACCTACCCAAGGTGTCCCCACAGACT
GTGGCCAGTCAATGGTTTTGCTCTTTCTGCTTGTGTAAACCCGGTTGGTGAGGTTCTTTCCCC

9.6.2. Prdm9

>WM msc09
TCAAGTATTGAAAGACAATGTGGGCAATATTTCAGTGATAAGTCAAATGTCAATGAGCACCAGAAGACACACACAGGGGAG
AAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAAGTCAGACCTCATCAAGCACCAGAGGACACACACAGG
GGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAAGTCAGTCCTCATCAAGCACCAGAGGACACACA
CAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAAGTCAGACCTCATCAAGCACCAGAGGACA
CACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAGCGAAGTCAAACCTCATCCAGCACCAGAG
GACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAGCGAAGTCAAACCTCATCCAGCACC
AGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAAGTCAGACCTCATCAA
GCACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAGCGAAGTCAAACCTC
ATCCAGCACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAAGTCAA
ACCTCATCAAGCACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAGAGAA
GTCAAACCTCATCCAGCACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAC
AGAAGTCAAACCTCATCAAGCACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTT
TACAGCGAAGTCAAACCTCATCCAGCACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGTGGG
GCTTTACACAGAAGTCAAACCTCATCAAGCACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGG
TGGGGCTTTACACAGAAGTCAGACCTCATCCAGCACCAGAGGACACATACAAGAGAGAAGTAA
>WM msc11
TCAAGTATTGAAAGACAATGTGGGCAATATTTCAGTGATAAGTCAAATGTCAATGAGCACCAGAAGACACACACAGGGGAG
AAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAGCGAAGTCAAACCTCATCCAGCACCAGAGGACACACACAGG
GGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAAGTCAGTCCTCATCCAGCACCAGAGGACACACA
CAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAAGTCAGACCTCATCAAGCACCAGAGGACA
CACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAGCGAAGTCAAACCTCATCCAGCACCAGAG
GACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAGAGAAGTCAAGCCTCATCAAGCACC
AGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGTGGGGCTTTACAGCGAAGTCAAACCTCATCCAG
CACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAAGTCAAACCTCAT
CAAGCACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAGCGAAGTCAAAC
CTCATCCAGCACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGTGGGGCTTTACACAGAAGTC
AAACCTCATCAAGCACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGTGGGGCTTTACACAGA
AGTCAGACCTCATCCAGCACCAGAGGACACATACAAGAGAGAAGTAA
>TM4 msc11
TCAAGTATTGAAAGACAATGTGGGCAATATTTCAGTGATAAGTCAAATGTCAATGAGCACCAGAAGACACACACAGGGGAG
AAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAGCGAAGTCAAACCTCATCCAGCACCAGAGGACACACACAGG
GGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAAGTCAGTCCTCATCCAGCACCAGAGGACACACA
CAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAAGTCAGACCTCATCAAGCACCAGAGGACA

CACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAGCGAAGTCAAACCTCATCCAGCACCAGAG
GACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAGAGAAGTCAAGCCTCATCAAGCACC
AGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGTGGGGCTTTACAGCGAAGTCAAACCTCATCCAG
CACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAAGTCAAACCTCAT
CAAGCACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAGCGAAGTCAAAC
CTCATCCAGCACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGTGGGGCTTTACACAGAAGTC
AAACCTCATCAAGCACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGTGGGGCTTTACACAGA
AGTCAGACCTCATCCAGCACCAGAGGACACATACAAGAGAGAAGTAA
>TM4 mmt1
TCAAGTATTGAAGGGCAATATTTCAGTGATAAGTCAAATGTCAATGAGCACCAGAAGACACACACAGGGGAGAAGCCCTAT
GTTTGCAGGGAGTGTGGGCGGGGCTTTACAGCGAAGTCAAGCCTCGTCCAGCACCAGAGGACACACACAGGGGAGAAGC
CCTATGTTTGCAGGGGGTGTGGGCGGGGCTTTACAGCGAAGTCAAGCCTCATCCAGCACCAGAGGACACACACAGGGGAG
AAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAACTCACACCTCATCAAGCACCAGAGGACACACACAGG
GGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTATAGCGAAGTCAAACCTCATCCAGCACCAGAGGACACACA
CAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAACGAAGTCAGACCTCATCAAGCACCAGAGGACA
CACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAACTCACACCTCATCAAGCACCAGAG
GACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTATAGCGAAGTCAAACCTCATCCAGCACC
AGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAACGAAGTCAGACCTCATCAA
GCACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAACTCACACCTCA
TCAAGCACCAAAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAAGTCAGA
CCTCATCCAGCACCAGAGGACACATACAAGAGAGAAGTAA
>TD5 dom11
TCAAGTATTGAAAGACAATGTGGGCAATATTTCAGTGATAAGTCAAATGTCAATGAGCACCAGAAGACACACACAGGGGAG
AAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAACTCACACCTCATCCAGCACCAGAGGACACACACAGG
GGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAAGTCAGACCTCATCAAGCACCAGAGGACACACA
CAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAAGTCAGACCTCATCAAGCACCAGAGGACA
CACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAAGTCAGTCCTCATCAAGCACCAGAG
GACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAAGTCAGTCCTCATCAAGCACC
AGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAGCGAAGTCAGTCCTCATCCAG
CACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAGCGAAGTCAAACCTCAT
CCAGCACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAAGTCAGTC
CTCATCCAGCACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAGCGAAGTC
AGTCCTCATCCAGCACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAGCGA
AGTCAGTCCTCATCCAGCACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACA
CAGAAGTCAAACCTCATCAAGCACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGTGGGGCTT
TACACAGAAGTCAGACCTCATCCAGCACCAGAGGACACATACAAGAGAGAAGTAA
>TD5 mmt1
TCAAGTATTGAAGGGCAATATTTCAGTGATAAGTCAAATGTCAATGAGCACCAGAAGACACACACAGGGGAGAAGCCCTAT
GTTTGCAGGGAGTGTGGGCGGGGCTTTACAGCGAAGTCAAGCCTCGTCCAGCACCAGAGGACACACACAGGGGAGAAGC
CCTATGTTTGCAGGGGGTGTGGGCGGGGCTTTACAGCGAAGTCAAGCCTCATCCAGCACCAGAGGACACACACAGGGGAG
AAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAACTCACACCTCATCAAGCACCAGAGGACACACACAGG
GGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTATAGCGAAGTCAAACCTCATCCAGCACCAGAGGACACACA
CAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAACGAAGTCAGACCTCATCAAGCACCAGAGGACA
CACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAACTCACACCTCATCAAGCACCAGAG
GACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTATAGCGAAGTCAAACCTCATCCAGCACC
AGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACAACGAAGTCAGACCTCATCAA
GCACCAGAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAACTCACACCTCA
TCAAGCACCAAAGGACACACACAGGGGAGAAGCCCTATGTTTGCAGGGAGTGTGGGCGGGGCTTTACACAGAAGTCAGA
CCTCATCCAGCACCAGAGGACACATACAAGAGAGAAGTAA

10. A Note of Thanks

Throughout the eventful time I have spent writing this thesis several people have helped me tremendously to get me where I am standing now.

First and foremost I would like to extend my gratitude to Dr. Linda Odenthal-Hesse who not only took a chance on me when I asked her for an internship last summer but also supported me throughout all steps. I have learned many valuable lessons from her and got to experience what it means to be a Scientist from here. I am incredibly thankful and proud of all she has helped me achieve.

Further I would like to thank Prof. Dr. Tal Dagan for agreeing to be my first examiner.

Very importantly my thanks also goes to Nicole Thomsen, who not only taught me the way around her lab in an especially patient way but helped me overcome the challenges I faced there.

Next I would like to thank Dr. Elena Damm who was always willing to help me with any questions and problems I faced both in and outside the lab and provided me with valuable insight.

Lastly I would like to thank my family, my Parents and my siblings, who did their very best to support me emotionally throughout this process. Thank you for listening to me, whatever and whenever it was.

Declaration

I hereby declare that I have prepared this thesis independently and without outside assistance.

I have not used any sources or aids other than those indicated.

The submitted written version of the thesis corresponds to the one on the electronic storage medium.

Furthermore, I certify that this work has not been submitted as a thesis elsewhere.

Kiel, 15. September 2022