Characterisation of the M-locus and functional analysis of the male-determining gene in the housefly
Wu, Yanli

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2018

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 2
Characterisation of the complex nature of $M$-loci in *Musca domestica*

2.1 Abstract

The housefly (*Musca domestica*) is a perfect model to study insect sex determination as it harbours various systems. An *M*-locus that contains the male-determining gene(s) is typically located on the Y-chromosome, but can also be present on any of the five autosomes or even the X-chromosome. Recently, based upon a differential transcriptome analysis of early male and female embryos, “orphan reads” (ORMs) were identified as possible transcripts from the male-determining gene, *Mdmd* (for *Musca domestica* male determiner), to reside in the *M*-locus. To further investigate the nature of the *M*-locus, I used these ORM sequences to find adjacent genomic DNA sequences. I found that the *M*III-locus (*M*-locus on chromosome III) and the *M*V-locus (*M*-locus on chromosome V) contain multiple copies of sequences, with various level of homology to each other. Cladogram analysis further demonstrated that sequences in the *M*III-locus and the *M*V-locus could be divided into different clades, with sequences within clades being more similar than sequences between clades. Interestingly, the *M*III-locus and the *M*V-locus share some similar sequences. These results are most easily explained by assuming that there have been independent amplification events before and after the translocation of the *M*-locus to autosomes III and V, possibly from the Y-chromosome.
2.2 Introduction

Various insect sex determination systems exist that can be variable even within species (Sánchez, 2004; Bachtrog et al., 2014; Beukeboom and Perrin, 2014; Blackmon et al., 2017). How this diversity of insect sex determination systems has evolved still remains unclear. The housefly, *Musca domestica*, harbours several sex determination systems and is therefore a perfect model to study the evolution of sex determination. An *M*-locus that contains the male-determining gene(s) is typically located on the Y-chromosome, but can also be present on any autosome or even the X-chromosome (Wagoner, 1969; Inoue and Hiroyoshi, 1982; Denholm et al., 1983; Inoue et al., 1986). *Md-transformer* (*Mdtra*) was identified as the female-determining gene in the *M. domestica* sex determination pathway (Hediger et al., 2010). *Mdtra* mRNA and *Mdtra2* mRNA are maternally provided to kick-start a positive autoregulatory feed-back loop of female-specific splicing of *Mdtra* mRNA in the zygote (Bopp, 2010). MdTRA protein leads to female-specific splicing of *Mdtra* mRNA with the assistance of other essential co-factors such as MdTRA2 protein (Hediger et al., 2010). The mRNA of the *M. domestica* doublesex homologue, *Mddsx*, is spliced by MdTRA protein and its co-factor MdTRA2 protein into the female variant, which leads to female development (Burghardt et al., 2005; Hediger et al., 2010). The action of the male-determining gene(s) is the interruption of this autoregulatory loop. This results in male-specific splicing of *Mdtra* mRNA, yielding a non-functional MdTRA truncated protein (Hediger et al., 2010). Hence, in the presence of a male-determining gene(s), *Mddsx* is spliced into its male-specific isoform, leading to male development.

A differential transcriptome analysis on early unisexual embryos identified four transcript parts among the top male-specifically expressed sequences that were also absent in the female genome (Scott et al., 2014; Sharma et al., 2017). These “orphan sequences” were termed ORM#1, ORM#2, ORM#3 and ORM#6 (Sharma et al., 2017). PCR amplification from the genomes of *M*III males (*M*-locus on autosome III) with primers located in these ORMs confirmed that all four ORMs belong to the same gene. This candidate male-determining gene was named *Mdmd* (for *Musca domestica* male determiner) (Sharma et al., 2017). *Mdmd* is only present in the male genome (Sharma et al., 2017). Fig. 2.1 shows the order of ORMs in the *Mdmd* assembly.
Silencing of *Mdmd* by RNAi confirmed that *Mdmd* is necessary for testes differentiation (Sharma et al., 2017). Moreover, knockout of *Mdmd* by CRISPR-Cas9 resulted in complete feminisation, indicating that *Mdmd* plays an important role in male development. Disruption of *Mdmd* also affected its downstream gene *Mdtra* and *Mddsx*. When *Mdmd* is disrupted by CRISPR-Cas9, *Mdtra* is spliced in the female variant in sex-reversed individuals (Sharma et al., 2017). Similarly, the female splice variant of *Mddsx* was also detected in sex-reversed individuals, in contrast to the male splice variant of control males (Sharma et al., 2017). These results confirmed that *Mdmd* plays an important role in male development and serves as the primary signal in the *M. domestica* sex determination pathway.

Several questions remain about the structure of the *M*-locus and the function of *Mdmd*. How *Mdmd* is embedded in the *M*-locus remains unknown, and the regions adjacent to the *Mdmd* ORMs have not been determined yet. Moreover, it is not yet known whether expression of *Mdmd* is sufficient to turn genotypic females into males or whether additional genes are involved. Identifying these genomic regions adjacent to the orphan contigs will provide molecular evidence for the organisation of the *M*-locus and help to characterize the complete sequence of *Mdmd*. In this chapter, I describe the genomic regions adjacent to ORM#1 and ORM#6 in two autosomal *M* strains, *M*III (*M*-locus on chromosome III) and *M*V (*M*-locus on chromosome V) by genome walking (Siebert et al., 1995). I present evidence that the *M*-loci in both strains contain multiple copies of sequences, that all show various level of homology to each other. I further investigate whether the *M*-loci also contain interspersed genomic sequences that exist both in the male and the female genome. In addition, I describe the common sequences shared by the *M*III-locus and the *M*V-locus. These results contribute to a further understanding of sex chromosome evolution in *M. domestica*.

Figure 2.1: The position of ORMs in *Mdmd*: ORM#1 is located on the 5’region and ORM#6 on the 3’region. ORM#3 is spanning a small intron and ORM#2 is located in the middle part of *Mdmd*. MIF4G and MA3 are two conserved domains.
2.3 Materials and Methods

2.3.1 Musca domestica strains and culturing

Two different *M. domestica* strains were used for genome walking analysis. (1) 3-6 *M*III strain: *M* is located on autosome III. Females have genotypes *X/X; pw bwb w/pw bwb w* and males *X/X; pw* *M*III *bwb* + *w* + *bwb w*. *pw* stands for pointed wings, *bwb* for brown body and *w* for white eyes, all being recessive visible markers on autosome III. Females have brown body, white eyes and notched wings. Males are heterozygous for *M* and they have black body, white eyes and normal wings. (2) 35-4 *M*V strain: *M* is located on autosome V. Females are *X/X; bwb/bwb; ocra/ocra*, males are *X/X; bwb/bwb; M*V *ocra* + / + *ocra*. *ocra* is a recessive yellow eye colour marker on autosome V. Females are phenotypically brown body with yellow eyes. Males are heterozygous for *M* and they have brown body with red eyes. Strains were reared at 25°C as described previously (Schmidt et al., 1997).

2.3.2 Genome walking

DNA of single adult males from the *M*III and *M*V strains was extracted by NucleoSpin® Tissue Genomic DNA purification kit from Macherey Nagel (Düren, Germany). Genome walking was performed according to Universal GenomeWalker™ 2.0 User Manuel from Clontech (Fig. 2.2; California, United States). The concentration of experimental genomic DNA was checked in Nanodrop from Thermo Fisher Scientific (Massachusetts, United States). The size and the quality of genomic DNA were checked on a 0.6% agarose/EtBr gel and the size of genomic DNA should be larger than 50kb with minimum smearing. Subsequently, to test whether the genomic DNA can be digested by restriction enzymes, the experimental genomic DNA was digested by DraI (TTT|AAA) with the following concentrations: 5 µL Experimental genomic DNA (0.1 µg/µL), 1.6 µL DraI (10 units/µL), 2 µL 10×DraI Restriction Buffer in a total volume of 20 µL. After incubation at 37°C overnight, 5 µL of digested products were analysed on a 0.6% agarose/EtBr gel along with 0.5 µL of undigested experimental genomic DNA as a control. A smear was observed in the gel, indicating that the experimental genomic DNA can be digested by restriction enzymes. Subsequently, genomic DNA was digested separately by four enzymes provided by the kit: DraI (TTT|AAA), EcoRV (GAT|ATC), PvuII (CAG|CTG) and Stul (AGG|CTT). Each enzyme digested the genomic DNA separately with the following concentrations: 25 µL Genomic DNA (0.1 µg/µL), 8 µL Restriction enzyme (10 units/µL), 10 µL 10×Restriction enzyme buffer in a total volume of 100 µL. After incubation at 37°C for 2 hrs, the reaction was vortexed at slow
speed for 5-10 sec and incubated at 37°C overnight afterwards. The reaction products were checked on 0.6% agarose/EtBr gel.

After digestion, DNA fragments were purified by the NucleoSpin® Gel and PCR Clean-up kit from Macherey Nagel (Düren, Germany) and ligated to the GenomeWalker adaptors to establish so-called GenomeWalker™ “libraries”. After building four libraries, a primary “touchdown” PCR was performed with primer pairs GSP1 and AP1 for 5’_genome walking and GSP3 and AP1 for 3’_genome walking (Fig. 2.2, primer sequences were shown in appendix). The following concentrations and conditions were used for the primary PCR: 1 µL DNA library, 0.5 µL 10 µM forward primer, 0.5 µL 10 µM reverse primer, 0.5 µL 10 mM dNTP, 2.5 µL 10×Advantage 2 PCR Buffer and 0.5 µL Advantage 2 Polymerase Mix (50×) in a total volume of 25 µL; 7 cycles of 94°C denaturation for 25 sec, annealing/extension for 6 min at 72°C, followed by 32 cycles of 94°C denaturation for 25 sec, annealing/extension at 67°C for 6 min, and finally extension at 67°C for 7 min.

After primary PCR, a secondary (nested) “touchdown” PCR was performed by taking 1 µL of 50×diluted primary PCR product. The primers for the secondary PCR are AP2 and GSP2a or GSP2b, respectively, for 5’_genome walking and AP2 and GSP4a or GSP4b, respectively, for 3’_genome walking. Compound concentrations were the same as for the primary PCR. The following conditions were used for the secondary PCR: 5 cycles of 94°C denaturation for 25 sec, annealing/extension for 6 min at 72°C, followed by 20 cycles of 94°C denaturation for 25 sec, annealing/extension at 67°C for 6 min, and finally extension at 67°C for 7 min. PCR products were analysed on a 1% agarose/EtBr gel.

Target fragments were purified with the NucleoSpin® Gel and PCR clean-up kit from Macherey Nagel (Düren, Germany) and subsequently cloned according to the TA Cloning® Kit, with pCR®II vector from Clontech (California, United States) under the following concentrations and conditions: 1-5.5 µL DNA (DNA from gel purification was diluted in 20 µL water), 2 µL 5×Express Link™T4 DNA Ligase Buffer, 1.5 µL linearised pCR®II vector (25 ng/µL) with a total volume of 9 µL, 1 µL Express Link™T4 DNA Ligase (5 Weiss units) was added into the reaction to reach the final volume of 10 µL. Alternatively, if the PCR products only show a single fragment, 1 µL of PCR products can be directly ligated into the pCR®II vector without gel and PCR purification. The ligation concentrations for the rest of the components were the same as above. Ligation was performed at 16°C overnight. The construct was used to transform competent E. coli DH5α. The pCR®II vector contains the lacZα gene that allows for blue-white screening of
positive colonies by α-complementation. White colonies were cultured in Luria-Bertani (LB) medium that contain 100 μg/mL ampicillin at 37°C overnight. Plasmids were extracted the following day and the size of inserted DNA fragments was checked by EcoRI-HF® (G|AATTC) from NEB (Massachusetts, United States) digestion. LGC Genomics (Berlin, Germany) carried out sequencing of the candidate fragments by using the primers M13F and M13R located in the vector.

The primers GSP_Dra52_R2 and GSP_Dra52_R1 combined with AP1 and AP2, respectively, were used for a second round of 5′_genome walking of both the $M^{III}$ and $M^{V}$ strains. $M^{III}_{GSP}$_Stu93_F1 and $M^{III}_{GSP}$_Stu93_F2 combined with AP1 and AP2, respectively, were used for a second round of 3′_genome walking of the $M^{III}$ strain. $M^{V}_{GSP}$_Dra13B_F1 and $M^{V}_{GSP}$_Dra13B_F2 combined with AP1 and AP2, respectively, were used for another second round of 3′_genome walking of the $M^{V}$ strain. A third round of genome walking was performed after having obtained new sequences from the second round of genome walking. The primers $M^{III}_{GSP}$_Pvu3B_R1 and $M^{III}_{GSP}$_Pvu3B_R2 combined with AP1 and AP2, respectively, were used for a third round of 5′_genome walking of the $M^{III}$ strain. $M^{V}_{GSP}$_Pvu7B_R1 and $M^{V}_{GSP}$_Pvu7B_R2 combined with AP1 and AP2, respectively, were used for another third round of 5′_genome walking of the $M^{V}$ strain.

![Genome walking diagram](image)

Figure 2.2: Genome walking. Genomic DNA was digested by four enzymes: DraI, EcoRV, PvuII and Stul. Each enzyme digested the genomic DNA separately. After digestion, GenomeWalker™ adaptors were annealed to the DNA. Gene specific primer GSP1 and adaptor primer AP1 are primers for primary PCR. GSP2 and AP2 are primers for secondary PCR. N: Amine group blocks extension of the 3′ end of the adaptor-ligated genomic fragments, preventing the generation of an AP1 binding site in lower adaptor strand (if double-stranded adaptor sequences are present at both ends, they will form a “panhandle” structure that cannot be extended) (modified from Universal GenomeWalker™ 2.0 User Manual from Clontech, California, United States).
2.3.3 Rapid amplification of cDNA ends (RACE)

5’_RACE PCR and 3’_RACE PCR were performed using the SMARTer™ RACE cDNA Amplification Kit from Clontech (California, United States) according to the SMARTer™ RACE cDNA Amplification Kit User Manual. RNA was purified from 0-24 hrs embryos from the MIII strain with the ZR Tissue & Insect RNA Micro Prep™ kit from Zymo Research (California, United States). First-strand of the 5’_RACE_Ready cDNA and the 3’_RACE_Ready cDNA was synthesised according to the SMARTer™ RACE cDNA Amplification Kit User Manual.

After synthesis of first-strand of the 5’_RACE_Ready cDNA and the 3’_RACE_Ready cDNA, a “touchdown” PCR was performed with primer pairs Universal Primer A Mix (UPM) and GSP1 or GSP2b, respectively, for 5’_RACE PCR and UPM and GSP4b for 3’_RACE PCR. The following concentrations and conditions were used for the PCR: 2.5 µL RACE_Ready cDNA, 1 µL 50×UPM, 1 µL 10 µM primer, 1 µL 10 mM dNTP, 5 µL 10×Advantage 2 PCR Buffer and 1 µL Advantage 2 Polymerase Mix (50×) in a total volume of 50 µL; 5 cycles of 94°C denaturation for 30 sec, annealing/extension for 5 min at 72°C, followed by 5 cycles of 94°C denaturation for 30 sec, annealing at 70°C for 30 sec and extension at 72°C for 5 min, and finally 25 cycles of 94°C denaturation for 30 sec, annealing at 68°C for 30 sec and extension at 72°C for 5 min. The PCR products were checked on 1% agarose/EtBr gel. The cloning procedure was the same as the cloning step in genome walking. Plasmids from white colonies were extracted and the size of inserted DNA fragments was checked by EcoRI-HF® (G|AATTC) from NEB (Massachusetts, United States) digestion. Sequencing of the candidate fragments with the primers M13F and M13R in the vector was carried out by LGC Genomics (Berlin, Germany).

2.3.4 PCR amplification of M-locus sequences

PCR was performed on single male housefly gDNA and cDNA with the primer combinations GSP2b-Dra-52-F or GSP1-9-F and GSP3-R or GSP4b-R, respectively. GSP2b-Dra-52-F and GSP1-9-F are located on the newly yield sequences from genome walking and RACE. The following concentrations and conditions were used in gDNA PCR: 100 ng gDNA, 0.5 µL 10 µM forward primer, 0.5 µL 10 µM reverse primer, 3 µL 2.5 mM dNTP, 3 µL 10×Advantage 2 PCR Buffer and 0.5 µL Advantage 2 Polymerase Mix (50×) in a total volume of 30 µL; followed by denaturation at 94°C for 2 min, then 30 cycles of 94°C denaturation for 30 sec, annealing at 70°C for 30 sec and extension at 72°C for 7 min, and lastly extension at 72°C for 10 min. For the cDNA PCR, cDNA was first synthesised with the Thermo Fisher Scientific (Massachusetts, United States) Maxima First Strand
cDNA Synthesis Kit with the following concentrations and conditions: 4 µL 5×Reaction Mix, 2 µL Maxima Enzyme Mix and 1.5 µL template RNA (1.2 µg/µL) in a total volume of 20 µL. The mixture was incubated at 25°C for 10 min followed by 30 min at 50°C. The reaction was terminated by incubating at 85°C for 5 min. The cDNA was diluted 5× with nuclease-free water after synthesis and 1 µL cDNA was used in each PCR reaction. The cDNA PCR was performed under the same conditions as gDNA PCR.

PCR products were analysed on a 1% agarose/EtBr gel. The cloning procedure was the same as the cloning step in genome walking. Plasmids from white colonies were extracted and the size of inserted DNA fragments was checked by EcoRI-HF® (G|AATTC) from NEB (Massachusetts, United States) digestion. The candidate fragments from positive plasmids were sequenced with the primers M13F and M13R in the vector combined with PCR primers by LGC Genomics (Berlin, Germany).

2.3.5 Sequence analysis

Sequencing data were analysed with “Geneious” (Kearse et al., 2012). Nucleotide multiple alignment was used in the data processing. A BLAST survey of all unique Mdm sequences against database Genome (Musca_domestica-2.0.2 reference Annotation Release 102) and organism Musca domestica (taxid: 7370) (Scott et al., 2014) was performed using the NCBI on-line blast tool. Phylogenetic trees were built with the Geneious tree builder that was based on the Jukes-Cantor genetic distance model and the Neighbor-joining method combined with the bootstrap resampling method (1000 replicates).

2.4 Results

2.4.1 The MIII-locus consists of multiple copies of Mdm

The sequences obtained from genome walking revealed that the MIII-locus consists of multiple copies of sequences, with various level of homology to each other (Fig. 2.3). The first round of genome walking yielded four new and different sequences (sequences #1-4) with ORM#1 based primers and six (sequences #5-10) with ORM#6 based primers. Each sequence from 3’_genome walking might be connected to any sequence from 5’_genome walking. Hence, the MIII-locus contains at least six copies of Mdm. Among the obtained sequences, the longest is around 2.4kb (sequence #5), and the shortest 260bp (sequence #2). Sequences #1, #9 and #10 contain genomic sequences that exist in both the male and the female genome. The genomic sequences in sequences #9 and #10
Chapter 2

share identical parts. Among these ten newly obtained sequences, some are completely different from each other, but some are partly similar (displayed in similar colours: from dark green to light green (sequences #4 and #5) in Fig. 2.4A, from dark blue to light blue (sequences #6, #7 and #8) in Fig. 2.4B and from red to light red (sequences #9 and #10) in Fig. 2.4C. Alignment of these partly similar sequences reveals that they have indels (insertion or deletion mutations) and nucleotide variations. Since the genomic DNA comes from a single male, nucleotide variations cannot be population polymorphism, but most come from independent repeats of the *M*III-locus.

Alignment of these partly similar sequences reveals that they have indels (insertion or deletion mutations) and nucleotide variations. Since the genomic DNA comes from a single male, nucleotide variations cannot be population polymorphism, but most come from independent repeats of the *M*III-locus.

Figure 2.3: The *M*III-locus consists of multiple copies of sequences with various levels of homology to each other. 5’ genome walking started from ORM#1 and 3’ genome walking started from ORM#6 on the *M*III strain genomic DNA. GSP1 and GSP2b are primers for 5’ genome on ORM#1. GSP3 and GSP4b are primers for 3’ genome walking on ORM#6. Sequences #1-4 overlapped with ORM#1. Sequences #5-10 overlapped with ORM#6. Different sequences are colour coded (5’ yellow, orange, grey and green, 3’ green, blue and red). Partly similar sequences marked with different colour intensities (e.g. sequences marked from dark green to light green) and labeled with A, B and C. The dotted line represents potential *Mdmd* homologous sequences of so far unknown variations. The shaded boxes indicate sequences that exist in both the male and the female genome.

Figure 2.3: The *M*III-locus consists of multiple copies of sequences with various levels of homology to each other. 5’ genome walking started from ORM#1 and 3’ genome walking started from ORM#6 on the *M*III strain genomic DNA. GSP1 and GSP2b are primers for 5’ genome on ORM#1. GSP3 and GSP4b are primers for 3’ genome walking on ORM#6. Sequences #1-4 overlapped with ORM#1. Sequences #5-10 overlapped with ORM#6. Different sequences are colour coded (5’ yellow, orange, grey and green, 3’ green, blue and red). Partly similar sequences marked with different colour intensities (e.g. sequences marked from dark green to light green) and labeled with A, B and C. The dotted line represents potential *Mdmd* homologous sequences of so far unknown variations. The shaded boxes indicate sequences that exist in both the male and the female genome.
Characterisation of the complex nature of M-loci in Musca domestica

Figure 2.4: Alignment of partly similar sequences obtained from genome walking of the $M^{III}$ strain.

A: Alignment of similar sequences marked from dark green to light green (sequences #4 and #5).

B: Alignment of similar sequences marked from dark blue to light blue (sequences #6, #7 and #8).

C: Alignment of similar sequences marked from red to light red (sequences #9 and #10).

Sequence #4 was obtained from 5’_genome walking with the primers GSP1 and GSP2b in ORM#1. Sequences #5, #6, #7, #8, #9 and #10 were obtained from 3’_genome walking with the primers GSP3 and GSP4b in ORM#6. All these partly similar sequences can align with each other but with indels (insertion or deletion mutations) and nucleotide variations. The horizontal bars indicate the presence of the same sequences and the lines indicate indels. The vertical lines in the bars indicate nucleotide variations among sequences.

More evidence indicated that the $M^{III}$-locus contains multiple copies of Mdmd. Sequence #4 might be a “bridge” between Mdmd homologous sequences as it went out from ORM#1 and into ORM#6 (Fig. 2.5). Also, sequence #5, which is similar to sequence #4, went out from ORM#6 and into ORM#1 and part of ORM#3 (Fig. 2.5). These results demonstrate that the $M^{III}$-locus consists of tandem copies of Mdmd repeats and some of the copies are quite similar.

Figure 2.5: Sequences from genome walking linked several copies of Mdmd. Sequence #4 went from ORM#1 into ORM#6. Sequence #5 went from ORM#6 into ORM#1 and part of ORM#3. Sequence #4 was obtained from 5’_genome walking with the primers GSP1 and GSP2b in ORM#1. Sequence #5 was obtained from 3’_genome walking with the primers GSP3 and GSP4b in ORM#6. The dotted lines are undefined sequences of potential Mdmd repeats, and the solid lines are known sequences.

The $M^{III}$-locus contains interspersed genomic sequences that exist in both the male and the female genome. However, those sequences represent mostly repetitive sequences that cannot be used to design primers for further genome
walking. For example, sequence #1 contains such sequences, as shown in Fig. 2.3. To characterise the whole \(M^{III}\)-locus, I performed a second and third round of 5' genome walking based on some of the male specific sequences acquired in the previous round. I obtained three new sequences containing genomic sequences that exist in both the male and the female genome. Walking out from sequence #2 yielded sequences #2-1 and #2-2 (Fig. 2.6A). Interestingly, sequence #2-1 overlapped with \textit{Musca_Mariner_Like_Elements} (MLEs), and sequence #2-2 with sequences from tapeworms and trematodes. In the third round of 5' genome walking, I used primers in the male specific part of sequence #2-2, yielding sequence #2-2-1 (Fig. 2.6B). This sequence also overlapped with \textit{Musca_Mariner_Like_Elements} (MLEs).

When walking out from sequence #9 in the second round of 3' genome walking, I acquired four diverse sequences. All of them contained genomic sequences that exist in both the male and the female genome (Fig. 2.6C). Sequences #9-3 and #9-4 are very similar with two nucleotide differences (Fig. 2.6C). Sequences #9-2, #9-3 and #9-4 overlapped with the \textit{Musca domestica} pre-mRNA-splicing factor \textit{CWC22} homolog, which contains two conserved domains MIF4G and MA3. Genomic sequences “b” in #9-3 and “c” in #9-4 are identical, which are also similar with “a” in #9-2. Since ORM#3 overlapped with conserved domain MIF4G, I also aligned sequences #9-2, #9-3 and #9-4 with ORM#3. I found that sequence #9-2 overlapped with part of ORM#3 (Fig. 2.7A). Sequences #9-3 and #9-4 overlapped with part of ORM#3 and with part of sequence #9 (Fig. 2.7B). These results confirmed that the \(M^{III}\)-locus consists of multiple tandemly repeated, partially truncated copies of \textit{Mdmd} interspersed by genomic sequences that exist in both the male and the female genome.

In addition, I performed RACE PCR to characterise the \(M^{III}\)-locus at the cDNA level. I obtained eight sequences (Fig. 2.8A). Interestingly, sequences #1-RACE and #2-RACE are identical with sequences #1 and #2 from genome walking, respectively. When I compared sequence #3-RACE with sequence #2 from genome walking, I only found one nucleotide change (data not shown). Sequence #4-RACE is partly similar with sequence #2 from genome walking and sequence #5-RACE is partly similar with sequence #4 from genome walking. The results from RACE PCR indicated that some of the copies in the \(M\)-locus are transcribed into RNA. I also performed PCR with primers in the newly obtained sequences and in ORM#6, yielding two sequences (Fig. 2.8B). Alignment of these two partly similar sequences shows that they have indels (insertion or deletion mutations) and nucleotide variations (Fig. 2.8C).
Figure 2.6: Results of second and third round of genome walking of the \( M^{\text{III}} \) strain. The \( M^{\text{III}} \)-locus contains interspersed genomic sequences that exist in both the male and the female genome. A: Second round of 5’-genome walking started from sequence #2, yielding two new sequences: sequences #2-1 and #2-2. B: Third round of 5’-genome walking started from sequence #2-2, yielding one new sequence: sequence #2-2-1. C: Second round of 3’-genome walking started from sequence #9, yielding four new sequences: sequences #9-1, #9-2, #9-3 and #9-4. “a”, “b” and “c” are similar genomic sequences. GSP_Dra52_R2 and GSP_Dra52_R1 are primers for the second round of 5’-genome walking. \( M^{\text{III}} \_\text{GSP_Pvu3B_R1} \) and \( M^{\text{III}} \_\text{GSP_Pvu3B_R2} \) are primers for the third round of 5’-genome walking. \( M^{\text{III}} \_\text{GSP_Stu93_F1} \) and \( M^{\text{III}} \_\text{GSP_Stu93_F2} \) are primers for the second round of 3’-genome walking. The shaded boxes indicate sequences that exist in both the male and the female genome. The dotted lines are undefined sequences of potential \( Mdmd \) repeats. In sequence alignment, the horizontal bars indicate the presence of the same sequences and the vertical lines in the bars indicate nucleotide variations among sequences.
Figure 2.7: Sequences from genome walking showing the complexity of the $M_{III}$ locus. A: Sequence #9-2 overlapped with part of ORM#3. B: Sequences #9-3 and #9-4 overlapped with part of ORM#3 and with part of sequence #9. Sequences #9-2, #9-3 and #9-4 were obtained from the second round of 3' genomic DNA walking with the primers $M_{III}$ GSP_Stu93_F1 and $M_{III}$ GSP_Stu93_F2. The shaded boxes indicate sequences that exist in both the male and the female genome. The dotted lines are undefined sequences, and the solid lines are known sequences of potential Mdmd repeats.
Figure 2.8: RACE PCR to determine the $M^{III}$-locus on the cDNA level. A: The 5’_RACE PCR started from ORM#1. GSP1 or GSP2b are the primers for the 5’_RACE PCR on ORM#1. Sequences #1-6-RACE overlapped with ORM#1. The 3’_RACE PCR started from ORM#6. GSP4b is the primer for the 3’_RACE PCR on ORM#6. Sequences #7-8-RACE overlapped with ORM#6. B: PCR with primers GSP1-9-F and GSP2b-Dra52-F in the newly obtained sequences and primers GSP4b-R and GSP3-R in ORM#6, respectively, yielding two sequences. C: Alignment of these two partly similar sequences shows that they have indels (insertion or deletion mutations) and nucleotide variations. The dotted lines are undefined sequences, and the solid lines are known sequences of potential $Mdmd$ repeats. In sequence alignment, the horizontal bars indicate the presence of the same sequences and the lines indicate indels. The vertical lines in the bars indicate nucleotide variations among sequences.
2.4.2 Multiple copies of Mdmd exist in $M^{II}$, $M^{III}$, $M^{V}$ and $M^{Y}$ males

Genomic DNA from $M^{I}$, $M^{II}$, $M^{III}$, $M^{V}$ and $M^{Y}$ males was amplified with divergent primers localised at ORM#1 and ORM#6. The results are displayed in Fig. 2.9, which was kindly provided by Claudia Brunner. It revealed that there are multiple copies of Mdmd in males from the $M^{II}$, $M^{III}$, $M^{V}$ and $M^{Y}$ strains but not the $M^{I}$ strain, which probably has a different male-determining gene(s). The $M^{III}$-locus contains at least six copies of Mdmd. The $M^{Y}$-locus seems less complicated, as there are only two fragments amplified with divergent primers. For finding sequences adjacent to ORM#1 and ORM#6 in the $M^{Y}$ strain, I also performed genome walking in this strain.

![Image of Figure 2.9](image)

Figure 2.9: Multiple copies of Mdmd exist in $M^{II}$, $M^{III}$, $M^{V}$ and $M^{Y}$ males. A: Sequences between Mdmd were amplified by divergent primers 1as localised at ORM#1 and 6as localised at ORM#6. B: Multiple fragments were amplified in $M^{II}$, $M^{III}$, $M^{V}$ and $M^{Y}$ males indicating that there were multiple tandemly repeated copies in the $M$-loci. The dotted lines are undefined sequences. This figure was kindly provided by Claudia Brunner from University of Zürich.
2.4.3 The $M^v$-locus consists of multiple copies of Mdmd

Genome walking in the $M^v$ strain revealed that the $M^v$-locus also contains repetitive sequences (Fig. 2.10). The first round of genome walking yielded six new and different sequences (sequences #11-16) with ORM#1 based primers and four (sequences #17-20) with ORM#6 based primers. Each sequence from 5’_genome walking might be connected to any sequence from 3’_genome walking. Hence, the $M^v$-locus contains at least six copies of Mdmd. Among the obtained sequences, the longest is around 2.1kb (sequence #20), and the shortest is 195bp (sequence #16). Sequence #20 contains genomic sequences that exist in both the male and the female genome. Similar sequences are displayed in similar colours: from dark purple to light purple (sequences #12, #13 and #14) in Fig. 2.11A, from dark orange to light orange (sequences #15 and #16) in Fig. 2.11B and from dark pink to light pink (sequences #18, #19 and #20) in Fig. 2.11C. Alignment of these partly similar sequences shows that they have indels (insertion or deletion mutations) and nucleotide variations. Since the genomic DNA comes from a single male, nucleotide variations cannot be population polymorphism, but most come from independent repeats of the $M^v$-locus.

Figure 2.10: The $M^v$-locus consists of multiple copies of sequences with various level of homology to each other. 5’_genome walking started from ORM#1 and 3’_genome walking started from ORM#6 on the $M^v$ strain genomic DNA. The primer GSP1 combined with GSP2a and GSP2b, respectively, were used for 5’_genome walking on ORM#1 and the primer GSP3 combined with GSP4a and GSP4b, respectively, were used for 3’_genome walking on ORM#6. Sequences #11-16 overlapped with ORM#1. Sequences #17-20 overlapped with ORM#6. Different sequences are colour coded. Partly similar sequences are marked with different colour intensities (e.g. sequences marked from dark purple to light purple) and labeled with A, B and C. The dotted line is an undefined sequence. The shaded boxes indicate sequences that exist in both the male and the female genome.
Figure 2.11: Alignment of partly similar sequences obtained from genome walking of the $M^V$ strain. A: Alignment of similar sequences marked from dark purple to light purple (sequences #12, #13 and #14). B: Alignment of similar sequences marked from dark orange to light orange (sequences #15 and #16). C: Alignment of similar sequences marked from dark pink to light pink (sequences #18, #19 and #20). Sequences #12, #14 and #16 were obtained from 5’_genome walking with the primers GSP1 and GSP2a in ORM#1. Sequences #13 and #15 were obtained from 5’_genome walking with the primers GSP1 and GSP2b in ORM#1. Sequence #18 was obtained from 3’_genome walking with the primers GSP3 and GSP4a in ORM#6. Sequences #19 and #20 were obtained from 3’_genome walking with the primers GSP3 and GSP4b in ORM#6. All these partly similar sequences can be aligned but with indels (insertion or deletion mutations) and nucleotide variations. The horizontal bars indicate the presence of the same sequences and the lines indicate indels. The vertical lines in the bars indicate nucleotide variations among sequences.

To characterise the whole $M^V$-locus, I performed a second and third round of 5’_genome walking based on some of the male specific sequences acquired in the previous round, particularly chose those sequences that are shared with the $M^{III}$-locus (see following part). I obtained five new sequences. Walking out from sequence #15 yielded sequences #15-1 and #15-2 (Fig. 2.12A). Sequences #15-1 and #15-2 are very similar (Fig.2.12A). In the third round of 5’_genome walking, I used primers in the male specific part of sequence #15-2, yielding sequences #15-2-1, #15-2-2 and #15-2-3 (Fig. 2.12B). Sequences #15-2-2 and #15-2-3 are very similar with two nucleotide differences (data not shown). Interestingly, sequences #15-2-1, #15-2-2 and #15-2-3 overlapped with Musca_Mariner_Like_Elements (MLEs), and Musca domestica clone MdAG226 microsatellite sequences.
When walking out from sequence #20 in the second round of 3’-genome walking, I acquired three sequences. All of them include #20, which contains similar genomic sequences that exist in both the male and the female genome (Fig. 2.12C). Genomic sequences “a” in #20, “b” in #20-1, “c” in #20-2 and “d” in #20-3 are similar. Sequences #20-2 and #20-3 are very similar with four nucleotide differences (Fig. 2.12C). These results confirmed that the M-locus consists of repetitive sequences interspersed by genomic sequences that exist in both the male and the female genome.
Figure 2.12: Results of second and third round of genome walking of the $M^y$ strain. The $M^y$-locus contains interspersed genomic sequences that exist in both the male and the female genome. A: Second round of 5’_genome walking started from sequence #15, yielding two new sequences: sequences #15-1 and #15-2. B: Third round of 5’_genome walking started from sequence #15-2, yielding three new sequence sequences: #15-2-1, #15-2-2 and #15-2-3. C: Second round of 3’_genome walking started from sequence #20, yielding three new sequences: sequences #20-1, #20-2 and #20-3. “a”, “b”, “c” and “d” are similar genomic sequences. GSP_Dra52_R2 and GSP_Dra52_R1 are primers for the second round of 5’_genome walking. My_GSP_Pvu7B_R1 and My_GSP_Pvu7B_R2 are primers for the third round of 5’_genome walking. My_GSP_13B_F1 and My_GSP_13B_F2 are primers for the second round of 3’_genome walking. The shaded boxes indicate sequences that exist in both the male and the female genome. The dotted lines are undefined sequences. In sequence alignment, the horizontal bars indicate the presence of the same sequences and the lines indicate indels. The vertical lines in the bars indicate nucleotide variations among sequences.
2.4.4 $M_{III}$-locus and $M^V$-locus share intergenic sequences between $Mdmd$ repeats

Interestingly, in 5’_genome walking of the $M^V$ strain, I obtained sequences that were very similar to some of the sequences in the $M_{III}$-locus (Fig. 2.13A). Alignment of the sequences revealed only few nucleotide differences. In 3’_genome walking, I also found similarities between sequences of the $M_{III}$-locus and the $M^V$-locus (Fig. 2.13B). These results indicate that the $M_{III}$-locus and the $M^V$-locus share some similar sequences.

Figure 2.13: Sequence alignments from genome walking sequences of the $M_{III}$ and $M^V$ strains reveal similarity. A: Sequence #2 was obtained from 5’_genome walking of the $M_{III}$ strain and sequences #15 and #16 were obtained from 5’_genome walking of the $M^V$ strain. B: Sequences #9 and #10 were obtained from 3’_genome walking of the $M_{III}$ strain and sequences #18, #19 and #20 were obtained from 3’_genome walking of the $M^V$ strain. The horizontal bars indicate the presence of the same sequences and the lines indicate indels. The vertical lines in the bars indicate nucleotide variations among sequences.

I composed a cladogram of the sequences from 5’_genome walking in the $M_{III}$-locus and the $M^V$-locus and of sequence #5 from 3’_genome walking in the $M_{III}$-locus by trimming the variable end. It turns out that these sequences belong to six clades, which I labeled A-F. Sequences #2, #15 and #16 belong to clade A, sequences #12, #13 and #14 belong to clade B and sequences #4 and #5 belong to clade C (Fig. 2.14). Sequences #1, #3 and #11 form their own clades. Similarly, I also composed a cladogram of the sequences from 3’_genomic DNA walking in $M_{III}$-locus and the $M^V$-locus and of sequence #4 from 5’_genome walking in the $M_{III}$-locus by trimming the variable end. It turns out that these sequences belong to five clades, which I labeled A-E (N.B. different clades than from the 5’cladogram). Sequences #9, #18, #19 and #20 belong to clade A, sequences #6, #7 and #8 belong to clade C, and sequences #4 and #5 belong to clade D (Fig. 2.15). Sequences #10 and #17 form their own clades. Cladogram confirmed that the $M_{III}$-locus and the $M^V$-locus share some similar sequences.
Figure 2.14: Cladogram of 5’_genome walking sequences in the M_{III}-locus and the M_{V}-locus and of sequence #5 from 3’_genome walking in the M_{III}-locus. They belong to six clades, sequences #2, #15 and #16 to clade A, sequences #12, #13 and #14 to clade B, and sequences #4 and #5 to clade C. The branch labels show the percentage of consensus support. The scale bar indicates the number of substitutions per site. The bottom table shows the percentage of bases/residues that are identical between two sequences. High similarities between two sequences are indicated by white numbers in dark cells and low similarities by dark numbers in light cells.
Figure 2.15: Cladogram of 3’_genome walking sequences in the \textit{M}^{III}-locus and the \textit{M}^{V}-locus and of sequence #4 from 5’_genome walking in the \textit{M}^{III}-locus. They belong to five clades, sequences #9, #18, #19 and #20 to clade A, sequences #6, #7 and #8 to clade C, and sequences #4 and #5 to clade D. The branch labels show the percentage of consensus support. The scale bar indicates the number of substitutions per site. The bottom table shows the percentage of bases/residues that are identical between two sequences. High similarities between two sequences are indicated by white numbers in dark cells and low similarities by dark numbers in light cells.
2.5 Discussion

The objective of this study was to determine the structure of the M-loci in two autosomal M strains, $M^\text{III}$ and $M^\text{V}$, through genome walking. I first performed genome walking to identify genomic regions adjacent to the $Mdmd$ ORMs in the $M^\text{III}$ strain. I found that the $M^\text{III}$-locus consists of multiple copies of sequences, which all show homology to each other. The $M^\text{III}$-locus contains at least six copies of $Mdmd$, which was confirmed by genomic DNA amplification with the divergent primers localised at ORM#1 and ORM#6 in the $M^\text{III}$ strain. The $M^\text{V}$-locus seems to have fewer copies, as there were only two fragments amplified with divergent primers in Fig. 2.9, indicating that it might contain a minimum of three copies of $Mdmd$. However, genome walking on the $M^\text{V}$ strain revealed that it contains at least six copies of $Mdmd$. The different results obtained from genome walking and genomic DNA amplification with divergent primers indicate that various methods are required to determine the structure of the M-loci in other $M. domestica$ strains.

Cladogram analysis further illustrated that sequences in the $M^\text{III}$-locus and the $M^\text{V}$-locus could be divided into different clades, with sequences within clades being more similar than sequences between clades. Interestingly, the $M^\text{III}$-locus and the $M^\text{V}$-locus share some similar sequences. These results are most easily explained by assuming that there have been independent amplification events before and after translocation of the M-locus to autosomes III and V, possibly from the Y-chromosome. In addition, some sequences are always interspersed by identical or similar genomic sequences that exist in both the male and the female genome, indicating that amplification of $Mdmd$ occurred with inclusion of their flanking genomic regions. Also, it is still not known how many repetitive sequences exist in the M-loci from different $M. domestica$ strains. Further characterisation of the M-loci by Pacific Biosciences (PacBio) sequencing that produces long reads will be required to determine the precise structure of the M-loci in different $M. domestica$ strains.

One of the most striking findings regarding the structure of the M-loci is the presence of transposable element sequences that are homologous to $\text{Musca}_\text{Mariner}_\text{Like}_\text{Elements}$ (MLEs). MLEs belong to class II transposons/DNA transposons that are characterised by cut-and-paste transposition. DNA transposons are known to play a role in gene duplication and translocation (Feschotte and Pritham, 2007). M-loci that contain multiple copies of $Mdmd$ flanked by MLEs, may suggest that TEs have played a role in generating $Mdmd$ duplications. In addition, transposons may be involved in translocation of the M-loci. For example, the $Tc1/mariner$ element was found to be capable of
transposing host sequences in *Drosophila melanogaster*, *Lucilia cuprina* and *Bactrocera tryoni* (Coates et al., 1997). An alternative explanation is that they are not functionally involved in the *M*-locus evolution, but merely have landed there after amplification and translocation of *Mdmd*. Besides the presence of transposable element sequences in the *Mdmd* region, I also observed the insertion of microsatellite sequences in the *M*-locus. Microsatellites are simple sequence repeats (SSR), which have high mutation rates (Li et al., 2002). It is currently unknown how the transposable element sequences and microsatellite sequences inserted in the *M*-loci and what their role might have been in *Mdmd* amplification and translocation. Further study is required to determine the causes and effects of the observed association between the *M*-loci and these repetitive sequences in the housefly genome.

The question of how sex chromosomes evolve is currently receiving a lot of attention given that we now have the genomic tools to address this question in a number of systems (Beukeboom and Perrin, 2014). The housefly polymorphic sex determination systems can be uniquely used to study Y-chromosome evolution. Based on my results, I am able to formulate a hypothesis for the *M*-loci evolution in the context of the generally accepted model for Y-chromosome evolution (Fig. 2.16; Charlesworth, 1996; Rice, 1996; Beukeboom and Perrin, 2014). The initial stage of the Y-chromosome evolution is considered to be the acquisition of a sex-determining gene by a standard chromosome. First, *Mdmd* must have evolved as a new male-determining gene and taken up a position at the top of the *M. domestica* sex determination hierarchy. Whether this happened on the ancestral Y or an autosome pair that was not yet involved in sex determination cannot be answered at this moment. The next stage would be the reduction of recombination in the surrounding *Mdmd* region, as predicted by the theory of sex chromosome evolution (Rice, 1996). This would be followed by accumulation of transposable elements and deleterious mutations, including repetitive DNA sequences and transposons due to a lack of recombination on the proto-sex chromosomes (Bachtrog, 2005, 2006, 2013). I indeed found that *M*-loci contain transposable elements and repetitive sequences. Insertions of transposons may play a dynamic and early role in proto-Y chromosome degeneration and may cause functional genes to gradually lose their function (Bachtrog, 2005). Also, accumulation of transposable elements and related repeats can induce heterochromatin (Lippman et al., 2004). In at least one other study of novel sex determination genes in the fish *Oryzias latipes*, it was found that the young Y-chromosome accumulated inactive repetitive elements and transposable element-like sequences in the male-specific region (Nanda et al., 2002; Kondo et al., 2004). The finding of transposable element insertions close to *Mdmd* homologous sequences in my study is consistent with this model of
Y-chromosome degeneration.

Over evolutionary time, lack of recombination and accumulation of deleterious mutations was counteracted by the amplification of *Mdmd* on the Y-chromosome, thus forming the *M*-locus that contains multiple copies of sequences, with various level of homology to each other. In a more advanced phase of Y-chromosome evolution, the *M*-locus may translocate again to an autosome and form a new proto-Y chromosome, starting the whole cycle over again. The finding of multiple copies of *Mdmd* in *MIII* and *MIV* males may reflect this process. The *M*-locus may have translocated multiple times from the Y to an autosome and/or subsequently between autosomes. In addition, my cladogram analysis of sequences obtained from genome walking revealed that to some extent different sequences exist in different autosomes, indicating that after translocation, the *M*-locus underwent further independent amplification on each autosome. The existence of multiple different autosomal *M* variants in the housefly provides a unique opportunity for further study of early stages of sex chromosome evolution.
Characterisation of the complex nature of M-loci in Musca domestica

Figure 2.16: Model for the evolution of M-loci. Mdmd evolved as a new male-determining gene, generating a proto-Y chromosome. Lack of recombination and accumulation of deleterious mutations was counteracted by the amplification of Mdmd on the Y and formed the M-locus. After amplification, the M-locus translocated from the Y to autosomes, either multiple times and/or subsequently between autosomes. After translocation, the M-locus underwent independent amplification on each autosome. The differently coloured boxes indicate repetitive sequences in the M-loci that are partly shared between chromosomes. The shaded boxes indicate sequences that exist in both the male and the female genome.

2.6 Acknowledgements

I acknowledge Akash Sharma for providing the four orphan reads of the male-biased sequences and Claudia Brunner for the gel picture in Fig. 2.9.
2.7 Appendix

2.7.1 Primer sequences

GSP1: 5’-TCTACTGGGTGTTCATTTGAATCCGTTGTG-3’
GSP2b: 5’-CCATACGACTTCCCTTTGCCCTGATAG-3’
GSP2a: 5’-TTCGAGATTCGGCGTGGCR(A/G)TTCAT-3’
GSP3: 5’-GGTW(A/T)GACGCGGACAATACAGGATATT-3’
GSP4b: 5’-AGTGAATTAAAGACGCGGGAAGGC-3’
GSP4a: 5’-R(A/G)GCAGAATCATGAATATCGAATACGCAG-3’
AP1: 5’-GTAATACGACTCACTATAGGGC-3’
AP2: 5’-ACTATAGGCGACGCGGTG-3’
GSP_Dra52_R1: 5’-TCCCTAATTATAGGGTGGCTCAGAACATCG-3’
GSP_Dra52_R2: 5’-CCGCTCTTTAATACCCAAAGTTCTGG-3’
MIII_GSP_Stu93_F1: 5’-CTTCTGTTGTTGGCCCTTCCACCTTTAG-3’
MIII_GSP_Stu93_F2: 5’-GCTGCAATGTCAGATTGTGCATGG-3’
MV_GSP_Dra13B_F1: 5’-AAAGCTGTTCTCTCATCCATACAATTCGTG-3’
MV_GSP_Dra13B_F2: 5’-GCTGCAATGTCAGATTGTGCATGGGTTAC-3’
MIII_GSP_Pvu3B_R1: 5’-AGAAACATTTAACGGCACCGGGACACCTC-3’
MIII_GSP_Pvu3B_R2: 5’-GCTGCAATGTCAGATTGTGCATGG-3’
MIV_GSP_Pvu7B_R1: 5’-TTGGGCTTGACTTGTGTATTTTTTCTGC-3’
MIV_GSP_Pvu7B_R2: 5’-AAACTTGTTGGCCGAAATGGAACGCTGG-3’
UPM for RACE:
Long: 5’-ATTAACCCTCACTAAAGGAAAGCAGTGATCAACGCAGAT-3’
Short: 5’-ATTAACCCTCACTAAAGGAAAGCAGTGATCAACGCAGAT-3’
GSP2b-Dra-52-F: 5’-TGGAATATACGATGTGCTTGAGCCACCTA-3’
GSP3-R: 5’-AATATCTGGTGATTGTTGCCCTGCAACC-3’
GSP1-9-F: 5’-CAACACCTGACGACCAGAAGATGAGT-3’
GSP4b-R: 5’-GCTCTCCGGCGGTCTTTAATTTGACT-3’
M13F: 5’-GTTAACGACGGCCAGTG-3’
M13R: 5’-CAGGAAACAGCTATGAC-3’
1as: 5’-GATTGGCTCAAGTCCGGCTGA-3’
6as: 5’-GGTTGACGGCAGACATACAA-3’