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 3
Comparative analysis of *Mdmd* genes located on different chromosomes of *Musca domestica*

3.1 Abstract

The primary signals on top of the sex determination cascade vary greatly between insect species. How this diversity of primary signals evolved remains unclear. In the housefly, *Musca domestica*, the absence or presence of a male-determining gene(s) is the primary signal for sexual differentiation. A locus that contains the male-determining gene(s) (*M*-locus) 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 differential expression analysis, a male-determining gene was identified and termed *Mdmd* (*Musca domestica* male determiner). The *M*<sup>III</sup>-locus and the *M*<sup>V</sup>-locus have a complex organisation and contain repetitive sequences. A cladogram analysis, based on the *M*-locus sequences, revealed that the *M*<sup>III</sup>-locus and the *M*<sup>V</sup>-locus share some highly similar sequences. Here, I identified an open reading frame (ORF) that is assumed to be the coding sequence of *Mdmd* on the basis of these shared sequences. To further investigate whether this ORF is also present in the *M*-loci located on other chromosomes, I cloned the *Mdmd*<sup>V</sup>cDNA containing this ORF located on autosome V. Cloning of the *Mdmd*<sup>V</sup>cDNA allowed the design of specific primers to amplify *Mdmd* cDNA from other *M*-loci. I found high sequence similarity for *Mdmd* ORF sequences in *M. domestica* strains with the *M*-locus on autosomes II, III and V and on the Y-chromosome but not on autosome I, which apparently has a different male-determining gene(s). This high sequence similarity suggests that all *Mdmd* genes originated from a common ancestral sequence. Comparison of Mdmd protein sequences and its paralog CWC22/NCM suggests a scenario of *M*-locus evolution, whereby the male-determining gene *Mdmd* evolved from a single duplication event of *Md-ncm* generating a proto-Y chromosome, with subsequent translocation to other chromosomes.
3.2 Introduction

Sex determination systems vary strongly among insect species (Sánchez, 2004; Bachtrog et al., 2014; Beukeboom and Perrin, 2014; Blackmon et al., 2017). Three general components can be distinguished in the sex determination pathway in insects: a primary signal, a transductory gene in the middle that memorises the selected fate and a switch gene at the bottom, that together form a cascade of regulatory genes (Bopp et al., 2013). In particular, the bottom gene and to some extent the transductory gene seem to be conserved, whereas the primary signals on top of the cascade vary greatly between different insect species (Bopp et al., 2013). For example, in the housefly, *Musca domestica* (Diptera), the presence of a dominant male-determining gene(s) is the primary signal for male differentiation (Hiroyoshi, 1964). In the honeybee, *Apis mellifera* (Hymenoptera), allelic composition of the *csd* gene is the primary signal for sex determination (Beye et al., 2003). In the silkworm, *Bombyx mori* (Lepidoptera), a single non-coding RNA (piRNA) is the primary signal for female development (Kiuchi et al., 2014). How this diversity of primary signals has evolved still remains unclear. Genetic and molecular study of the various primary signals may aid in understanding the evolution of sex determination diversity among insect species.

In *M. domestica*, a locus that contains the male-determining gene(s) (the *M*-locus) 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). Recently, based upon differential expression analysis, a male-determining gene was identified and termed *Mdmd* (*Musca domestica male determiner*) (Sharma et al., 2017). Knockdown of *Mdmd* by RNAi silencing confirmed that *Mdmd* is necessary for male development and knockout of *Mdmd* by CRISPR-Cas9 resulted in complete feminisation (Sharma et al., 2017). This shows that *Mdmd* plays a crucial role in male development.

Structural analysis of the *M*-loci on autosomes III and V, as described in the previous chapter, demonstrated that the *M*-loci in both strains contain multiple copies of sequences that show various degrees of homology to each other. A cladogram analysis, based on the *M*-locus sequences, revealed that the *M*III-locus and the *M*V-locus share some highly similar sequences. On the basis of these common sequences, I try to identify an open reading frame (ORF) that is assumed to be the coding sequence of *Mdmd*.

An additional question was whether this ORF is also present in the *M*-loci located on the other chromosomes. Cloning the *Mdmd*III cDNA containing a complete ORF
will help to design specific primers to amplify *Mdmd* sequences from other *M*-loci and make functional analysis of *Mdmd* more attainable. In this chapter, I describe the cloning of the *Mdmd* cDNA containing the complete ORF of the *M*- locus on chromosome V using a two-step procedure based on intron-spanning primers. The *M*-locus on chromosome V was chosen because of its potentially lower complexity compared to other *M*-loci (see chapter 2). Next, to gain a better understanding of the origin of *Mdmd*, I compare *Mdmd* nucleic acid and protein sequences derived from the ORFs on autosomes II, III, V and the Y-chromosome. I further compare *Mdmd* protein sequences with NCM from several insect species and CWC22 protein sequences from vertebrate and yeast species, as *Mdmd* and CWC22/ncm are paralogs. The results are discussed in the light of *Mdmd* evolution.

### 3.3 Material and Methods

#### 3.3.1 *Musca domestica* strains

Four different *M. domestica* strains were used in this chapter. (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+ MIII bwb+ w/pw + 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 pointed wings. Males are heterozygous for *M* and 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; MIV 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. (3) 23-1 *M*II strain: *M* is located on autosome II. Females are *X/X; ar/ar; pw bwb/pw bwb*, males are *X/X; MII ar+/+ ar; pw bwb/pw bwb*. ar stands for aristapedia, a recessive mutation on chromosome II. Females from this strain have brown body, pointed wings and leg-like antennae. Males are heterozygous for *M* and they have brown body, pointed wings and normal antennae. (4) 0-3 *M*Y strain: *M* is located on the Y-chromosome. Males are *MIV/X; +/+* and females are *X/X; +/+*. Strains were reared at 25 °C as described previously (Schmidt et al., 1997).

#### 3.3.2 Cloning of *Mdmd* cDNA

Total RNA was purified from 0-8hrs old embryos of the *MIV* strain according to the ZR Tissue & Insect RNA Micro Prep™ Kit from Zymo Research (California, United States). cDNA was 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 µL template RNA (1.8 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 TA Cloning® Kit Dual Promoter with pCR®II vector from Clontech (California, United States) was chosen for the cloning experiment because it allows for direct cloning of the PCR products into the vector at the insertion site flanked by T7 and SP6 promoters, enabling transcription to start from either side into the insert to produce sense or anti-sense products. Advantage 2 polymerase mix was used in the separate amplification of the 5’ and the 3’ parts of MdmdV because it adds a single deoxyadenosine(A) to the 3’ end of the PCR products, which enable the direct ligation of the PCR products to the linearised pCR®II vector with single 3’ deoxythymidine(T) overhang, and it also possesses proofreading ability to minimize PCR induced mutations. In addition, the pCR®II vector contains the lacZα gene that allows for blue-white screening of positive colonies by α-complementation.

The MdmdV cDNA was amplified and ligated into the pCR®II vector in two parts separately. The first 5’ 1702bp containing part was amplified by the primer combinations Mdmd_Intron_as and F1 or F2, respectively, (primer sequences were shown in appendix). The next 1823bp part towards the 3’ end was amplified by the primer combination Mdmd_Intron and R4. F1, F2 and R4 are specific primers for amplifying the ORF of Mdmd. PCR was carried out with the Advantage 2 polymerase Mix of Clontech (California, United States) under the following concentrations and cycling conditions: 0.5 µL MdmdV cDNA, 0.5 µL 10 µM forward primer, 0.5 µL 10 µM reverse primer, 2 µL 2.5 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; followed by denaturation at 94°C for 2 min, then 35 cycles of 94°C denaturation for 30 sec, annealing at 59°C for 30 sec and 72°C for 2:30 min, and finally extension at 72°C for 10 min.

The PCR products were checked on 1% agarose/EtBr gel and the 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 the pCR®II vector from Clontech (California, United States). The constructs were used to transform competent E. coli DH5α. White colonies were selected and cultured in Luria-Bertani (LB) medium that containing 100 µg/mL ampicillin at 37°C overnight. Plasmids from white colonies were extracted and the size of inserted DNA fragments was checked by EcoRI (G|AATTC) from NEB (Massachusetts, United States) digestion. LGC Genomics
(Berlin, Germany) carried out sequencing of the candidate fragments with the primers M13F and M13R located in the vector.

Plasmids from pCR®II+5’ part and pCR®II+3’ part were first digested independently by SmiI (ATTT|AAAT) from Takara with the following concentrations and conditions: 1 µg plasmid DNA, 1 µL SmiI (10 u/µL), 2 µL 10×H Buffer in a total volume of 20 µL; followed by incubating at 30°C overnight. The digestions were checked by agarose electrophoresis and the target fragments were isolated by NucleoSpin® gel and PCR clean-up kit from Macherey-Nagel (Düren, Germany). The target fragment from pCR®II+5’ part with SmiI digestion was digested again by SpeI (A|CTAGT; 10 u/µL) from Takara (California, United States) with the following concentrations and conditions: 1 µg plasmid DNA, 1 µL restriction SpeI, 2 µL 10×M Buffer in a total volume of 20 µL. The target fragment from pCR®II+3’ part with SmiI digestion was digested again by XbaI (T|CTAGA; 15 u/µL) from Takara (California, United States) with the following concentrations and conditions: 1 µg plasmid DNA, 1 µL XbaI, 2 µL 10×M Buffer, 2 µL 0.1% BSA in a total volume of 20 µL. Both digestions were incubated at 37°C for 3 hrs and were checked by agarose electrophoresis and the target fragments were isolated by NucleoSpin® gel and PCR clean-up kit from Macherey-Nagel (Düren, Germany).

After isolating the fragment containing the 5’ part by electrophoresis and gel extraction by NucleoSpin® plasmid kit from Macherey-Nagel (Düren, Germany), it was ligated into the plasmid containing the 3’ part with the following concentration and conditions: 100 ng insert DNA, 100 ng vector DNA, 1 µL T4 ligation buffer in a total volume of 9 µL, 1 µL T4 ligase from NEB (Massachusetts, United States) was added into the reaction to reach the final volum of 10 µL. The ligation was performed at 16°C overnight. The two halves ligated at the exon-exon junction made the full length of MdmdV ORF. After transformation, colony PCR was performed by picking up white colonies. Primers in the 5’ part (46-GSP2b-Dra52-F) and the 3’ part (46-GSP2b-Dra52-R) were used to amply the target fragment. The following concentrations and conditions were used for PCR: 0.2 µL 10 µM 46-GSP2b-Dra52-F, 0.2 µL 10 µM 46-GSP2b-Dra52-R, 0.8 µL 2.5 mM dNTP, 1 µL 10×Advantage 2 PCR Buffer and 0.1 µL Advantage 2 Polymerase Mix (50×) in a total volume of 10 µL. A single white colony was added to each reaction with pipette tip. PCR was performed by denaturation at 94°C for 2 min, followed by 30 cycles of denaturation at 94°C for 30 sec, annealing at 55°C for 30 sec and extension at 72°C for 2 min, and lastly extension at 72°C for 10 min. PCR products were analysed on a 1% agarose/EtBr gel. When the colony revealed positive PCR results, plasmid DNA was extracted with the NucleoSpin® plasmid kit from Macherey-Nagel (Düren, Germany). Plasmids were
checked by EcoRI (G|AATTC) and Sacl (GAGCT|C) from NEB (Massachusetts, United States) independent digestion. LGC Genomics (Berlin, Germany) carried out sequencing of the candidate fragments from positive plasmids with the vector primers M13F and M13R combined with 46-GSP2b-Dra52-F, 46-GSP2b-Dra52-R and cDNA_R_MIII_MV.

3.3.3 Sequence analysis

Sequence alignments of Mdmd from different M. domestica strains were performed with Geneious (Kearse et al., 2012). Protein sequences of Mdmd from different M. domestica strains, NCM from several insect species and CWC22 from vertebrate and yeast species were also aligned. NCM protein sequences from ten insect species (Musca domestica, Stomoxys calcitrans, Bactrocera dorsalis, Glossina morsitans, Ceratitis capitate, Drosophila melanogaster, Aedes aegypti, Bombyx mori, Tribolium castaneum and Nasonia vitripennis), five vertebrates (Homo sapiens, Mus musculus, Gallus gallus, Xenopus laevis, Danio rerio) and two yeast species (Schizosaccharomyces pombe and Saccharomyces cerevisiae) were included. Phylogenetic trees were constructed with the Geneious tree builder based on the Jukes-Cantor genetic distance model and the Neighbor-joining method combined with bootstrap resampling (1000 replicates).

3.4 Results

3.4.1 Identification of Mdmd open reading frame

The cloning of the Mdmd cDNA is difficult owing to the short intron and the existence of many expressed pseudogenes in the M-locus. Using specific primers for highly similar sequences in the MIII-locus and the MV-locus, only intron containing cDNA was amplified. To specifically select for intron-less cDNA, the 5’ part that contains the first exon of Mdmdv ORF and the 3’ part that contains the second exon of the Mdmdv ORF were amplified separately by applying reverse-complementary primers that span the intron (Fig. 3.1).
Figure 3.1: Sequences of primers Mdmd_Intron and Mdmd_Intron_as spanning the intron. The first seven nucleotides of primer Mdmd_Intron anneal to the 3’ end of the first exon. The next nineteen nucleotides of primer Mdmd_Intron anneal to the 5’ end of the second exon. Mdmd_Intron_as is the reverse complement of Mdmd_Intron. Both primers contain the recognition site ATTTAAAT of SmaI marked in red.

To amplify the 5’ part of the \textit{Mdmd\textsuperscript{v}} cDNA, I carried out a PCR with primer combinations F1 and Mdmd_Intron_as that yielded a 1.9kb target fragment (Fig. 3.2, lane 1) and F2 and Mdmd_Intron_as that yielded a 1.8kb target fragment (Fig. 3.2, lane 2). To amplify the 3’ part of the \textit{Mdmd\textsuperscript{v}} cDNA, I used primer combination Mdmd_Intron and R4 that yielded a 1.8kb target fragment (Fig. 3.2, lane 3). Each target fragment was inserted into the pCR\textsuperscript{®} II vector. The fragments from successful insertions were checked by sequencing.

Figure 3.2: PCR amplification of the 5’ and the 3’ part of the \textit{Mdmd\textsuperscript{v}} cDNA. The fragment in lane 1 is amplified by primer combination F1 + Mdmd_Intron_as, in lane 2 by F2 + Mdmd_Intron_as and in lane 3 by Mdmd_Intron + R4. Primers used for PCR amplification were kindly provided by Dr. Daniel Bopp.

Sequencing revealed that the 5’ and the 3’ parts of the \textit{Mdmd\textsuperscript{v}} cDNA were successfully inserted into the pCR\textsuperscript{®} II vector. The \textit{Mdmd\textsuperscript{v}} ORF contains 3525 nucleotides of mRNA coding sequence. The first exon has 1702 nucleotides and the second exon has 1823 nucleotides. A BLAST search of the \textit{Mdmd\textsuperscript{v}} sequence against the published female genome (Scott et al., 2014) showed that \textit{Mdmd} has a
Comparative analysis of Mdmd genes located on different chromosomes of Musca domestica

high sequence similarity with the splicing regulatory gene CWC22/ncm (nucampholin).

3.4.2 Molecular cloning of the MdmdV cDNA

The MdmdV cDNA was reconstructed by cloning the 5’ part and the 3’ part of the MdmdV cDNA into the same pCR®II vector. Plasmid candidates #101 for the 5’ part and #20 for the 3’ part were chosen to reconstruct the MdmdV cDNA as these two plasmid candidates showed that there was no nucleotide mutation generated throughout the cloning process. Geneious revealed a unique Smil (ATTT|AAAT) site in the reverse-complementary primers sequences that span the intron but not in the pCR®II vector (Fig. 3.1). The pCR®II vector contains Spel (A|CTAGT) and XbaI (T|CTAGA) sites on opposite sides of the insertion. Sequencing revealed that #101 and #20 inserted into of pCR®II vector with different orientations: pCR®II+5’ part was oriented from Spel site to XbaI site in #101, whereas pCR®II+3’ part was oriented from XbaI site to Spel site in #20 (Fig. 3.3). Sequence analysis by Geneious revealed that there is one XbaI site in the 5’ part of the MdmdV cDNA.

![Diagram](image)

Figure 3.3: Sequences #101 and #20 inserted into the pCR®II vector with different orientations. #101 is a candidate for the 5’ part of the MdmdV cDNA and #20 for the 3’ part of the MdmdV cDNA.

After the 5’ part and the 3’ part of the MdmdV cDNA were successfully inserted into the pCR®II vector, the 5’ part of the MdmdV cDNA was cut out from pCR®II vector and ligated into pCR®II+3’ part to reconstruct the MdmdV cDNA (Fig. 3.4). Two unique enzymes Smil and Spel were chosen to digest plasmid #101 that yielded a 1.8kb target fragment. Two unique enzymes Smil and XbaI were chosen to digest plasmid #20 that yielded a 5.7kb target fragment. As Spel and XbaI have compatible cohesive ends after digestion, the 5’ part that was digested by Smil and Spel and the 3’ part that was digested by Smil and XbaI can be ligated together.
Figure 3.4: Cloning strategy of the *Mdmd* cDNA. Primers 46-GSP2b-Dra52-F in the 5’ part and 46-GSP2b-Dra52-R in the 3’ part were chosen to perform colony PCR for successful ligation. EcoRI and SacI were chosen to perform the test digestion for selecting positive cloning candidates.

I first digested plasmids #20 and #101 by Smal, yielding target fragments of 5.8kb (Fig. 3.5A). The upper fragments were purified for the following digestions as the lower fragments were supposed to be undigested plasmid DNA because of incomplete digestion of Smal. Afterwards, the linearised plasmid #20 was digested by XbaI, yielding a 5.7kb target fragment, while the linearised plasmid #101 was digested by SpeI, yielding two fragments: one was 4kb and the other one was 1.8kb (Fig. 3.5B). The 5.7kb fragment containing the pCR®II+3’ part digested by the enzymes Smal and XbaI and the 1.8kb fragment containing the 5’ part digested by enzymes Smal and SpeI were purified for the following ligation.
Comparative analysis of Mdmd genes located on different chromosomes of Musca domestica

Figure 3.5: Restriction digestion of #20 and #101. A: Lane 1 is undigested plasmid DNA of #20. Lanes 2-4 are plasmid DNA of #20 digested by Smil. Lane 5 is undigested plasmid DNA of #101. Lanes 6-8 are plasmid DNA of #101 digested by Smil. B: Lane 1 is the SpeI digestion of #20 that was linearised by Smil and lane 2 is the XbaI digestion of #101 that was linearised by Smil. The target fragments marked by arrows were purified for final ligation.

The Mdmd\textsuperscript{V} cDNA containing the complete ORF was successfully reconstructed (Fig. 3.6). Successful ligation was checked using colony PCR, in which fragments of 1.9kb were amplified with primers 46-GSP2b-Dra52-F in the 5' part and 46-GSP2b-Dra52-R in the 3' part (Fig. 3.4). The plasmids from these positive colonies were extracted and the size of inserted DNA fragments was checked by EcoRI and SaeI digestion. There are two EcoRI sites in the Mdmd\textsuperscript{V} cDNA and two EcoRI cutting sites in the pCR\textsuperscript{®}II vector (Fig. 3.4). Mdmd\textsuperscript{V} cDNA containing plasmids should give four fragments of 3.9kb, 2.7kb, 798bp and 107bp upon digestion with EcoRI. There is one Sacl site in the Mdmd\textsuperscript{V}cDNA and one in the pCR\textsuperscript{®}II vector (Fig. 3.4). Mdmd\textsuperscript{V} cDNA containing plasmids should give two fragments of 4.4kb and 3.2kb upon digestion with Sacl. Two colonies were obtained that contained a complete ORF as insert (Fig. 3.6). Sequencing of #136 confirmed that there was no nucleotide mutation generated throughout the cloning process. Hence, Mdmd\textsuperscript{V} cDNA was successfully cloned.
Figure 3.6: Reconstruction of the \textit{Mdmd} cDNA. A: Colony PCR: successful ligation yielded fragments of 1.9kb in the colony PCR and some colonies among #132, #133, #134, #135, #136, #137, #148, #149 and #150 showed the target fragments. B: EcoRI test digestion yielded four fragments (3.9kb, 2.7kb, 798bp and 107bp), confirming candidates #133 and #136 (the 107bp fragment is not visible in the gel picture because it ran out of the gel). C: SacI test digestion yielded two fragments (4.4kb and 3.2kb), confirming candidates #133 and #136.
3.4.3 Comparison of \textit{Mdmd} ORFs from different chromosomes

High sequence similarity for \textit{Mdmd} sequences in \textit{M}II, \textit{M}III and \textit{M}V \textit{M. domestica} strains with autosomal or Y-chromosomal \textit{M}-loci was found, with the exception of \textit{M}I strain, which apparently has a different male-determining gene(s) (Supplementary Fig. 3.1). \textit{Mdmd} ORFs have in total 3525 nucleotides with two parts coding for the conserved domains MIF4G (from alignment position 1045 to alignment position 1590) and MA3 (from alignment position 1924 to alignment position 2244).

Most of the mutations exist in the flanking regions of the conserved domain coding sequences in supplementary Fig. 3.1 (Table 3.1). Four nucleotide variations are found in the 5’ flanking region and eight nucleotide variations in the 3’ flanking region. There are four nucleotide variations in the middle region located between the domain coding sequences. Only one nucleotide variation is found in the MIF4G domain coding sequence and none in the MA3 domain coding sequence.

The protein sequence of Mdmd consists of 1174 amino acids. Alignment of Mdmd protein sequences from \textit{M}II, \textit{M}III, \textit{M}V and \textit{M}Y strains revealed that they are very conserved with only a few amino acid variations (Fig. 3.7).

Table 3.1: Nucleotide variations in \textit{Mdmd} sequences from the \textit{M}Y, \textit{M}II, \textit{M}III, \textit{M}V strains. Non-synonymous mutations are marked in red.

<table>
<thead>
<tr>
<th></th>
<th>\textit{Mdmd}II</th>
<th>\textit{Mdmd}III</th>
<th>\textit{Mdmd}V</th>
<th>\textit{Mdmd}Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>#310</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>#512</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>#527</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>#1020</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>#1184</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>#1596</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>#1712</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>#1765</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>#1853</td>
<td>G</td>
<td>T</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>#2671</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>#2683</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>#2942</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>#3093</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>#3257</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>#3272</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>#3282</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>#3452</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>C</td>
</tr>
</tbody>
</table>
Figure 3.7: Protein sequence alignments for Mdmd from the M\text{II}, M\text{III}, M\text{V} and M\text{Y} strains revealed conservation of Mdmd protein sequences from different \textit{M. domestica} strains (Mdmd protein sequences from M\text{II}, M\text{III} and M\text{V} strain were kindly provided by Dr. Daniel Bopp). The differences are marked with *.

### 3.4.4 Comparison of Mdmd protein sequences and its paralog CWC22/NCM

As \textit{Mdmd} has a high sequence similarity with the splicing regulatory gene \textit{CWC22/ncm} (\textit{nucampholin}), Mdmd\text{II}, Mdmd\text{III}, Mdmd\text{V}, Mdmd\text{Y} and CWC22/NCM protein sequences from ten insect species, five vertebrates and two yeast species were aligned. Sequence similarities were found between the MIF4G domain and MA3 domain, whereas the amino-terminal region and the carboxy-terminal region displayed substantial sequence divergence (Supplementary Fig. 2).

I performed a phylogenetic analysis of the conserved central parts of these protein sequences by trimming the variable N (up to alignment position 472) and C (from alignment position 1038) termini. The phylogenetic tree in Fig. 3.8 shows clustering of the NCM protein sequences into clade that is generally consistent with evolutionary relationships. CWC22 protein sequences of vertebrates and yeasts are distinct from the NCM protein sequences. All the Mdmd protein sequences are similar and belong to one clade, which has the closest phylogenetic relationship with Md-NCM. This suggests that \textit{Mdmd} arose by \textit{Md-ncm} duplication.
Comparative analysis of Mdmd genes located on different chromosomes of Musca domestica

Figure 3.8: Phylogenetic analysis of conserved central parts of Mdmd, NCM and CWC22 protein sequences by trimming the variable N (up to alignment position 472) and C (from alignment position 1038) termini (indicated by black lines) (see Supplementary Fig. 3.2). The scale bar displays the number of substitutions per site. Full species and strain names are with accession numbers in brackets: Musca domestica (Md, ART29446.1, ART29448.1, ART29445.1, ART29447.1, XP_005185085.2), Stomoxys calcitrans (Stc, XP_013100431.1), Ceratitis capitata (Cc, XP_004537789.1), Bactrocera dorsalis (Bd, XP_011197544.1), Glossina morsitans (Gm, GMY010984-RA), Drosophila melanogaster (Dm, NP_609877.2), Aedes aegypti (Aa, XP_001650254.1), Bombyx mori (Bm, XP_012548441.1), Tribolium castaneum (Tc, XP_001811305.1), Nasonia vitripennis (Nv, XP_001601117.3), Homo sapiens (Hs, NP_065994.1), Mus musculus (Mm, XP_006500496.1), Gallus gallus (Gg, NP_001026421.2), Xenopus laevis (Xl, NP_001089418.1), Danio rerio (Dr, NP_001071037.2), Schizosaccharomyces pombe (Sp, NP_596256.4) and Saccharomyces cerevisiae (Sc, AJR85293.1).

3.5 Discussion

Four achieved goals are described in this chapter. The first objective was to find the potential open reading frame for Mdmd. Second, I cloned the Mdmd cDNA sequence from autosome V. Third, I made a comparative analysis of Mdmd genes located on different chromosomes of the M. domestica genome. Finally, I compared Mdmd protein sequences and its paralog NCM/CWC22 from ten insect species, five vertebrates and two yeast species. Comparative analysis of Mdmd
aids in understanding the evolution of the male-determining gene(s) in *M. domestica*.

Cloning of the *Mdmd* cDNA was hindered by the multiple tandemly arranged copies of *Mdmd* pseudogenes with high sequence variation, the 3.5kb length of the *Mdmd* ORF, and the fact that the *Mdmd* cDNAs contain only one small intron of 56-58bp. By applying a two-step cloning procedure, based on intron spanning primers, the *Mdmd* cDNA of the *M* strain was cloned successfully. Separate amplification of the 5’ part and the 3’ part of the *Mdmd* cDNA by intron spanning primers efficiently reduced the chance of gDNA contamination and enabled the amplification of the 5’ part and the 3’ part of the *Mdmd* cDNA. My successful strategy for cloning of the *Mdmd* cDNA can also be applied to clone long cDNA fragments from other complicated genomic regions in *M. domestica* or other species.

Sequences with high similarity to the *Mdmd* ORF in all *M. domestica* strains with autosomal or Y-chromosomal *M*-loci were found, with the exception of autosome I, which apparently has a different male-determining gene(s). This high sequence similarity suggests that all *Mdmd* genes of autosomes II, III, V and the Y-chromosome originated from a common ancestral sequence. The male-determining gene(s) on autosome I might have taken a different path in evolution. The mutation Ag is hypothesised to be a variant of a male-determining gene on autosome I, which is too weak to repress *Mdtra* activity in the soma, but strong enough to suppress maternal *Mdtra* activity in the germ line (Este and Rovati, 1982; Dübendorfer et al., 2003; Hediger et al., 2010). However, the evolutionary relationships between *Ag* and the male-determining gene(s) on autosome I are still unclear as the sequence of *Ag* and the male-determining gene(s) on autosome I have not yet been characterised. Identification of the male-determining gene(s) on autosome I will shed more light on the evolution of various sex-determining genes within *M. domestica*.

Further experiments on the molecular function of *Mdmd* are required. As the *Mdmd* ORF is conserved throughout *M. domestica* strains, I hypothesise that *Mdmd* is supposed to be solely sufficient to perform the male-determining function in *M. domestica* strains. To further investigate the function of *Mdmd* in the *M. domestica* sex determination pathway, I will start a functional analysis by transiently expressing *Mdmd* in female embryos. I will use the cloned *Mdmd* gene to introduce it into early blastoderm embryos by injecting capped polyadenylated RNA after *in vitro* synthesis. Transient expression of *Mdmd* in the early blastoderm embryos will tell whether *Mdmd* is sufficient for male development in *M. domestica*. The results of this study are reported in chapter 4.
Comparative analysis of Mdmd genes located on different chromosomes of Musca domestica

My phylogenetic comparisons provided more insight into the possible evolution of the male-determining gene(s) of *M. domestica*. Comparison of Mdmd protein sequences and its paralog CWC22/NCM revealed that Mdmd protein sequences have a closer phylogenetic relationship with Md-NCM, suggesting that the male-determining gene *Mdmd* evolved from a single duplication event of *Md-ncm* on a proto-Y chromosome. The duplicated copy of *Md-ncm* somehow acquired a male sex determination function and became *Mdmd*. Subsequent amplification events yielded additional *Mdmd* copies, thereby establishing the complex *M*-locus that subsequently translocated to other autosomal locations.

Gene duplication plays an important role in the origin of new genes (Lynch and Katju, 2004). There is some evidence that sex determination genes arose from a hormone-producing gene or immunity-related gene by duplication. For example, the *Amh* (anti-Müllerian hormone) in vertebrates is required for regression of the Müllerian duct during male fetal development (Rey et al., 2003). However, in the teleost fish *Odontesthes hatcheri*, *amh* duplicated and obtained a new role in being necessary for male testicular differentiation (Hattori et al., 2012). In the rainbow trout *Oncorhynchus mykiss*, the gene *sdY* (sexual dimorphic on the *Y*-chromosome) is expressed only in male testis and has been shown to be not only necessary but also sufficient to trigger testicular differentiation (Yano et al., 2012). *sdY* is supposed to be a truncated and divergent copy of *irf9* (interferon regulatory factor 9). In addition, in *Xenopus laevis*, *Dmrt1* (double sex and mab-3 related transcription factor 1) obtained a new function in sex determination that suppresses male development by duplication, translocation and truncation (Yoshimoto et al., 2008). All these examples provide evidence that autosomal genes may adopt a new function in sex determination systems by gene duplication. The origin of *Mdmd* from duplication of *Md-ncm* in the housefly provides another convincing example for this evolutionary path.

### 3.6 Acknowledgements

I acknowledge Dr. Daniel Bopp for providing the sequences of *Mdmd*II, *Mdmd*III and *Mdmd*V. I would like to thank Claudia Brunner for fruitful discussion about *Mdmd*V cloning and provision of primers F1, F2, R4, Mdmd_Intron and Mdmd_Intron_as. I also want to thank Akash Sharma for providing four orphan reads of the male-biased sequences.
3.7 Appendix

3.7.1 Primer sequences

F1: 5’-CACTCGTTTCAGAACCTTTGGT-3’
F2: 5’-CACGTAACACCGCAGTTTATC-3’
R4: 5’-GTGTTTGATAGCAAGAATTAGGAGT-3’
Mdmd_Intron: 5’-ATATTGAACGAATTTAAATTGCACGA-3’
Mdmd_Intron_as: 5’-TCGTCAATTTAAATTGCTTAGATAT-3’
cDNA_R_MIII_MV: 5’-CGGAATTACTACCCGAACCGAAGGAGC-3’
M13F: 5’-GTAAACGACGGCCAGTG-3’
M13R: 5’-CAGGAAACAGCTATGAC-3’
46-GSP2b-Dra52-F-GSP4b-R-F: 5’-CCAGGGACAAGGACAATCGACTAAGACG-3’
46-GSP2b-Dra52-F-GSP4b-R-R: 5’-AAGAACTTGATGAGGACGAGGATGC-3’
Comparative analysis of Mdm genes located on different chromosomes of Musca domestica

Supplementary Figure 3.1: Sequence alignment of Mdm, Mdm, Mdm and Mdm reveals only few nucleotide differences. Mdm ORFs have in total 3525 nucleotides with two parts coding for the conserved domains MIF4G (from alignment position 1045 to alignment position 1590) and MA3 (from alignment position 1924 to alignment position 2244). Sequences of Mdm, Mdm and Mdm were kindly provided by Dr. Daniel Bopp. The nucleotide differences are marked with gaps.
Supplementary Figure 3.2: Sequence conservation of *Mdmd*, NCM and CWC22. Multiple protein sequence alignment of *Mdmd*, NCM and CWC22 from seventeen different insect, vertebrate and yeast species. The variable N (up to alignment position 472) and C (from alignment position 1038) termini (indicated by black lines) from this alignment were trimmed, to use only the conserved central parts of the proteins for phylogenetic analysis.