University of Groningen

Identification and characterization of the male-determining gene of the housefly, Musca domestica

Sharma, Akash

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):
Sharma, A. (2018). *Identification and characterization of the male-determining gene of the housefly, Musca domestica*. University of Groningen.

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).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

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.

Download date: 12-10-2022

Identification and characterization of the male-determining gene of the housefly, *Musca domestica*

Akash Sharma

The research described in this thesis was carried out in the Evolutionary Genetics, Development and Behaviour group at the Center of Ecology and Evolutionary Studies (CEES) - from 2015 onwards known as the Groningen Institute for Evolutionary Life Sciences (GELIFES) - of the University of Groningen, The Netherlands, according to the requirements of the Graduate School of Science (Faculty of Science and Engineering, University of Groningen).

The research was funded by the Netherlands Organisation for Scientific Research (grant no. ALW 822.02.009). The printing of this thesis was supported by the University of Groningen and the Faculty of Science and Engineering.

Cover design and layout: Akash Sharma
Printed by: Glideprint – The Netherlands
ISBN (printed): 978-94-034-0329-8
ISBN (digital): 978-94-034-0328-1

university of groningen

Identification and characterization of the male-determining gene of the housefly, *Musca domestica*

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. E. Sterken
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Friday 5 January 2018 at 12.45 hours

by

Akash Sharma

born on 19 July 1985
in Lalitpur, India

Supervisor
Prof. L.W. Beukeboom

Co-supervisors
Dr. L. van de Zande
Dr. D. Bopp

Assessment committee
Prof. L. Rabinow
Prof. L. Sánchez
Prof. I.R. Pen

Contents

CHAPTER 1	General introduction and thesis overview	7
CHAPTER 2	A candidate male-determining gene in the housefly, *Musca domestica*	21
CHAPTER 3	Different sex determination strains in *M. domestica* evolved by translocations of *Mdmd*, a paralog of the spliceosomal factor *Md-ncm*	39
CHAPTER 4	Functional analysis of *Mdmd* in *M. domestica* by embryonic RNAi	65
Box 1	Functional analysis of *Mdmd* by CRISPR/Cas9 approach	82
CHAPTER 5	Summarizing discussion	95
References		105
English summary		115
Dutch summary		121
Acknowledgements		127

Chapter 1

General introduction and thesis overview

SEX DETERMINATION: GENERAL PERSPECTIVE

Sex determination is a fundamentally vital process in the development of sexually reproducing organisms, but the underlying genetic mechanisms are remarkably diverse (Valenzuela *et al.*, 2003; Beukeboom and Perrin, 2014). There are two types of sex determination distinguished in nature (Bull, 1983), environmental sex determination (ESD) and genotypic sex determination (GSD). Under ESD, sex is determined by external environmental factors, such as temperature, pH and nutrient availability (Bull, 1983). In ESD sex cannot be predicted by the zygotic genotype if any genetic difference exists between the sexes (Bull, 1983; Solari, 1994). Under GSD, sex is genetically determined by instructive genes. In many organismal groups, these genes are contained in sex chromosomes, but there are also groups with GSD without sex chromosomes (e.g., haplodiploids) (Sarre *et al.*, 2004; Beukeboom and Perrin, 2014). Sex chromosomes can contain both sex-related and non-sex-related genes (Mawaribuchi *et al.*, 2012). Besides, many species, are known to possess genes that are only expressed in one sex or are differentially expressed between the sexes (Mawaribuchi *et al.*, 2012). In mammals (including humans), females are the homogametic sex and have two X chromosomes, whereas males are heterogametic and have an X and a Y chromosome, the latter containing a dominant male-determining gene *Sry* (Gubbay *et al.*, 1990). Under female heterogamety (ZW-ZZ system), females are the heterogametic sex (ZW), and males are the homogametic sex (ZZ), which occurs for example in birds, butterflies, and snakes (Bull, 1983). Some species have a polygenic sex determination system, e.g., zebrafish (*Danio rerio*), male or female sex is controlled by a quantitative threshold trait determined by multiple sex-associated regions in the genome (Bradley *et al.*, 2011; Anderson *et al.*, 2012). In some amphipods, sex of the offspring is determined by cytoplasmic factors in combination with their nuclear genotype (Bull, 1983; Werren and Beukeboom, 1998).

Insects exhibit a great variety of genetic systems to determine sex (Sánchez, 2008). Systems range from male heterogamety (XX-XY), female heterogamety (ZW-ZZ) to haplodiploidy (male 1n-female 2n), and some less abundant and more peculiar systems (Beukeboom and Perrin, 2014; Blackmon *et al.*, 2015). Male heterogamety is most abundant, in such species male sex determination is often accomplished by a dominant male-determining factor, located on the Y chromosome (Schmidt, *et al.*, 1997; Pane *et al.*, 2002). In XX/XO sex determination systems, which is found in for example grasshoppers, the females are homogametic (XX), and the males have only one sex chromosome (XO) (Traut *et al.*, 2008). Female heterogamety occurs in a number of groups, including all Lepidoptera. In *Bombyx mori* female sex is determined by a dominant feminizing factor, *Feminizer* (*Fem*), which is present on the W chromosome (Tanaka, 1916). Similar to XX/XO male heterogamety the W chromosome can also be absent, and females carry one Z only (ZO-ZZ). In these sex determination systems, sex is determined by the dose of X (or

Z) chromosome linked-genes. Sex chromosome dose-dependent sex determination can also occur under XX-XY, like in *Drosophila* species (Erickson and Quintero, 2007). Under haplodiploid sex determination, in, e.g., Hymenoptera, males develop from unfertilized eggs and are haploid, whereas females develop from fertilized eggs and, are diploid (Crozier, 1971; Normark, 2003). Sex is often determined by complementary sex determination (CSD) in which individuals heterozygous at one (or more) sex loci develop into females, and homozygous or hemizygous individuals develop into males (Beye *et al.*, 2003). In some insects, the sex of offspring depends entirely on the maternal genotype, i.e., females produce only male or only female offspring, e.g., the dipterans *Sciara coprophila* and *Chrysomya rufifacies* (Bull, 1983). This phenomenon is known as monogeny.

SEX DETERMINATION CASCADES

Under GSD, sex is often determined by an interaction of multiple genes in a cascade of gene regulatory events leading to the permanent establishment of the sexual phenotype (Bull, 1983; Sarre *et al.*, 2004). The regulation of such genes in either the female or the male mode is subject to a primary signal: an initial bias between the embryos of distinct sexes. Sexual differentiation, the development of sex-specific morphology, physiology and behavior directly follows sex determination in the early stage of development.

In insects, the primary instructive signals of the sex determination cascade appear to be highly divergent (Gempe and Beye, 2011; Bopp *et al.*, 2014). Examples of primary signals include a dominant male-determining gene, like in the flies *Musca domestica* and *Ceratitis capitata*, and dose of X-chromosome-linked genes, like in *Drosophila*. Recently, the first two male-determining factors have been identified from insects; *Nix* (Hall *et al.*, 2015) and *Yob* (Krzywinska *et al.*, 2016) in mosquitoes. Both genes are distinctly different from each other. In the lepidopteran *Bombyx mori*, a single non-coding female-specific PIWI-interacting RNA (piRNA) acts as the primary signal for sex determination (Kiuchi *et al.*, 2014).

The primary signal is transduced by the binary genetic switch gene *transformer* (*tra*) which is conserved among higher dipterans (Sánchez, 2008; Verhulst *et al.*, 2010; Bopp, 2010). The bottom-most gene in the cascade is *double-sex* (*dsx*) of which the male- and female-specific products are both functional protein isoforms and ultimately regulate the genes responsible for sexual differentiation on a transcriptional level (Dübendorfer *et al.*, 2002; Saccone *et al.*, 2002; Sánchez, 2008). *Dsx* is a transcription factor that regulates several target genes; this final step of the cascade is conserved, as opposed to the uppermost primary signal (Gempe and Beye, 2011). The regulatory events downstream of *dsx* have also considerably diversified in the course of insect evolution. These pathways are directly

responsible for establishing the enormous variety of dimorphic phenotypes, by shaping morphology, anatomy, physiology, and behavior of both sexes (Carroll *et al*., 2008). Wilkins proposed that sex determination mechanisms are the result of the hierarchical addition of genes to the top of the cascade rather than the loss or modification of genes from an ancient multi-step pathway (Wilkins, 1995). Thus, the genes at the top of the pathway are those that have been added most recently, and the genes at the bottom of the pathway are the most ancient.

SEX DETERMINATION IN *MUSCA DOMESTICA* AND *DROSOPHILA MELANOGASTER*

The sex determination pathways of the dipterans *Musca domestica* and *Drosophila melanogaster* have evolved separately for roughly 100 million years (Beverley and Wilson, 1984) and are consequently quite different. In *M. domestica* the primary signal for sex determination is the presence or absence of one or more male-determining factors (*M*) (Hiroyoshi, 1964; Franco *et al.*, 1982; Denholm *et al.*, 1983) whereas in *Drosophila* it is the dose of X-chromosome-linked gene products that controls sex determination (Erickson and Quintero, 2007). XX, XXY, and XXYY drosophila are females, but XY, and XO flies are males.

A second difference is the presence of the key sex-determining gene *Sex-lethal* (*Sxl*) in the *Drosophila* pathway. During early development, it is active in females and inactive in males (Cline, 1983; Parkhurst and Meneely, 1994) and directly targeted by the X:A signal. The expression of *Sxl* is first transcriptionally controlled by an early promoter (P_E) and later by differential processing of RNA from a late promoter (P_L). The early promoter is activated only in XX embryos, and the late promoter is constitutively active in both XX and XY individuals (Keyes *et al.*, 1992). In XX individuals, a double dose of X chromosome-linked factors activates *Sxl* at the early blastoderm stage producing an early supply of SXL proteins. In XY individuals, *Sxl* remains inactive, and maleness follows (Erickson and Quintero, 2007). In contrast to *Drosophila*, *Sxl* is not a component of the cascade in *M. domestica* (Bopp, 2010). Whereas in *Drosophila* females the female-specific splicing of *Sxl* is maintained by an autoregulatory feedback loop where SXL splices its own pre-mRNA and subsequently *tra*, in *M. domestica* a maternal supply of *Md-tra* (*M. domestica-transformer*) activity engages the autoregulatory feedback loop of *Md-tra* in the zygote. Continuous expression of *Md-tra* products is required to specify and maintain the female fate. The *Md-tra* products are required to direct and maintain female development. If Md-TRA is lost during embryogenesis male development ensues (Hediger *et al.*, 2010).

A third difference is that the autoregulatory *Md-tra* loop is interrupted by a dominant male-determining factor, provided by the male genome, in the housefly. The identity of this male-determining factor was as of yet unknown. The final step of the cascade is conserved in both species: the sex-specific splicing of the transcription factor gene *doublesex* (*dsx*). *Dsx* is a direct target of *tra* and regulates many downstream genes to implement the selected sexual program (Gempe and Beye, 2011) (Fig. 1.1).

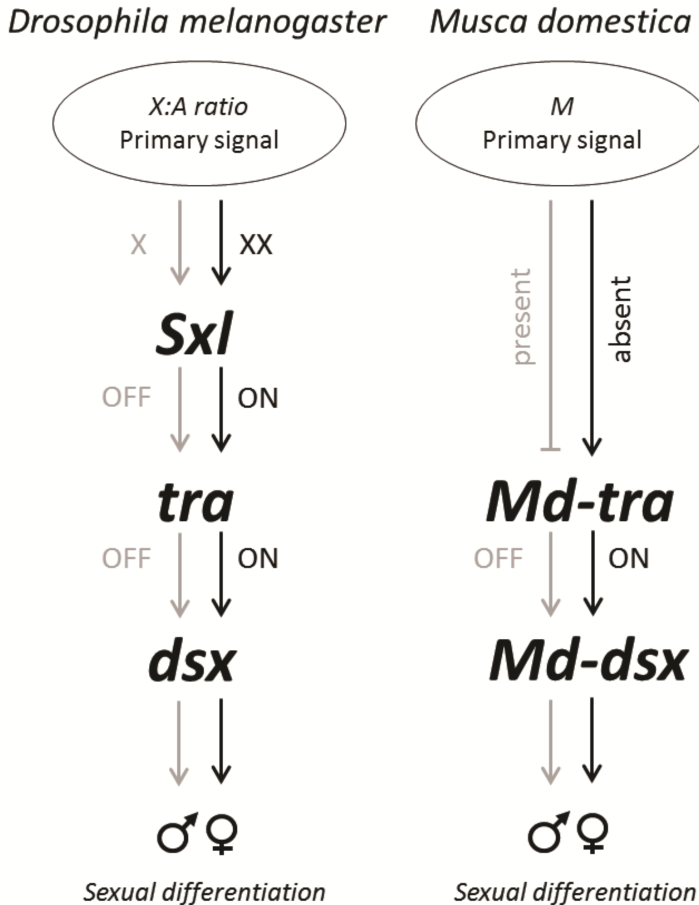

Figure 1.1: Comparison of the sex-determining cascades of *D. melanogaster* and *M. domestica* (Figure modified from Dübendorfer *et al.*, 2002).

THE HOUSEFLY AS A MODEL ORGANISM

The housefly, *Musca domestica* L. (Diptera: Muscidae) is a cosmopolitan species that can transmit many life-threatening diseases in cattle and humans, such as anthrax, typhoid fever, tuberculosis, cholera, and diarrhea (Greenberg, 1965; Fotedar *et al.*, 1992). Its polymorphic and dynamic sex determination system (Dübendorfer *et al.*, 2002) makes the housefly particularly suitable for experimental research on sex determination evolution. It has many advantages as a model organism for biological research. It can be easily cultured in the laboratory on standard media, and it takes around two weeks at 25°C to develop from egg to adult (Schmidt *et al.*, 1997). A number of genetically defined housefly strains are available and used in different laboratories worldwide. Its genome is sequenced, and a first draft has been published (Scott *et al.*, 2009). A linkage map with some visible markers for the five autosomes is available (Hiroyoshi, 1961; Nickel and Wagoner, 1973; Hiroyoshi, 1977). Gene functional analysis tools like germline transformation, embryonic RNAi, and CRISPR/Cas9 mutagenesis have also been developed.

VARIATION IN SEX DETERMINATION MECHANISM IN *M. DOMESTICA*

Different sex-determining mechanisms have been observed in natural housefly populations. Male heterogamety (XX-XY genotype) appears most abundant with the Y chromosome carrying a dominant male determiner *M* (Hiroyoshi, 1964). One deviation from this standard sex determination pathway is the location of the male-determining factor *M*. Besides on the Y it can be found at different genomic sites in natural populations. *M*-factors have been reported from all autosomes (M^{I}, M^{II}, M^{III}, M^{IV} and M^{V}) and even the X chromosome (M^{X}) (Hiroyoshi, 1964; Franco *et al.*, 1982; Denholm *et al.*, 1983; Inoue *et al.*, 1983). It has been a longstanding enigma whether these variants represent different male-determining factors or one and the same factor that somehow translocated between chromosomes.

In addition to variation in the chromosomal position of *M*, the strength of its male-determining activity also varies among *M*-carrying strains. *M*-factors that are located on autosomes II, III and V and Y (M^{II}, M^{III}, M^{V}, and M^{Y}) show effects in the soma as well as in the female germ line. The *M*-factor mapped on autosome I (M^{I}) has weak masculinizing activity resulting in some yolk protein production in the fat body of fertile M^{I}/+ males (Schmidt, *et al.*, 1997; Hediger *et al.*, 1998).

The Y chromosome of *M. domestica* contains two copies of the *M*-factor: one of which has been assigned to the short arm, the other to the long arm of the Y (Hediger *et al.*, 1998). The whole Y chromosome ensures a masculinizing effect, but each region alone

behaves as a hypomorphic *M*-factor, meaning that the activity or expression level of *M* is reduced, which causes many individuals to develop as intersexes instead of males. The *M*-factor located on the euchromatic region of the Y chromosome has stronger male-determining activity than the *M*-factor on the long arm, which consists of constitutive heterochromatin (Hediger *et al.*, 1998).

In *M. domestica*, alternative sex determination instruction signals exist such as dominant autosomal male-determining factors, dominant autosomal female-determining factors, and maternal effect determiners (Hiroyoshi, 1964; Franco *et al.*, 1982; Vanossi Este and Rovati 1982; Denholm *et al.*, 1983; Inoue *et al.*, 1983; Dübendorfer *et al.*, 2002). An allelic variant of the *Md-tra* gene is the dominant female-determining *Md-tra*^{D} allele that is present in some natural populations. *Md-tra*^{D} is a gain-of-function mutation allele of *Md-tra* and insensitive to one or more *M*-factors (Franco *et al.*, 1982; Tomita and Wada, 1989; Hilfiker-Kleiner *et al.*, 1993; Çakır and Kence, 1996; Dübendorfer *et al.*, 2002; Hamm *et al.*, 2005). The *Md-tra*^{D} allele evolved from multiple nucleotide deletions and insertions in the intron sequences of *Md-tra* (Hediger *et al.*, 2010) (Fig. 1.2).

Sex determination variants of the housefly show a peculiar geographic distribution. On several continents, the XX/XY sex-determining system is abundant at higher latitudes whereas autosomal-*M* systems occur closer to the equator (Franco *et al.*, 1982; Denholm *et al.*, 1983; Tomita and Wada, 1989; Çakır and Kence, 1996; Hamm *et al.*, 2005; Kozielska *et al.*, 2008). At lower latitudes, the frequency of the *Md-tra*^{D} allele also increases (Kozielska *et al.*, 2008; Feldmeyer *et al.*, 2008). The cause of this clinal distribution is as of yet unknown, but in a meta-analysis, temperature was found to be a significant factor (Feldmeyer *et al.*, 2008).

Additional variation in sex determination mechanisms has been found in laboratory populations of the housefly. One such strain carries a dominant maternal effect mutation on autosome I named *Arrhenogenic* (*Ag*) (Vanossi Este and Rovati 1982; Dübendorfer *et al.*, 2002). In this strain, the sex of the offspring depends on the genotype of the mother. Heterozygous females (*Ag*/+) carrying this *Ag* allele produce mostly male and intersex offspring because the maternal activity of *Ag* prevents the production of maternal *Md-tra* mRNA in the female germ line; thus females are devoid of maternal *Md-tra*. The resulting males have no *M*-factor and are referred to as no-*M* males (Hilfiker-kleiner *et al.*, 1994; Hediger *et al.*, 2010). It has been proposed that this *Ag* mutation is a derivative of an *M*-factor that is not expressed in somatic tissues but active in the germ line (Hediger *et al.*, 2010). This arrhenogenic mutant is also known as maternal effect sex determiner (Fig. 1.2).

In another laboratory strain, a recessive mutation *masculinizer* (*man*) has been described, which is probably a partial loss of function mutation of *Md-tra* (Schmidt, *et al.*,

1997). Homozygous individuals (*Md-tra*^{man}/*Md-tra*^{man}) produce substantially reduced levels of the *Md-tra* product and develop as males. In heterozygous individuals (*Md-tra*^{man}/+) the wild-type allele of *Md-tra* is sufficient to engage and sustain the feedback loop, and they develop as females (Fig. 1.2).

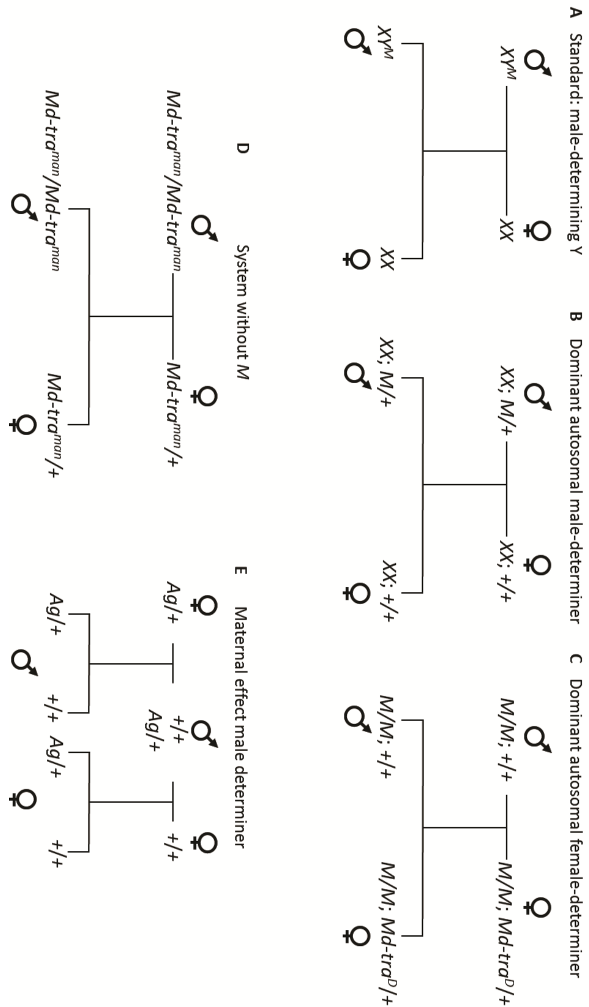

Figure 1.2: Sex determination variants in the housefly. (A) A dominant male-determining factor (*M*) located on the Y chromosome. (B) A dominant male-determining factor (*M*) located on an autosome. (C) A dominant female-determining factor *Md-tra*^{D} is insensitive to *M*. (D) Female heterogamety with the dominant feminizing *Md-tra*$^{+}$ allele from heterozygous individuals (*Md-tra*^{man}/+). (E) Maternal sex determination with the maternal-effect determiner *Arrhenogenic* (*Ag*). Figure modified from Dübendorfer *et al.*, (2002).

MOLECULAR REGULATION OF SEX DETERMINATION IN *M. DOMESTICA*

The sex determination system of *M. domestica* is not only polymorphic at the primary signal level but also at the transducing gene *transformer* (Dübendorfer *et al.*, 2002). A central position in the sex determination pathway is the self-regulatory loop of posttranscriptional regulation of *Md-tra*. *Md-tra* acts as a binary switch that directs female differentiation, when active, whereas male differentiation ensues, when inactive. Maternal *Md-tra* activates zygotic *Md-tra* which, in turn, upholds its activity by a positive feedback loop (Fig. 1.3). The complex of Md-TRA/Md-TRA2 directs splicing of transcripts of the downstream target gene *Md-dsx* into female mode, *Md-dsx*^{F}, leading to female development. The instructive signal for male development is the male-determining factor (*M*-factor). *M* prevents maternal activation of the zygotic *Md-tra* self-regulatory splicing loop; as a result, a non-functional truncated Md-TRA protein is produced and by default *Md-dsx* transcripts are spliced into male mode, *Md-dsx*^{M}, leading to male development (Hediger *et al.*, 2010). Hence, *Md-tra* is the main switch in the manifestation of the sexual fate in *M. domestica* (Dübendorfer *et al.*, 2002; Burghardt *et al.*, 2005). However, as no *M*-factor has been molecularly identified in the housefly, the precise regulation of transformer suppression by *M* remains unknown.

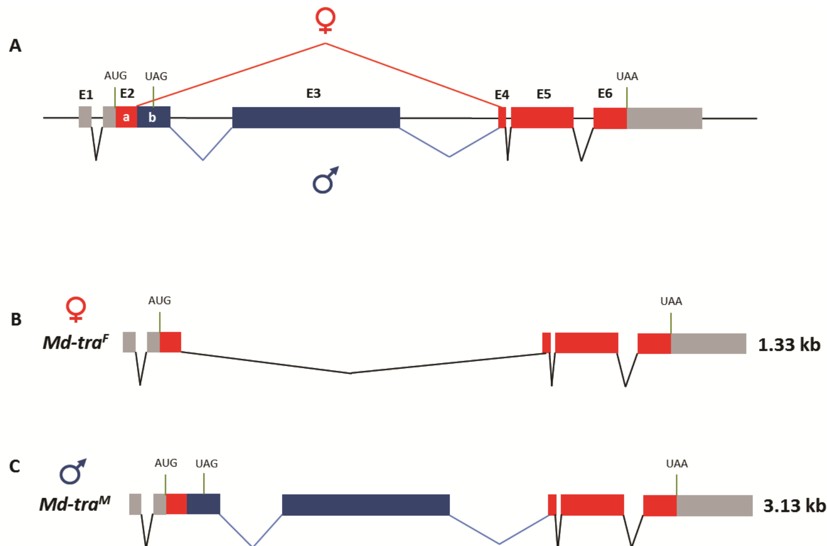

Figure 1.3: Sex-specific splicing of *Md-tra* transcripts in *M. domestica*. (A) Exons in red (E2a, E4, E5, and E6) contain the long ORF. Exons in blue (E2b and E3) are male-specific, and exon E2b contains the premature stop codon. In the female splicing, the male-specific exon (E2b) carrying a stop codon is skipped from the mRNA, and a full-length protein can be translated. In the male splicing, the exon of male-specific *Md-tra* mRNA (E2b) incorporates an in-frame stop codon (UAG) that leads to premature termination of translation. (B) The female-specific splice variant *Md-tra*^{F} has an intact open reading frame that codes for a functional Md-TRA protein (367 aa, 1.33 kb). (C) The male-specific splice variant *Md-tra*^{M} leads to a truncated, non-functional Md-TRA protein (<80 aa, 3.13 kb).

AIM OF THIS RESEARCH

The main objective of my Ph.D. research is to identify and characterize the male-determining (*M*) factor of the housefly, *M. domestica*. Identification of *M*-factors would allow us to answer the question whether *M*-factors that are present at distinct genomic sites are different genes or one and the same gene. At the start of this project, no male-determining gene had been identified in any insect species. This investigation aimed at finding the first male sex-determining gene in the Brachycera sub-order of Diptera. As the evolutionary forces that drive the diversity of sex determination mechanisms are not well understood, I will also address an important evolutionary question, i.e., how *M*-factors can evolve on different locations in the genome. Another aim is to gain insight into the causes of the remarkable diversity in sex chromosomes and sex-determining pathways. A third aim is to gain more knowledge on the molecular regulation of sex determination in the housefly, in particular how *M*-factors may interact with the *transformer* gene.

THESIS OVERVIEW

The experimental approaches and results towards the identification and characterization of the *M*-factor are described in chapters 2 to 4. In the final chapter 5, I summarise the conclusions of these data chapters and discuss future perspectives resulting from my Ph.D. research.

Chapter 2

In this chapter, transcriptomes of early male and female embryos were compared to sort out transcripts that are only present in male embryos. Amongst these, a putative candidate for the male-determining signal in the housefly was identified. We named this candidate gene *Mdmd* (for *Musca domestica male determiner*). *Mdmd* was identified in the M^{III} strain of *M. domestica* in which males carry the *M*-factor on the third chromosome. Sequence analysis of *Mdmd* revealed that it shares a high degree of structural similarity to Complexed with Cef-1/Nucampholin (CWC22/NCM), a well-conserved spliceosome-associated protein which is required for pre-mRNA splicing process and exon junction complex (EJC) assembly.

Chapter 3

In this chapter, I describe that *Mdmd* is a duplication of the *CWC22* ortholog in *M. domestica* which we named *Md-ncm* (*Musca domestica-nucampholin*). To answer the longstanding question whether different *M*-factors that are present on the Y and autosomes are translocations of the same gene, I tested whether *Mdmd* is present in M^{I}, M^{II}, M^{III}, M^{V} and M^{Y} strains. Nucleotide and protein sequence comparison demonstrates that M^{II}, M^{III}, M^{V} and M^{Y} strains contain identical *Mdmd* sequences. *Mdmd* is present on different autosomes and the Y chromosome, confirming that *M* is one and the same gene and has translocated to different genomic sites. The notable exception is strain M^{I} which indicates the existence of another *M* that is different from *Mdmd*. Expression profiling of developmental stages revealed that both *Mdmd* and *Md-ncm* are constitutively expressed genes. Based on molecular evidence (genomic and expression data) we confirmed that *Mdmd* is male-specific and *Md-ncm* is present in both males and females. A phylogenetic analysis including *Mdmd* and *Md-ncm* of the housefly and other metazoans indicates that *Mdmd* rapidly diverged from *Md-ncm* after duplication before translocating to new genomic sites in the *M. domestica* genome.

Chapter 4

In this chapter, I tested whether *Mdmd* acts as the male determiner by specifically disrupting its function either by embryonic RNAi or by CRISPR/Cas9. Embryonic RNAi-based silencing of *Mdmd* leads to differentiation of ovaries in males of the M^{II}, M^{III}, M^{V} and M^{Y} strains, suggesting that *Mdmd* is required for differentiation of male gonads in these strains. Embryonic silencing of *Md-ncm* causes lethality in both male and female embryos, demonstrating that *Md-ncm* is the true ortholog of CWC22 providing its essential functions in general RNA splicing. In Box 1, I describe my attempts to generate loss-of-function alleles of *Mdmd* with the CRISPR/Cas9 method. Though unsuccessful, these attempts were helpful to develop new strategies to deliver Cas9 activity to induce non-homologous end joining (NHEJ) mediated *Mdmd* disruption.

Based on the results of these RNAi functional analyses, I conclude that *Mdmd* has a conserved role in the M^{II}, M^{III}, M^{V} and M^{Y} strains in specifying the male fate of the gonads. No phenotypes were observed in the M^{I} strain which is consistent with our hypothesis that the M^{I} strain uses a different *M*-factor. Alternatively, lack of a phenotype could be due to a high degree of divergence of the *Mdmd* sequence in M^{I} males targeted by dsRNA.

Chapter 5

In this chapter, I discuss the significance of my results for our understanding of housefly sex determination and the evolution of insect sex determination systems. The outcome from this study provides insight into the important evolutionary question how *Mdmd* duplicated and acquired new male-determining function, after rapid divergence from *Md-ncm* and before translocating to new genome sites in the *M. domestica* genome. It yields evidence for the model of co-option of existing genes with different function into the sex determination pathway.

ACKNOWLEDGEMENTS

I am grateful to Leo W. Beukeboom, Daniel Bopp and Louis van de Zande for comments on this chapter.

Chapter 2

A candidate male-determining gene in the housefly, *Musca domestica*

This chapter is published as part of:
Sharma, A., Heinze, S.D., Wu, Y., Kohlbrenner, T., Morilla, I., Brunner, C., Wimmer, E.A., van de Zande, L., Robinson, M.D., Beukeboom, L.W., Bopp, D. (2017). Male sex in houseflies is determined by *Mdmd*, a paralog of the generic splice factor gene *CWC22*. *Science* ***356***, 642–645.

ABSTRACT

The genetic mechanisms of sex determination specifying male or female development are unexpectedly diverse. Sex determination pathways in insects are composed of a cascade of regulatory genes with increasing diversification towards the top of this hierarchy, culminating in a plethora of primary signals. In the housefly, *Musca domestica*, male development requires the presence of a male-determining factor, *M*, that can be located on the Y chromosome or on any of the autosomes. If *M* is expressed, the autoregulatory feedback loop that ensures female-specific expression of the *transformer* gene (*Md-tra*) is disrupted. However, the molecular nature and mode of action of this *M*-factor is not known. Here, we report the identification of a potential *M*-factor candidate, which we call *Mdmd* (for *Musca domestica male determiner*) by comparing the transcriptomes of early male and female embryos. *Mdmd* is a paralog of the spliceosomal gene *CWC22/nucampholin* which is required for pre-mRNA splicing.

INTRODUCTION

Sex determination has been widely investigated in mammals, birds, reptiles, and insects (Manolakou *et al.*, 2006; Bachtrog, 2014). Sex determination mechanisms in insects are diverse and include male heterogamety (XX/XY), female heterogamety (ZW/ZZ) and haplodiploidy (1n/2n) (Beukeboom and Perrin, 2014; Blackmon *et al.*, 2015). Little is still known about the upstream genes in the sex determination cascade that underlie these different chromosomal systems (Schütt and Nöthiger, 2000; Sánchez, 2008; Gempe and Beye, 2011; Herpin and Schartl, 2015).

The insect sex determination pathway consists of a hierarchy of genes that evolves by the addition of genes towards the top of the cascade (Wilkins, 1995). Hence, genes at the top vary more between species than those at the bottom of the cascade (Gempe and Beye, 2011; Bopp *et al.*, 2014). In all insects investigated thus far the *doublesex* gene is present at the bottom of the cascade (Burtis and Baker, 1989; Ohbayashi *et al.*, 2001; Geuverink and Beukeboom, 2014). In several groups, including Diptera and Hymenoptera, the function of the *transformer* gene is to regulate female-specific splicing of *doublesex* (*dsx*) and *fruitless* (*fru*) (Verhulst *et al.*, 2010). In *M. domestica*, male-determining (*M*) factors prevent this *transformer* action, by as yet unknown mechanisms. Recently, two male-determining genes have been identified in mosquitoes in which the *transformer* gene appears to be absent (Salvemini *et al.*, 2013): *Nix* (Hall *et al.*, 2015) and *Yob* (Krzywinska *et al.*, 2016). *Nix* encodes a 288–amino acid polypeptide with two RNA recognition motifs, and *Yob* encodes a short, 56–amino acid protein. These two genes do not show sequence homology to each other, emphasizing the variation in primary signals at the top of the cascade, even between closely related species.

The common housefly (*M. domestica*) is a dipteran species with six pairs of chromosomes. Houseflies have a polymorphic sex determination system, both male (XY) and female (ZW) heterogametic systems are known (Dübendorfer *et al.*, 2002). Mapping crosses have shown that houseflies have multiple male-determining (*M*) factors that vary among natural populations. These *M*-factors can be present at different genomic sites in different natural populations (Hiroyoshi, 1964; Franco *et al.*, 1982; Denholm *et al.*, 1983). Despite the fact that the existence of multiple *M*-factors has been known for over 50 years, no male-determining genes have yet been identified in houseflies.

In some populations of *M. domestica* a gain-of-function allele of *Md-tra,* a dominant female-determining factor $\textit{Md-tra}^{D}$ is present that is insensitive to suppression by *M*. In such populations both females and males are homozygous for *M,* but the $\textit{Md-tra}^{D}$ allele is only present in females which causes female development even in the presence o

several *M*-factors in the same individual (Franco *et al.*, 1982; Tomita and Wada, 1989; Çakır and Kence, 1996; Dübendorfer *et al.*, 2002; Hamm *et al.*, 2005).

A laboratory strain of *M. domestica* carries a maternal effect mutation called *Arrhenogenic* (*Ag*) that is present on chromosome I (Vanossi Este and Rovati, 1982; Dübendorfer *et al.*, 2002). Females carrying this *Ag* allele mutation produce only male offspring because the maternal activity of *Ag* prevents the production of maternal *Md-tra*. The resulting males are referred to as no-*M* males (Hediger *et al.*, 2010). This *Ag* strain resembles a mechanism that is usually observed in insects with maternal sex determination (Dübendorfer *et al.*, 2002).

Sex determination in *M. domestica* is composed of three constituents: an *M*-factor as a primary signal for male development, the regulatory gene *Md-tra*, which is an ortholog of the *transformer* (*tra*) gene of *Drosophila melanogaster*, and the executor gene *doublesex* (*Md-dsx*), which is an ortholog of the *Drosophila dsx*. Only female-specific splicing of *Md-tra* transcripts leads to a functional Md-TRA protein. The maternal transcript of *Md-tra* (*Md-tra^{mat}*) provides the initial source of activity needed to engage the feedback loop. Once the loop is activated, it will sustain the female mode of splicing of *Md-tra* thereby securing a continuous production of active Md-TRA products. Md-TRA/Md-TRA2 and RNA-Binding Protein 1 (RBP1) subsequently splice transcripts of the downstream target gene *Md-dsx* into the female mode, which then instructs female development. We hypothesized that the *M*-factor prevents the maternal activation of the zygotic *Md-tra* self-regulatory loop, resulting in a non-functional form of Md-TRA. As a result, transcripts of *Md-dsx* are spliced in the male mode (Hediger *et al.*, 2010). However, no *M*-factor candidate genes have been identified for *M. domestica* thus far.

The goal of this study was to identify and characterize male-specific sequences as potential candidates for the *M*-factor of the housefly. Our approach was based on the following three premises. First, *M* should be present and expressed only in males. Second, the *M*-factor should be expressed during an early stage of embryogenesis before cellular blastoderm when sexual differentiation occurs. Third, presence and expression of *M* should shift splicing of zygotic *Md-tra* transcripts into male isoforms in the early pre-blastoderm stages (Hediger *et al.*, 2010). Based on these predictions, we conducted a differential gene expression analysis of the syncytial embryonic stage in only-male and only-female progeny.

MATERIALS AND METHODS

M. domestica strains and culturing

The following strains were used (Schmidt *et al.*, 1997; Dübendorfer *et al.*, 2002) (I) *Arrhenogenic*: females carry the *Ag* mutation on chromosome I and produce only sons due to lack of maternal *Md-tra* transcripts (Vanossi Este and Rovati, 1982; Dübendorfer *et al.*, 2002) (II) *Md-tra*^{D}: females carry a dominant gain-of-function *Md-tra*^{D} allele on chromosome IV. Females and males are homozygous for *M* in M^{III} strain (McDonald *et al.*, 1978; Franco *et al.*, 1982; Dübendorfer *et al.*, 2002) (III) M^{III}: males carry the *M*-factor on chromosome III linked to the wild-type alleles of *brown body* (*bwb*$^+$) and *pointed wings* (*pw*$^+$) (M^{III}/+; *bwb*$^+$/*bwb; pw*$^+$/*pw*). Females are homozygous (+/+; *bwb*/*bwb; pw/pw*). All strains were cultured at 25°C in beakers containing 140 g food-mixture (150 g flour, 50 g yeast, 120 g milk powder, 1000 g bran) dissolved in 185 ml water and 4 ml nipagin-solution (66 g nipagin, 123 g nipasol in 2 l of 96% ethanol) until hatching and then transferred to cages at room temperature or kept in beakers at 18°C.

Genomic DNA amplifications

Transcriptome analysis of early male and female embryos yielded RNA contigs. To extend RNA contigs (64077_c2, 64077_c5, 64077_c6.1, 64077_c6.2 and 64077_c3) of *Mdmd* in the M^{III} strain, genomic DNA was extracted from one male and one female of M^{III} strain in 1 ml extraction buffer (0.1 M Tris-HCL, pH 9.0; 0.1 M EDTA, 1% SDS, and 0.5%-1% Sodium N, N-dimethyldithiocarbamate (DMDC added freshly). After incubation at 70°C for 30 min, samples were lysed by extraction buffer, and potassium acetate (140 µl) added. The mixture re-incubated for 30 min on ice, centrifuged, precipitate in ½ volume isopropanol, washed with 70% ethanol and eluted with 100 µL 10 mM Tris+1 µl RNAse A (10 mg/ml stock). Polymerase chain reactions (PCRs) with genomic DNA were performed with different combinations of primers *Mdmd-F1*, *Md-MIII1*, *ORM1s*, *ORM1as*, *ORM3s*, *ORM3as*, *Md-ncm1s*, *Md-ncm2as*, *ORM2s*, *ORM6as*, *ORM6s* and *Mdmd-R4*. PCR was performed in 50 µl volume containing 10 µl 5x GoTaq Reaction Buffer, 1.5 µl 25 mM $MgCl_2$, 1.5 µl 10 mM of each dNTP, 0.3 µl GoTaq DNA Polymerase (Promega), 1.5 µl 10 µM of each primer and 2 µl genomic DNA. PCR thermal cycling consisted of 2 min initial denaturation at 94°C, followed by 35 cycles of 30 s at 92°C, 30 s at 59°C and 30 s at 72°C, and a final elongation of 3 min at 72°C. All PCR products with different combination of primers were purified with PCR Product Purification kit (JETquick, Spin Kit/250) and sanger sequenced by GATC company.

RNA isolation and cDNA amplifications

To check the sex-specific splicing in male and female embryos, total RNA was prepared from 0-6h and 1-8h unisexual embryos with TRIzol reagent (Invitrogen) according to the manufacturer's protocol. Reverse transcription was carried out with the Transcriptor High Fidelity Kit (Roche), using an oligo-dT primer. The cDNA was purified with the JetQuick Spin Column Kit (Brunschwig) and eluted with Tris (10 mM, pH 8) according to the manufacturer's protocol. For cDNA amplifications, primer pairs Mdtra12Bs and Mdtra20as were used for male-specific transcripts of *Md-tra*, and primers Mdtra9s and Mdtra24as for detection of female-specific transcripts of *Md-tra*. PCR was performed in 50 µl volume containing 10 µl 5× GoTaq Reaction Buffer, 1.5 µl 25 mM $MgCl_2$, 1.5 µl 10 mM of each dNTP, 0.3 µl GoTaq DNA Polymerase (Promega), 1.5 µl 10µM of each primer and 2 µl cDNA. PCR thermal cycling consisted of 2 min initial denaturation at 94°C, followed by 35 cycles of 30 s at 92°C, 30 s at 59°C and 30 s at 72°C, and a final elongation of 3 min at 72°C.

RNA-Seq and identification of male-specific transcripts

Libraries for sequencing were generated with the TruSeq RNA Sample Preparation kit (Illumina). cDNA clones were Illumina sequenced with 50 base-pair paired ends (GATC Biotech, Konstanz, Germany). 50bp paired-end reads span a longer region of the transcript, and each read represents one end of a ~200-300 base-pair RNA fragment. To create the transcriptome catalog, all reads from the four samples (female, 1-8h & 0-6h and male, 1-8h & 0-6h) were linked together into a single file and the Trinity tool (Grabherr *et al.*, 2011) used for assembling the male and female reads into 44064 transcripts. Transcript-level quantifications for each sample were done against this catalog with RNA-Seq data and analyzed by Expectation Maximization (RSEM) software (Li and Dewey, 2011) (Table 2.1). For quantification of replicated count data (transcripts), differential gene expression was performed with edgeR (Robinson *et al.*, 2009) by fitting a negative binomial generalized linear model (NB GLM) that accounts for time (0-6h/1-8h) as well as gender (male/female). NB GLM used for modeling count variables. To find male-specific transcripts, the main effect for gender was tested with the likelihood ratio test (McCarthy *et al.*, 2012) from the NB GLM within edgeR. Transcripts from the catalog were then mapped against the *M. domestica aabys* genome (Scott *et al.*, 2014) with GMAP (Wu and Watanabe, 2005) to provide an additional filter.

RESULTS

Generation of unisexual progeny for detection of *M*

Separation of male and female embryos is not feasible in *M. domestica* because there are no visible dimorphic markers in the embryonic stage. Therefore, we exploited the unique housefly polymorphic sex determination system to generate unisexual progenies. The arrhenogenic strain produces no-*M* males because the *Ag* mutation on chromosome I prevent maternal expression of *Md-tra* and, consequently, the maternal provision of *Md-tra* mRNA, required to activate the autoregulatory loop of *Md-tra* in the zygote. To produce female-only progeny, no-*M* males (XX; *Ag*/+ and XX; +/+) from the *Ag* strain were crossed to wild-type females (XX; +/+). To produce male-only progeny, we made use of the presence of the dominant gain-of-function *Md-tra*^{D} allele which overrides *M* repression by an unknown mechanism. In this strain, males carry an *M*-factor on chromosome III, and heterozygous *Md-tra*^{D} carrying individuals will always develop as females. Homozygous *M* males (XX; M^{III}/M^{III}) from the *Md-tra*^{D} strain were crossed to wild-type females (XX; +/+) (Fig. 2.1) and this cross yields only male offspring.

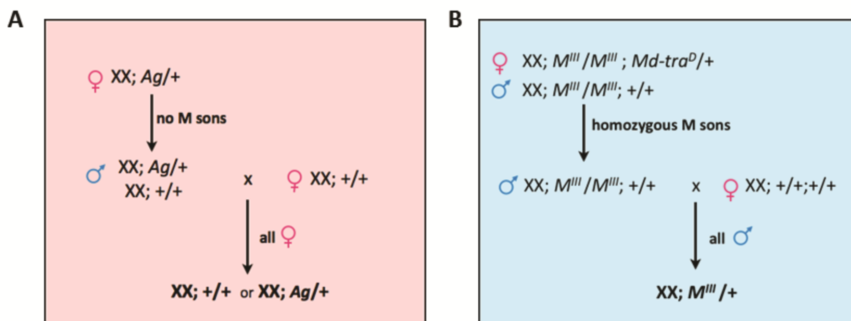

Figure 2.1: Crosses to produce female-only and male-only embryos. (A) Female-only progeny was obtained by crossing wild-type females to no-*M* males that were collected from the *Arrhenogenic* (*Ag*) strain. (B) Male-only progeny was obtained by crossing wild-type females to homozygous M^{III} males from the *Md-tra*^{D} strain.

RNA of all-male and all-female pools of embryos was used for transcriptomic analysis. RT-PCR of the female-specific transformer splice form (*Md-tra*^{F1}) yielded a product in the female RNA pools. *Md-tra*^{F1} also gave a product in male RNA pools because of the maternal provision of *Md-tra* transcripts. Male-specific transformer splice forms (*Md-tra*^{M5} and *Md-tra*^{M1}) were only observed in the male RNA pools which indicated that our crosses had yielded unisexual progenies (Fig. 2.2). Female splice variants (*Mdtra*^{F1}) and male splice variants (*Md-tra*^{M} and *Md-tra*^{M1}) are detected using female and male exon-specific primers.

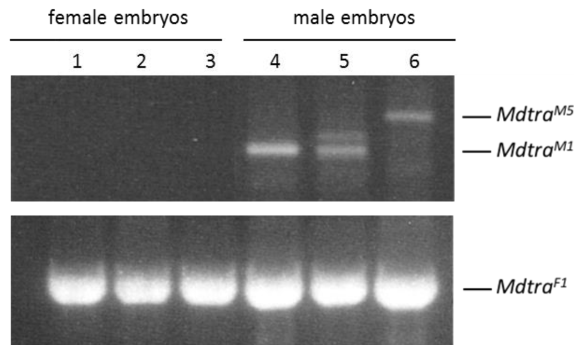

Figure 2.2: Validation of unisexual male and female embryos. RT-PCR amplifications of collected RNA sample pools (only female, lane numbers refer to hours after egg laying 1: 1-8h, 2: 1-5h, 3: 0-6h, only male 4: 1-8h, 5: 1-5h, 6: 0-6h). Male-specific transcripts of *Md-tra, Md-tra*^{M5} and *Md-tra*^{M1,} are only detected in male samples and absent in female samples. Female-specific transcripts of *Md-tra* and *Md-tra*^{F1} are present in all samples because of the maternal provision.

Transcriptome of early male and female embryos

For molecular identification of the *M*-factor, a differential expression analysis based on RNAseq of early stage male and female embryos was conducted. Illumina sequencing yielded 80 million female and 120 million male reads in the 1-8h samples and 140 million female and 120 million male reads in the 0-6h samples. The male and female sequences were assembled with Trinity into a total of 44,064 transcripts and mapped back against the female genome scaffold of *M. domestica* (Scott *et al.*, 2014). The analysis with edgeR revealed that more than 85% of male and female reads matched to the female genome scaffold. They had a similar level of expression in males and females, but we also retrieved more than 11,000 male-biased genes with significantly higher expression in males compared to females (Figure 2.3).

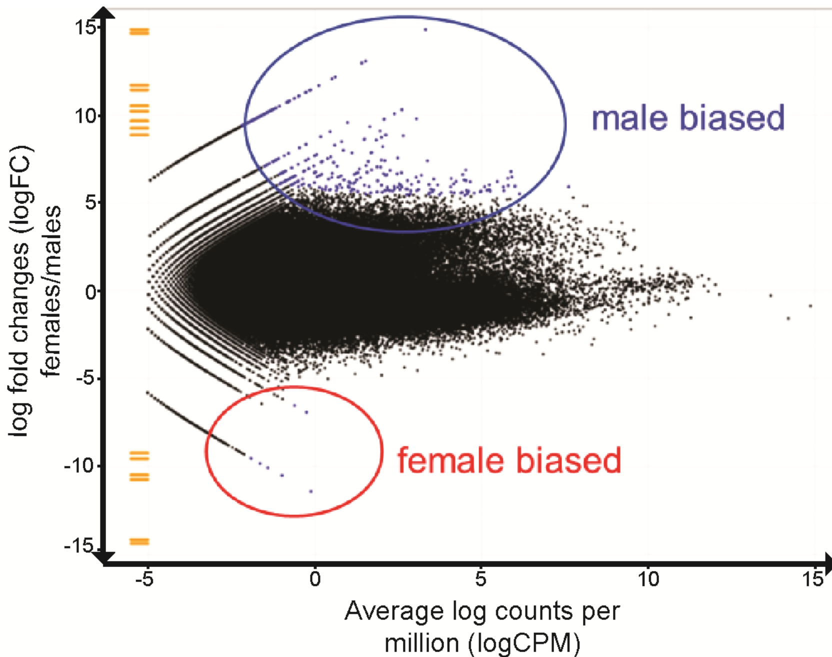

Figure 2.3: Transcriptome analysis by RNAseq. Most of the reads have a similar level of expression in males and females (black dots), but there are also male (blue dots, higher than 5 logFC value) and female-biased reads (red dots, lower than -5 logFC value). The positive and negative logFC (log fold changes) values indicate male and female-biased expression respectively. Average logCPM (counts per million) represent an increased level of male and female bias in gene expression.

Orphan reads in males (*ORMs*)

All reads from the four samples (0-6h/1-8h, male/female) were assembled into 14,392 contigs and transcripts quantified with RSEM software. Of these, we found 39 contigs (Table 2.1) that had a higher expression level in males (positive fold change), a false discovery rate (FDR) of less than 5%, and did not map to any sequence of the female *M. domestica* genome draft (Scott *et al.*, 2009). All males that were generated by the all-male cross (see above) have two X-chromosomes, which are represented in the genome draft. However, these males are also homozygous for M^{III}, which is not represented in the (female) genome draft. Therefore, these 39 contigs most likely represent transcripts from the genomic region linked to M^{III}. We refer to these sequences as "orphan reads in males" (ORMs) (Table 2.1).

A candidate male-determining gene, Mdmd

We did a BLASTN search of these 39 male-biased orphan reads in the NCBI genomic databank of all available organisms. Among the top 14 male-specifically expressed sequences (Table 2.1), we identified five orphan contigs of the same transcript that showed a high level of sequence similarity to Complexed with Cef-1/Nucampholin (CWC22/NCM), a spliceosome-associated protein that is required for pre-mRNA splicing and crucial for exon-junction complex (EJC) assembly (Alexandrov *et al.*, 2012; Steckelberg *et al.*, 2015). Having five copies of the same transcriptional unit among the top 14 genes that were only expressed in males, makes *Mdmd* a promising candidate for the male-determining gene. We named this *M* candidate gene, *Mdmd* (for *Musca domestica male determiner*) (Figure 2.4). Subsequent analysis by PCR extension with specific primers (see legend to Figure 2.4) showed that *Mdmd* codes for a protein that contains two domains, MIF4G (middle domain of eukaryotic initiation factor 4G, eIF4G) and MA3 (initiation factor eIF-4 gamma) (Marchler-Bauer and Bryant, 2004). *Mdmd* was identified in the M^{III} strain in which males carry the *M*-factor on the third chromosome. MDMD contains 1190 amino acid residues.

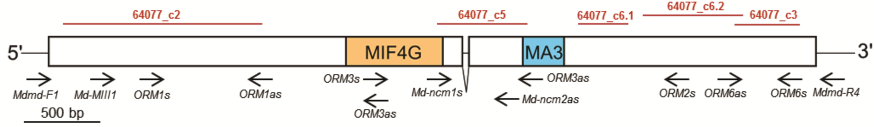

Figure 2.4: *Mdmd* is a male-specific sequence of *M. domestica*. *Mdmd* was identified by BLASTN search of male-specific RNA contigs (64077_c2, 64077_c5, 64077_c6.1, 64077_c6.2 and 64077_c3) into NCBI genomic database. These five RNA contigs were extended by genomic DNA PCR using different combinations of the primers *Mdmd-F1*, *Md-MIII1*, *ORM1s*, *ORM1as*, *ORM3s*, *ORM3as*, *Md-ncm1s*, *Md-ncm2as*, *ORM2s*, *ORM6as*, *ORM6s*, and *Mdmd-R4*. The RNA contigs were confirmed to be part of the same transcriptional unit. MDMD contains a highly conserved domain, MIF4G (in yellow) and MA3 (in blue) with an intron.

DISCUSSION

In this chapter, a potential male-determining gene *Mdmd* of *M. domestica* is described. It was identified based on the transcriptome analysis of early only-male and only-female embryos. *Mdmd* shows high similarity to the autosomal gene *CWC22/nucampholin*. In *D. melanogaster nucampholin* causes embryonic lethality when it is non-functional (Coelho *et al.*, 2005). CWC22 is an essential splicing factor that is required for exon junction complex (EJC) assembly and functionally links with post-transcriptional mRNA modifications (Barbosa *et al.*, 2012). EJC is a key regulator of mRNA localization, translation, and stability (Barbosa *et al.*, 2012) and highly conserved from yeast to human (Shiimori *et al.*, 2013).

In *M. domestica* sex determination is predominantly regulated at the splicing level (Hediger *et al.*, 2010). Males carry an *M*-factor that inhibits the activity of zygotic *Md-tra* yielding a non-functional form of Md-TRA (Hediger *et al.*, 2010). The *M*-factor is expressed in a very early stage of embryonic development, before the formation of cellular blastoderm and directs alternative splicing of zygotic *Md-tra* into male development. When *M* function is absent, *Mdtra* pre-mRNA is spliced into the functional mode retaining an intact ORF, and the embryo undergoes female development. These features suggest that *M* encodes a splice factor which is directly involved in regulating sex-specific splicing of *Md-tra* (Hediger *et al.*, 2010). The production of $Md\text{-}tra^{M}$ is a post-transcriptional event as a result of the alternative splicing process. CWC22 has a role in nonsense-mediated decay which is involved in detection and decay of mRNA transcripts that contain premature termination codons (PTCs) (Alexandrov *et al.*, 2012; Steckelberg *et al.*, 2015). *Mdmd* shows high similarity to CWC22 which has a post-transcriptional regulatory function. This crucial similarity makes *Mdmd* an excellent candidate for the male-determining gene in *M. domestica*.

Mdmd is the first *M*-factor identified from the Brachycera subclass of the Diptera. Recently, two *M*-factors were identified in mosquitoes as the primary signal for male development, *Yob* in *Anopheles gambiae* (Hall *et al.*, 2015) and *Nix* in *Ades aegypti* (Krzywinska *et al.*, 2016). They directly or indirectly induce male-specific splicing of *doublesex* (*dsx*). *Nix* is proposed to be a distant homolog of *transformer-2*, but *Yob* is an unusually short (56 aa) protein that does not show any homology to other genes. Knock-out of *Nix* with CRISPR-Cas9 resulted in partially feminized genetic males and yielded both male and female splice variants of *dsx*. Embryonic silencing of *Yob* has lethal effects in males due to misregulation of dosage compensation and inadequate transcription from the X chromosome, whereas misexpression of *Yob* in larvae indicates that it is sufficient to induce male splicing of *dsx*. These results suggest that both *Nix* and *Yob* have male-determining function. Both *Nix* and *Yob* do not show any homology to *Mdmd*. This emphasizes that primary signals at the top of insect sex-determination cascades are very diverse and not even conserved between closely related species, consistent with the 'bottom-up hypothesis' of Wilkins (Wilkins, 1995). It makes identification of sex determination genes in insects difficult by comparative analysis and detection of commonalities in primary signals a daunting task. Further identification of *M*-factors in dipterans will broaden our knowledge of how novel sex determination pathways and regulatory principles evolve.

ACKNOWLEDGEMENTS

This work was supported by the Netherlands Organisation for Scientific Research (grant no. ALW 822.02.009) and the Swiss National Science Foundation (grant 31003A_143883).

APPENDIX 2.1

Table 2.1: List of contigs of orphan reads in males. Ranked by the highest logFC (log fold changes) and lowest PValue (calculated probability), five orphan contigs (grey shading) of the same transcriptional unit were found among the top 14 RNA contigs. The following parameters are used to detect male-biased orphan reads: contig ID - male-biased orphan reads, female (1-8h, 0-6h) and male (1-8h, 0-6h) - replicated count data (transcripts), logFC (log fold changes) - difference in gene expression between samples, logCPM (counts per million of transcripts) - log2 counts per million normalized for library sizes, LR (likelihood ratio) and FDR (false discovery rate) for statistical significance and multiple comparisons for gene expression data.

Contig id	female 1-8h	male 1-8h	female 0-6h	male 0-6h	logFC	logCPM	LR	PValue	FDR
comp64077_c2_seq2	0	3.89	0	4.83	12.70694547	1.127387117	48.5834239	3.17E-12	4.56E-08
comp64077_c5_seq1	0	2.05	0	3.27	12.03311927	0.419659566	43.8389407	3.57E-11	2.57E-07
comp64077_c6_seq2	0	1.57	0	2.37	11.58931956	-0.014398535	41.4789742	1.19E-10	4.61E-07
comp64077_c3_seq1	0	0.9	0	1.6	10.95821666	-0.660700448	37.78868168	7.88E-10	2.27E-06
comp66948_c0_seq5	23.62	388.89	0.01	0.84	4.729860471	6.690970297	33.20015019	8.31E-09	1.71E-05
comp65705_c8_seq4	0.03	2.91	0.01	3.31	7.63560747	0.647002095	31.51819349	1.98E-08	2.84E-05
comp66112_c0_seq1	0	0.28	0	0.52	9.329689612	-2.259155615	29.59941192	5.31E-08	5.46E-05
comp66112_c1_seq1	0	0.16	0	0.85	9.739113106	-1.922779568	29.36842218	5.98E-08	5.74E-05
comp66112_c6_seq1	0	0.09	0	1.05	9.953201423	-1.732634656	27.98924966	1.22E-07	0.000103276
comp58560_c0_seq1	0.07	5.79	0	0.87	7.366389398	0.738006618	27.62889895	1.47E-07	0.000117508
comp56711_c1_seq1	0.03	3.04	0.05	3.82	6.453091959	0.799011746	27.21782292	1.82E-07	0.000127121
comp66112_c7_seq1	0	0.09	0	0.68	9.367798786	-2.283778992	26.25648374	2.99E-07	0.000175063
comp66112_c4_seq1	0	0.08	0	0.7	9.406084378	-2.249612933	26.10237573	3.24E-07	0.000179228

comp64077_c6_seq1	0	1.06	0.03	1.97	6.974017472	-0.372546802	25.33982814	4.81E-07	0.000230598
comp66112_c3_seq1	0	0.07	0	0.55	9.081017758	-2.552625368	24.60558664	7.03E-07	0.000326594
comp517191_c0_seq1	0	4	0	0	10.82152232	-0.026669173	22.94230608	1.67E-06	0.000686445
comp1087529_c0_seq1	0	0.07	0	0.31	8.335043956	-3.229131818	22.02499356	2.69E-06	0.001075894
comp44611_c0_seq1	0.05	23.47	0	0	8.495645732	2.550672689	21.37399994	3.78E-06	0.001414534
comp49039_c0_seq1	0	0.04	0	0.36	8.468150916	-3.120524882	21.34520583	3.84E-06	0.001414534
comp66112_c2_seq1	0	0.06	0	0.36	8.500193568	-3.08119601	20.90047449	4.84E-06	0.001619183
comp174825_c0_seq1	0.04	13.83	0	0	8.134000154	1.785296687	20.20511504	6.96E-06	0.00213022
comp64776_c0_seq1	0.03	7.61	0	0.01	7.829555161	0.918843148	20.07342357	7.45E-06	0.002234512
comp42865_c0_seq1	0	1.76	0	0	9.645406858	-1.238620477	19.96253838	7.90E-06	0.002241335
comp51856_c1_seq1	0	1.26	0	0	9.183623987	-1.735757111	19.95166507	7.94E-06	0.002241335
comp64150_c2_seq1	0.01	0.29	0.04	4.13	5.640569171	0.185097932	18.93394906	1.35E-05	0.003337824
comp63084_c1_seq2	0	1.12	0	0	8.984979441	-1.900883429	17.06280389	3.62E-05	0.006848266
comp23538_c0_seq1	0	0.01	0	0.18	7.415057104	-4.042122375	15.56594341	7.97E-05	0.011821751
comp61718_c0_seq1	5.16	139.29	0.01	0.03	3.945898406	5.173921113	15.22636345	9.54E-05	0.013724614
comp52867_c0_seq1	0.01	1.14	0	0.01	6.08453022	-1.849045938	15.11957543	0.000100911	0.014207114
comp3145_c0_seq1	0	0.01	0.01	0.54	5.952886899	-2.671278371	14.27317797	0.000158102	0.019447871

comp57279_c0_seq1	0.01	0.49	0	0.03	5.187918342	-2.970062849	14.19960475	0.000164405	0.019883339
comp59360_c0_seq1	2.83	52.39	0	0.02	4.298997469	3.785436854	13.57315248	0.000229444	0.024828224
comp39474_c0_seq1	0.03	1.33	0	0.01	5.425047803	-1.608560396	13.2758852	0.000268842	0.027636949
comp52318_c1_seq3	18.68	451.95	2.03	6.65	3.156510898	6.904581915	12.89108903	0.00033015	0.032996685
comp57141_c4_seq1	0.03	0.61	0.01	0.09	4.043548082	-2.491355462	12.76661691	0.00035286	0.034560967
comp53439_c0_seq1	0.03	0.53	0.04	0.36	3.729868824	-2.045511985	12.5128477	0.000404163	0.036193155
comp43327_c0_seq1	0.03	1.03	0	0.01	5.094334631	-1.966933993	12.49794762	0.000407399	0.036193155
comp63560_c12_seq1	0	0.09	0.03	0.4	4.264731554	-2.809482084	11.93266763	0.000551581	0.046423154
comp58952_c1_seq1	0.02	1.37	0	0	5.747997698	-1.58071779	11.917512	0.000556087	0.046530257

APPENDIX 2.2

Primer sequences

PRIMER NAME	PRIMER SEQUENCE (5'-3')	PURPOSE
Mdmd_F1	CACTCGTTTCAGAACTTTGGGT	*Mdmd* specific
Mdmd_R4	GTGTTTGATAGCAAGAATTAGGAGT	*Mdmd* specific
Md-MIII1	GTAGTACGTGATCTATCTTATACT	*Mdmd* specific
ORM1s	ATCAGGGCAAAGGGAAGTCG	*Mdmd* specific
ORM1as	GATTGGCTCAGATCGGCGTA	*Mdmd* specific
ORM2s	AAGAATCGTCGTCGGATGGT	*Mdmd* specific
ORM3s	CTTTGTTCAGCGCAGCAATC	*Mdmd* specific
ORM3as	AAATGCCTCCAACCCTATCCG	*Mdmd* specific
ORM6s	GCTCTTCCCGGCGTCTTTTA	*Mdmd* specific
ORM6as	GGTTGACGCGGACAATCAAC	*Mdmd* specific
Md-ncm_1s	CGCAGAGATGGCTTTAAGGA	cDNA amplifications
Md-ncm_2as	TTTTTGGGCACATTCCTCAT	cDNA amplifications

Chapter 3

Different sex determination strains in *M. domestica* evolved by translocations of *Mdmd*, a paralog of the spliceosomal factor *Md-ncm*

This chapter is published as part of:
Sharma, A., Heinze, S.D., Wu, Y., Kohlbrenner, T., Morilla, I., Brunner, C., Wimmer, E.A., van de Zande, L., Robinson, M.D., Beukeboom, L.W., Bopp, D. (2017). Male sex in houseflies is determined by *Mdmd*, a paralog of the generic splice factor gene *CWC22*. *Science* ***356***, 642–645.

ABSTRACT

Sex determination is a fundamental process involving a short cascade of regulatory genes that coordinate the implementation of sexual fate in different tissues of the body. Although the genetic pathways of sex determination in insects are diverse, they share a general principle. Different primary signals regulate the conserved transcription factor *doublesex* at the bottom of the cascade. In the housefly, male-determining factors (*M*) can be located on the Y chromosome, but also on other chromosomes. It has been a longstanding question whether these represent different genes or one and the same gene that translocated within the genome. In the previous chapter, we identified the putative primary signal in the housefly as *Mdmd*. Here, we report that BLASTN searches of *Mdmd* recovered a sequence of high similarity in the *M. domestica* female genome. This gene is an ortholog of the core spliceosomal factor *CWC22*/*nucampholin* and therefore named *Md-nucampholin* (*Md-ncm*). From our expression data, we postulate that *Mdmd* and *Md-ncm* are constitutively active during all developmental stages. Presence of *Mdmd* in different *M* strains suggests that these strains emerged from translocations of the same male determinant to different genomic sites.

INTRODUCTION

Insect sex determination involves a cascade of genes that evolve from the bottom up (Wilkins, 1995). The *doublesex* gene is located at the bottom of the cascade and most conserved, whereas at the top, a variety of genes can function as primary signals (Marin and Baker, 1998; Verhulst *et al.*, 2010; Gempe and Beye, 2011). In *Drosophila melanogaster* the gene *Sex-lethal* (*Sxl*) directly responds to the primary signal (X:A ratio) and acts as the main switch in the pathway (Erickson and Quintero, 2007). Although this gene is present in all insect species studied so far, it only has this key sex determination role in Drosophilids, illustrating the increasing diversity towards the top of the cascade (Traut *et al.*, 2006; Sánchez, 2008). Primary signals in other insects are often dominant male-determining genes, such as in the housefly *Musca domestica* (Franco *et al.*, 1982; Denholm *et al.*, 1983), the phorid fly *Megaselia* scalaris (Traut and Willhoeft, 1990), and the fruit fly *Ceratitis capitata* (Saccone *et al.*, 2002). Until recently, no male-determining factors had been molecularly characterized from any insect.

The housefly is an excellent model organism to study the evolutionary diversification of the sex determination pathway because of the presence of different sex determination systems within this species (Dübendorfer *et al.*, 2002; Hediger *et al.*, 2010). Male-determining factors have not only been localized on the Y chromosome but also on all other chromosomes (including the X chromosome) in natural populations (Hiroyoshi, 1964; Franco *et al.*, 1982; Denholm *et al.*, 1983; Inoue *et al.*, 1983). The latter are referred to as autosomal *M* strains. It has been suggested that the various autosomal *M*-factors are translocated copies of an *M*-factor which originally resided on the Y-chromosome (Hiroyoshi, 1964; Schmidt, *et al.*, 1997; Hediger *et al.*, 1998). Alternatively, they may represent different genes that have male-determining function. Identification of housefly male determiners will help to solve this longstanding question.

In *M. domestica*, the sexual fate is irreversibly set by ON/OFF regulation of the binary switch gene *Md-tra* (*Musca domestica-transformer*) during early embryonic development. Maternal *Md-tra* activates zygotic *Md-tra* and engages a self-propagating loop which results in the continuous production of the female splice variant of *Md-tra*, *Md-tra^F*, which encodes the functional TRA protein that directs female development. When a dominant male determiner (*M*) is present in the zygote, either located on the Y-chromosome or on any of the five autosomes, it prevents the activation of zygotic *Md-tra*. As a result a non-functional male transcript, *Md-tra^M*, is produced that has a truncated ORF, leading to a premature termination of protein elongation through the presence of a stop codon. In the absence of functional TRA protein male development follows (Hediger *et al.*, 2010).

In the previous chapter, a putative male determiner from the housefly was identified from the M^{III} strain (*M* on autosome III), which we called *Mdmd* (for *Musca domestica male determiner*). In this chapter, we perform BLASTN searches of *Mdmd* in the *M. domestica* genome (Scott *et al.*, 2014) to check whether *Mdmd* is indeed male-specific or present in both males and females. Another question addressed in this chapter is whether *Mdmd* is also present on the Y and other *M*-carrying autosomes. We try to answer this question by amplifying *Mdmd* from males of different *M*-carrying strains. We present molecular evidence (genomic and expression data) that *Mdmd* is male-specific and *Md-ncm* is present in both males and females. In addition, we perform a comparative analysis of *Mdmd* nucleotide and protein sequences to investigate its origin and phylogenetic relationship to *ncm* of different metazoans.

MATERIALS AND METHODS

Musca domestica strains

The following strains were used (Schmidt *et al.*, 1997; Dübendorfer *et al.*, 2002): (1) XY: wild-type strain from Siat, Switzerland. Males are XY and carry the *M*-factor on the Y; females are XX; (2) M^{II}: males carry the *M*-factor on chromosome II linked to the wild-type allele of *aristopedia* (ar^+); M^{II}, ar^+/ar. Females are homozygous *ar/ar*; (3) M^{III}: males carry the *M*-factor on chromosome III linked to the wild-type alleles of *brown body* (bwb^+) and *pointed wings* (pw^+); pw^+, M^{III}, bwb^+/pw, +, *bwb*. Females are homozygous *pw*, +, *bwb*/*pw*, +, *bwb*; (4) M^V: males carry the *M*-factor on chromosome V linked to the wild-type allele of *ocra* ($ocra^+$); M^V, $ocra^+/ocra$. (5) $Md\text{-}tra^D$: females carry the dominant gain-of-function $Md\text{-}tra^D$ allele on chromosome IV. Females and males are homozygous for M^{III}. All strains were cultured as described in chapter 2.

Genomic DNA and cDNA amplifications

To examine the presence of *Mdmd* and *Md-ncm* in different strains (M^{II}, M^{III}, M^V and M^Y), genomic DNA was extracted from one male and one female of each strain in 1 ml extraction buffer (0.1 M Tris-HCL, pH 9.0; 0.1 M EDTA; 1% SDS, and 0.5%-1% DMDC added freshly). After incubation at 70°C for 30 min, samples were lysed by adding 140 µl Potassium acetate, re-incubated for 30 min on ice, centrifuged, precipitated in ½ volume isopropanol, washed with 70% ethanol and eluted in 100 µL 10 mM Tris + 1 µl RNAse A (10 mg/ml stock). Genomic PCRs were performed with primers Mdmd_F1/ Mdmd_R4 for *Mdmd* and Md-ncm_6/ Md-ncm_7 for *Md-ncm*. For cDNA amplifications primer pairs

Mdmd_7s/Mdmd_8as were used for *Mdmd*, and Md-ncm_1s/Md-ncm_2as and Md-ncm_9s/Md-ncm_10as for *Md-ncm*. PCR was performed in 50 µl volume containing 10 µl 5x GoTaq Reaction Buffer, 1.5 µl 25 mM $MgCl_2$, 1.5 µl 10 mM of each dNTP, 0.3 µl GoTaq DNA Polymerase (Promega), 1.5 µl 10 µM of each primer and 2 µl genomic DNA. PCR thermal cycling consisted of 2 min initial denaturation at 94°C, followed by 35 cycles of 30 s at 92°C, 30 s at 59°C and 30 s at 72°C, and a final elongation of 3 min at 72°C.

For developmental expression profiling, we collected the following samples (eggs to adults) from the M^{III} strain: eggs, embryos (0-1h, 1-2h, 2-3h, 3-4h, 4-5h, 5-6h, 6-7h and 7-22h of age), larvae (mixed L1 stage, mixed L2 stage). L3 stage larvae, pupae and adults were collected from male-only and female-only progeny (Sharma *et al.*, 2017). For expression analysis of the downstream gene, *Md-tra*, total RNA was extracted and cDNA prepared from single flies as described in chapter 2. Amplifications were performed on the cDNA with primers Md-tra9/Md-tra20 for detecting *Md-tra*^{F} transcripts and primers Mdtra12Bs and Mdtra20as for detecting *Md-tra*^{M} transcripts.

Mdmd cloning

For phylogenetic analysis, we needed the open reading frame (ORF) of *Mdmd* from all *M*-strains. For cloning of *Mdmd*, genomic DNA and cDNA was prepared following the protocol described above. PCR products were purified with the Wizard® Genomic DNA Purification Kit (Promega), subcloned in the pGEM®-T Easy Vector (Promega). The plasmid was transformed in *Escherichia coli* culture medium following standard procedures and by blue-white screening, the recombinant colonies were picked for Midipreps preparation (Sigma-Aldrich). The Midipreps were outsourced to GATC company for sanger sequencing with primer pair Mdmd F1 and Mdmd R4.

Alignment of nucleotide and protein sequences

For nucleotide and protein sequence alignment CLUSTAL O (1.2.3) multiple sequence alignment tool (McWilliam *et al.*, 2013) was used. MDMD and Md-NCM domains (MIF4G and MA3) are aligned with the following insect species; *Musca domestica* (*Md*), *Drosophila melanogaster* (*Dm*), *Ceratitis capitata* (*Cc*), *Aedes aegypti* (*Aa*), *Nasonia vitripennis* (*Nv*), *Bombyx mori* (*Bm*) and *Tribolium castaneum* (*Tc*). Identification of conserved domain motifs was done with the NCBI conserved domain search program (Marchler-Bauer and Bryant, 2004). Protein sequence identity between *Mdmd*^{II}, *Mdmd*^{III}, *Mdmd*^{V}, *Mdmd*^{Y} and *Md-ncm* was analysed with the Geneious program with settings Cost matrix = BLOSUM, Gap open cost = 10 and Gap extend cost = 0.1; version 8.1.9 created by Biomatters (Kearse *et al.*, 2012).

Phylogenetic analysis

To determine the phylogenetic relationship of the *Mdmd* and *Md-ncm* genes, the protein sequences of *Mdmd*^{II}, *Mdmd*^{III}, *Mdmd*^{V}, *Mdmd*^{Y} and *Md-ncm* were compared to *ncm* genes of other species: the budding yeast *Saccharomyces cerevisiae (Sc),* roundworm *Caenorhabditis elegans* (*Ce*)*, zebrafish Danio rerio* (*Dr*), african clawed frog *Xenopus laevis* (*Xl*), domestic chicken *Gallus gallus* (*Gg*), house mouse *Mus musculus* (*Mm*), human *Homo sapiens* (*Hs*), jewel wasp *Nasonia vitripennis* (*Nv*), silk moth *Bombyx mori* (*Bm*), red flour beetle *Tribolium castaneum* (*Tc*), yellow fever mosquito *Aedes aegypti* (*Aa*), vinegar fly *Drosophila melanogaster* (*Dm*), tsetse fly *Glossina morsitans* (*Gm*), honey bees *Apis mellifera* (*Am*), and the mediterranean fruit fly *Ceratitis capitata* (*Cc*). The protein variants of the four different *M. domestica* strains, *Mdmd*^{II}, *Mdmd*^{III}, *Mdmd*^{V} and *Mdmd*^{Y}, were aligned with the identified CWC22/NCM homologs from these sixteen species with Geneious version 8.1.9 (Biomatters, Auckland, New Zealand) (Kearse *et al.*, 2012).

RESULTS

Md-ncm is an autosomal paralog of *Mdmd*

A BLASTN search of *Mdmd* in the female genome scaffolds of *M. domestica* (Scott *et al.*, 2009) identified a gene (LOC101896466) with a high degree of similarity. This gene is an ortholog of the phylogenetically deeply conserved core spliceosomal factor *CWC22*, named *nucampholin* in *D. melanogaster* (Fig. 3.1). We name it hereafter *Md-ncm* (*Musca domestica-nucampholin*). Amino acid residues of *Mdmd* (1190) and *Md-ncm* (1310) are mainly conserved in the central part and variations are mainly found at the amino and carboxyl terminal ends. MDMD and Md-NCM share two conserved domains, MIF4G and MA3, separated by an intron at the same position (Fig. 3.1). MDMD shares a high degree of identity with Md-NCM in the MIF4G (85%) and MA3 (79%) domains.

Md-ncm is located just downstream of a gene which is homologous to the *bicoid stability factor* (*bsf*) in *Drosophila*. This constellation corresponds to the situation in *D. melanogaster* where *ncm* resides downstream of *bsf*. Based on this conserved synteny we consider it very likely that *Md-ncm* is the true ortholog of *ncm* in *D. melanogaster*.

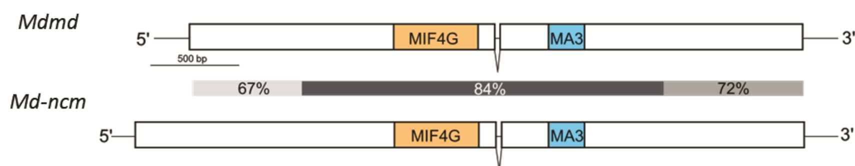

Figure 3.1. Structure of *Mdmd* and *Md-ncm*. Both genes contain two conserved domains, MIF4G (orange) and MA3 (blue), and share an intron at the same position. The amino and carboxy terminal ends diverge in sequence. The shorter ORF of *Mdmd* is due to several small in-frame deletions at the amino-terminal end. Nucleotide identity for the central part and both ends are indicated in percentages as different shades of grey.

Nucleotide sequence alignment of *Mdmd* and *Md-ncm*

To check the similarity between *Mdmd* ($Mdmd^{II}$, $Mdmd^{III}$, $Mdmd^{V}$ and $Mdmd^{Y}$), and *Md-ncm* sequences, we aligned the corresponding nucleotide sequences. *Mdmd* sequences of the different *M* strains are almost identical with very few substitutions. However, they considerably deviate from *Md-ncm* due to various deletions and many nonsynonymous nucleotide substitutions (Fig. 3.2).

Figure 3.2. Alignment of *Md-ncm* and *Mdmd* nucleotide sequences. $Mdmd^{II}$, $Mdmd^{III}$, $Mdmd^{V}$, $Mdmd^{Y}$ are compared with *Md-ncm*. Each sequence harbours a full-length open reading frame (ORF). Turquoise colours indicate intron boundaries. Asterisks indicate identical positions (see Appendix 3.2).

Protein sequence alignment of MDMD and Md-NCM

Comparison of MDMD and Md-NCM protein sequences show several differences in amino acid residues at the amino and carboxyl terminal ends. Interestingly, within MDMD all amino acid residues of both conserved domains MIF4G and MA3 are the same, but they all differ for several amino acids from Md-NCM in both domains (Fig. 3.3).

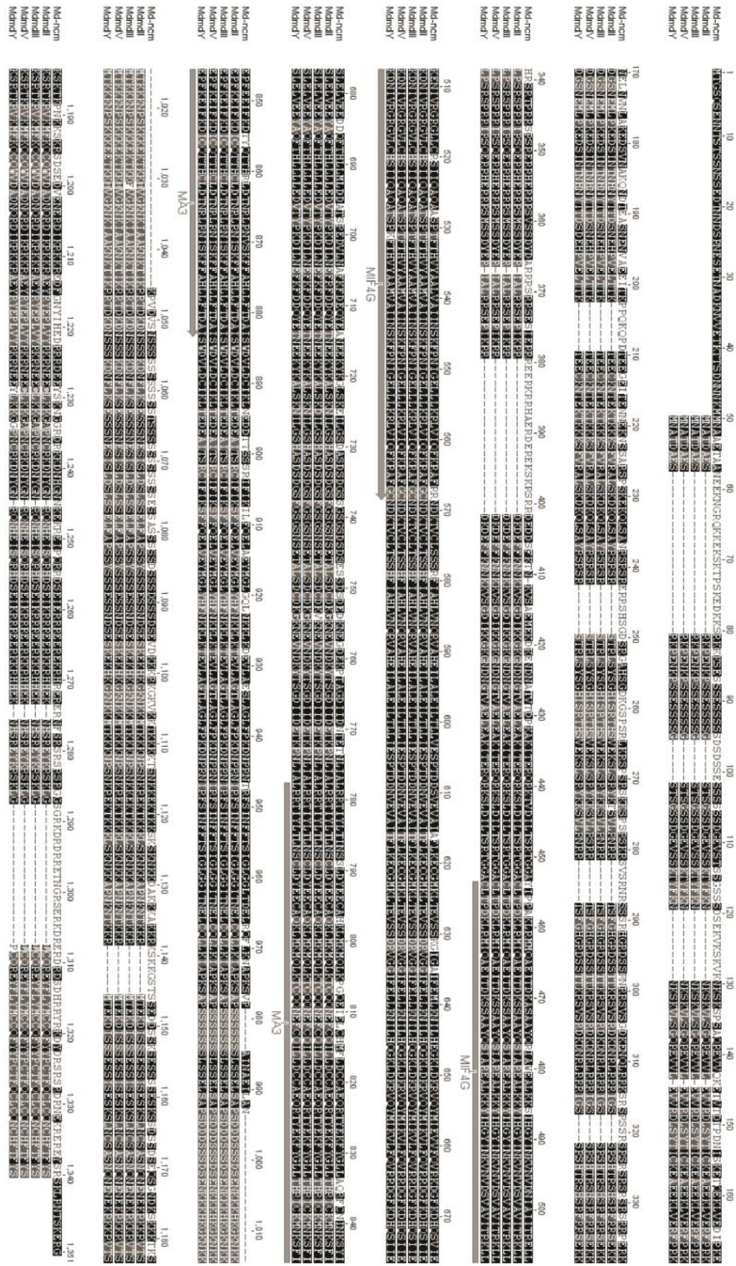

Figure 3.3. Alignment of Md-NCM and MDMD protein sequences. *Mdmd*^{II}, *Mdmd*^{III}, *Mdmd*^{V}, and *Mdmd*^{Y} are compared with *Md-ncm*. Differences in amino acid sequences between *Mdmd* copies and *Md-ncm* are shown in grey scale. Darker colours are showing higher conservation and lighter colours showing lower conservation. The MIF4G and MA3 domains are indicated by grey shading.

Genomic and expression analysis of *Mdmd* and *Md-ncm*

Genome amplifications indicated that *Mdmd* is present in males of *XY*, M^{II}, M^{III} and M^{V} strains but absent in females. In contrast, the predicted spliceosomal factor *Md-ncm* is present in both sexes of these strains (Fig. 3.4). Furthermore, cDNA amplifications show that *Mdmd* transcripts are present in 1-5 h old male embryos (M^{III}/+) and absent in female embryos (+/+) of the same stage. *Md-ncm* transcripts are found in 1-5 h old embryos of both males (M^{III}/+) and females (+/+) (Fig. 3.4). Based on this result, we can conclude that *Mdmd* is male-specific and *Md-ncm* is present in both males and females. RT-PCR analysis confirmed the results of chapter 2 that *Mdmd* is active only in males of the M^{III} strain and already at a very early stage of development, whereas *Md-ncm* is active in both in males and females consistent with its predicted essential function in RNA processing.

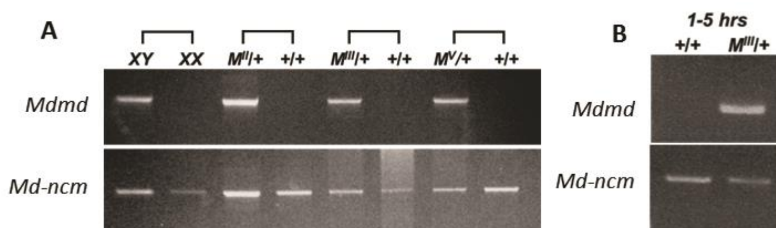

Figure 3.4. gDNA and cDNA amplifications of *Mdmd* and *Md-ncm*. (A) Genomic amplifications with *Mdmd*-specific primers (Mdmd_F1/Mdmd_R4) only show amplifications in adult males of the *XY*, M^{II}, M^{III} and M^{V} strains (upper panel), whereas *Md-ncm* (Md-ncm_6/ Md-ncm_7) is amplified in both adult males (M^{III}/+) and females (+/+) of each strain (lower panel) (B) RT-PCR showing *Mdmd* transcripts in 1-5 h old male embryos (M^{III}/+), but not in female embryos (+/+) (upper panel). *Md-ncm* is expressed in both male and female embryos (lower panel).

Developmental transcription profiles of *Mdmd*, *Md-ncm* and *Md-tra*

To examine when *Mdmd* and *Md-ncm* are active during development, we monitored expression of *Mdmd* and *Md-ncm* at different developmental stages collected from the M^{III} strain. Zygotic *Mdmd* transcripts can be detected from early zygote (2-3 h after egg laying) to adulthood only in male individuals. This includes the early stages when *Md-tra* pre-mRNA is first spliced into the male mode, resulting in a truncated, non-functional protein. A continuous expression pattern throughout all tested stages is observed for *Md-ncm* but in both male and female samples (Figure 3.5).

To monitor expression of *Md-tra* in the early zygote, we investigated the expression profile of the male-specific, non-functional form of *tra*, *Md-tra*^{M}, and female-specific, functional form of *tra*, *Md-tra*^{F} in the same M^{III} strain (Figure 3.5). *Md-tra*^{M} transcripts first appear around 2-3 h after fertilization and from then on are present in male samples up to the adult stage. *Md-tra*^{F} transcripts are present in eggs and L1 larvae, as well as in in female larvae L3, pupae, and adults. Note that the L1 and L2 larval stages are of mixed sex. *Md-tra*^{F} transcripts are already present in unfertilized eggs because they are maternally provided to eggs by the mother. The sizes of the two bands in *Mdmd*, *Md-ncm*, and *Md-tra*^{M} transcripts correspond to the predicted sizes of cDNA derived from the spliced mRNA versus cDNA derived from unprocessed transcripts and/or genomic DNA. We conclude that *Mdmd* and *Md-ncm* are constitutively active during development and that the presence of *Mdmd* transcripts coincides with male-specific splicing of *tra* (*Md-tra*).

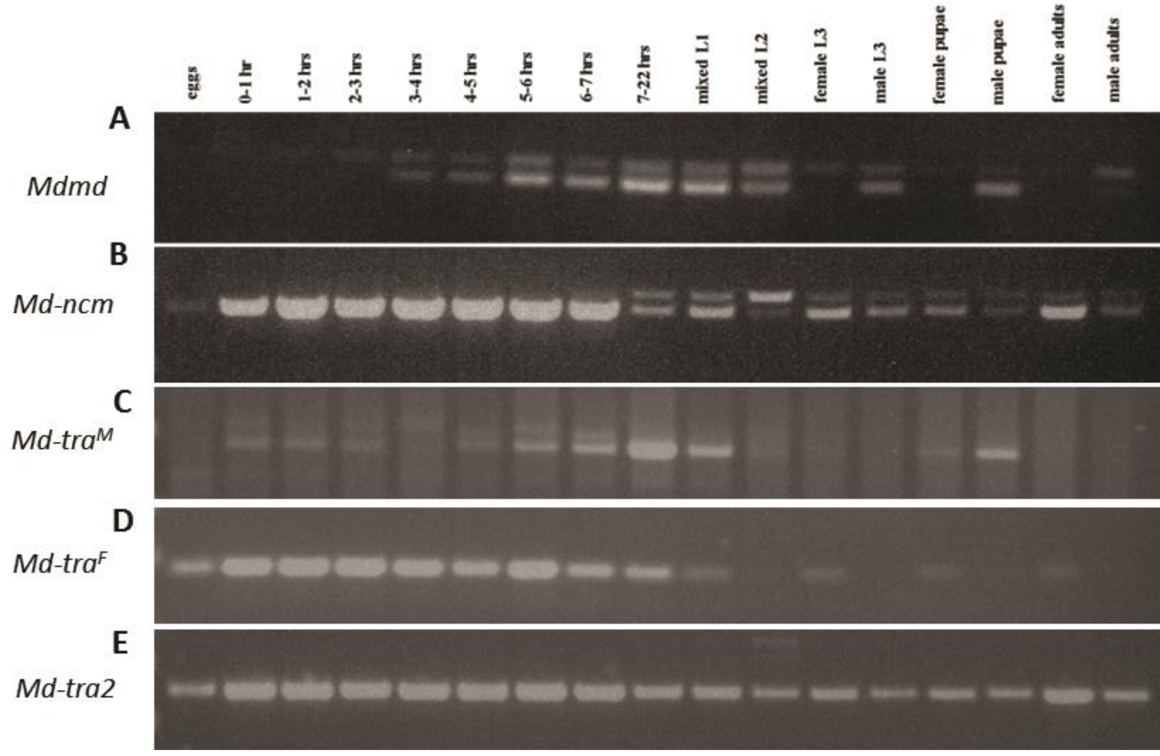

Figure 3.5. Developmental expression profiles of *Mdmd*, *Md-ncm* and *Md-tra* in the M^{III} strain. (A) *Mdmd* is expressed in 2-3 h old embryos (cellularized blastoderm stage). (B) *Md-ncm* is expressed in both males and females in all developmental stages (eggs to adults). (C) *Md-tra*^{M} can be detected in male larvae (L3) and male pupae. (D) *Md-tra*^{F} s only expressed in female larvae L3, female pupae, and female adults. (E) *Md-tra2* is a positive control for the presence of mRNA and continuously expressed during all developmental stages of both males and females.

Alignment of domains MIF4G and MA3, with *Ncm* of other insect species

To determine the phylogenetic relationship between *Mdmd* and *Md-ncm*, we aligned the conserved domains MIF4G and MA3 of $Mdmd^{II}$, $Mdmd^{III}$, $Mdmd^{V}$, $Mdmd^{Y}$ and *Md-ncm* protein sequences with the same domains of other insect species. The amino acids of the two domains are identical in *Mdmd* but, despite being clearly homologous, have some differences with the other insect species (Fig. 3.6).

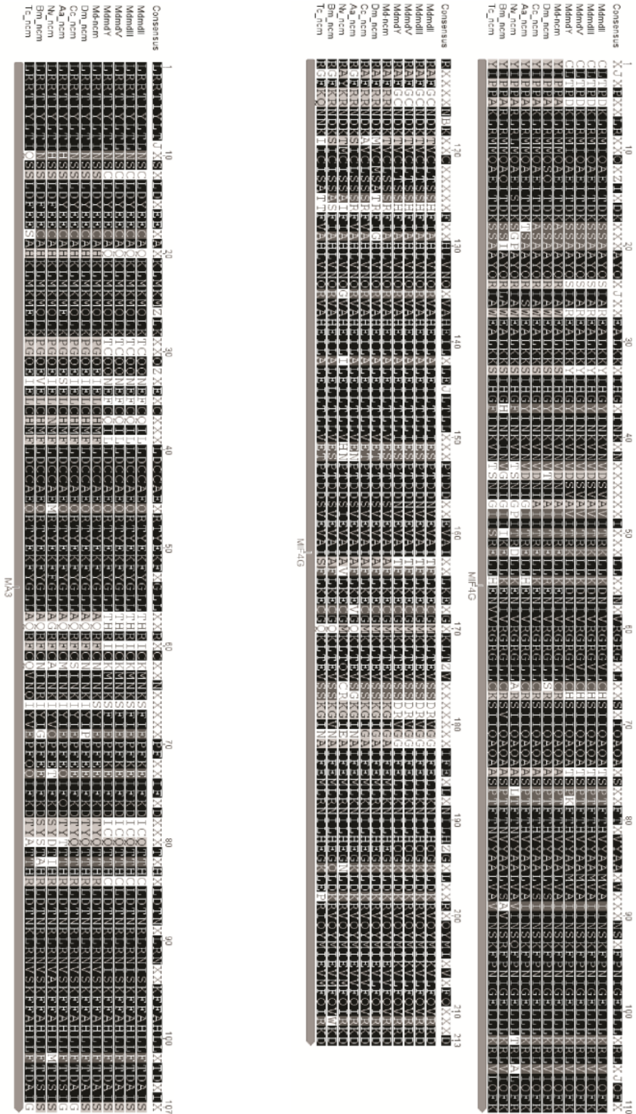

Figure 3.6. Alignment of conserved domains MIF4G and MA3 of *Mdmd* and *Md-ncm* with *Ncm* of other insects. Conserved domains (MIF4G and MA3 in grey shading) of *Mdmd* and *Md-ncm* are aligned with the following insect species; *M. domestica* (*Md*), *D. melanogaster* (*Dm*), *C. capitate* (*Cc*) *A. aegypti* (*Aa*), *N. vitripennis* (*Nv*), *B. mori* (*Bm*) and *T. castaneum* (*Tc*). Conservation of amino acid sequences of MIF4G and MA3 domains of *Mdmd, Md-ncm* and *Ncm* are shown in grey scale. Darker colours are showing higher conservation and lighter colours showing lower conservation.

Phylogeny of *Mdmd* and *Md-ncm*

A phylogenetic analysis of *Mdmd* (*Mdmd*^{II}, *Mdmd*^{III}, *Mdmd*^{V} and *Mdmd*^{Y}) and *Md-ncm* protein sequences revealed an order of similarity that is consistent with their phylogenetic distance. Thus, *Md-ncm* groups with prototype *ncm* genes of other insect species but *Mdmd* sequences show distinct divergence from other metazoans. No ortholog to *Mdmd* was found in the entire NCBI database (Fig. 3.7).

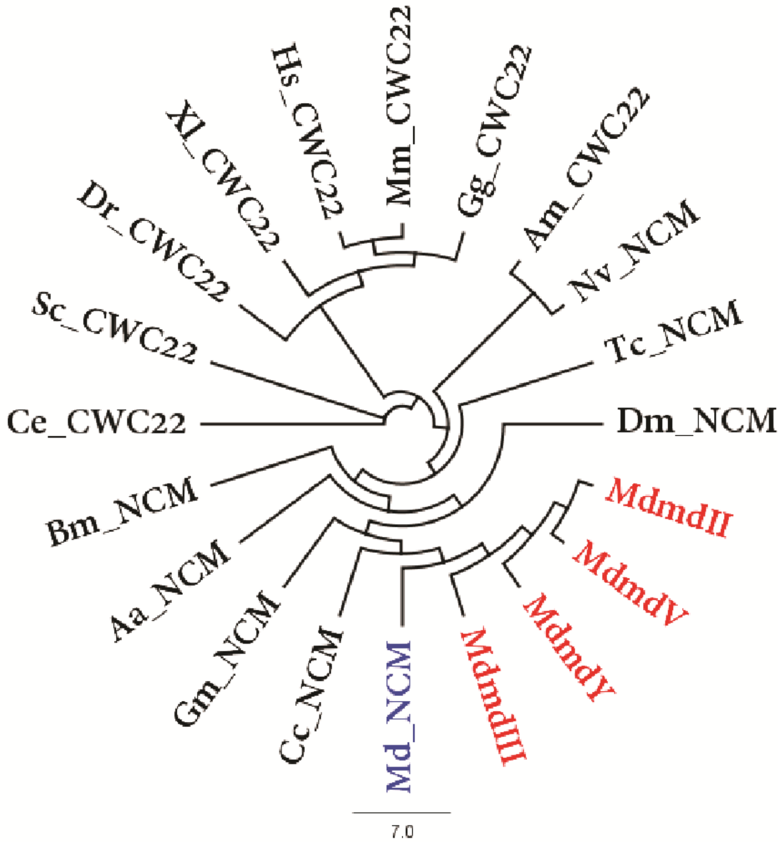

Figure 3.7. Phylogenetic relation of *Mdmd* and *Md-ncm* in several metazoans. Phylogenetic analysis indicates that *Mdmd* originated from a gene duplication of *M. domestica-ncm* (*Md-ncm*) and is unique to *M. domestica*. Species abbreviations are: (Diptera: *Musca domestica* (*Md*), *Glossina morsitans* (*Gm*), *Ceratitis capitata* (*Cc*), *Aedes aegypti* (*Aa*) and *Drosophila melanogaster* (*Dm*); Hymenoptera: *Apis mellifera* (*Am*) and *Nasonia vitripennis (Nv)*; Coleoptera: *Tribolium castaneum* (*Tc*); Lepidoptera: *Bombyx mori* (*Bm*); Rodentia: *Mus musculus* (*Mm*); Primates: *Homo sapiens* (*Hs*); Anura: *Xenopus laevis* (*Xl*); Galliformes: *Gallus gallus* (*Gg*); Cypriniformes: *Danio rerio* (*Dr*); Rhabditida: *Caenorhabditis elegans* (*Ce*) and Saccharomycetales: *Saccharomyces cerevisiae* (*Sc*).

DISCUSSION

Mdmd is a paralog of *Md-ncm*

In the previous chapter 2, we identified a potential candidate for the male-determining signal in *M. domestica* from the M^{III} strain which we named *Mdmd*. One immediate follow-up question was whether *Mdmd* is also present in strains with *M* on the Y chromosome or on any of the other autosomes. In this chapter, we found that *Mdmd* is present in males of different *M*-strains (M^{II}, M^{III}, M^{V}, and M^{Y}) and is the same gene. These results suggest that *Mdmd* translocated to different genomic sites in the *M. domestica* genome. *Mdmd* sequences of different *M*-strains share a high level of identity with only very few nucleotide substitutions. So based on this, we can infer that Y-linked and autosomal *M*-factors are phylogenetically very closely related and must have emerged from recent translocations.

The other interesting part is that BLASTN searches of *Mdmd* into the female genome of *M. domestica* identified its paralogous gene *Md-ncm*. The high similarity between *Mdmd* and *Md-ncm* is an indication of a duplication event. We also compared the nucleotide and protein sequence similarity of *Mdmd* and *Md-ncm*. The nucleotide sequence of *Mdmd* considerably deviates from *Md-ncm* due to many deletions and nonsynonymous nucleotide substitutions. These variations are mainly found at the amino and carboxyl terminal ends. This analysis further revealed that *Mdmd* has a common origin and subsequently translocated to different genomic sites.

Our phylogenetic analysis revealed a distinct divergence pattern of *Mdmd, Md-ncm,* and different *ncm* sequences among the different phyla of metazoans. *Mdmd* sequences from different *M* strains form a separate outgroup suggesting that after the duplication event *Mdmd* rapidly diverged from *Md-ncm* before translocating to new genome sites. We also conclude that recruitment of *Mdmd* may be unique for the housefly as *ncm* paralogs have thus far not been found in other higher dipterans.

Mdmd and *Md-ncm* share similar sequences (>85% identical amino acid residues) and are constitutively expressed in a very early stage of development. The main difference between both genes is the sex-specific mode of activation. *Mdmd* is male-specific, but *Md-ncm* is expressed in both males and females. These findings further suggest that *Mdmd* is a paralog of essential splicing factor *Md-ncm*.

Based on the results presented in this chapter we can draw the following conclusions: (1) translocation of the male determiner candidate *Mdmd* to different genomic locations is the likely cause for the evolution of different autosomal *M* strains, (2) *Mdmd*

expression profile coincides with presence of *Md-tra*^{M} transcripts consistent with the predicted role of *M* as a repressor of female splicing of *Md-tra*, and (3) *Mdmd* emerged as a recent duplication of the spliceosomal factor *Md-ncm*.

ACKNOWLEDGEMENTS

We are thankful to Claudia Brunner for technical support and Prof. Ernst Wimmer and Yanli Wu for providing *Mdmd*^{V} sequence information.

APPENDIX 3.1

Primer sequences

PRIMER NAME	PRIMER SEQUENCE (5'-3')	PURPOSE
Mdmd_F1	CACTCGTTTCAGAACTTTGGGT	*Mdmd* specific
Mdmd_R4	GTGTTTGATAGCAAGAATTAGGAGT	*Mdmd* specific
Md-ncm_1s	CGCAGAGATGGCTTTAAGGA	cDNA amplifications
Md-ncm_2as	TTTTTGGGCACATTCCTCAT	cDNA amplifications
Md-ncm_6	AGAAGAATGGACGCCAGAAA	*Md-ncm* specific
Md-ncm_7	GTCAGGTTGCTTCTGAGGCG	*Md-ncm* specific
Md-ncm_9s	CTCATCGAAGGGCATTGGAGC	cDNA amplifications
Md-ncm_10as	AATAATATCGCCAGCTGTGGGTTT	cDNA amplifications
Mdmd_7s	ATCATCGGATAGGGTTGGAGG	cDNA amplifications
Mdmd_8as	ATAATGTCACCACTCGTGTATTTA	cDNA amplifications
Mdtra-12Bs	GTGATCCTAACAATCAGCTAG	cDNA amplifications
Mdtra-20as	TTGCTGCTGGGGGAATGTG	cDNA amplifications
Mdtra9	CTGCTACAGAAAAGAAAGGCC	*Md-traF*
Md-tra20	TTGCTGCTGGGGGAATGTG	*Md-traF*

APPENDIX 3.2

Nucleotide sequence alignment of *Mdmd* and *Md-ncm* (Figure 3.2).

```
            1        10        20        30        40        50        60
            |         |         |         |         |         |         |

Md-ncm      CATATCCCTACTGTGTGCTGTTGGCGGAGATATTTTACTTGGGACTCGAAACGTTTTGTG
MdmdIII     -----TGGAAAATTACGATGTTCTGAGCCACCCTATAATTAGGGAAAA------------
MdmdY       ------------------------------------------------------------
MdmdII      ------------------------------------------------------------
MdmdV       ------------------------------------------------------------

Md-ncm      AAGATATTAGGAAAAGAAATCTAGAGAAACCACGGCCATGACTGG---CAGTCAATCTGA
MdmdIII     --AATCGATGCTATCGATAGGTAGCAACGTTACTAACTTCACTCGTTTCAGAACTTTGGG
MdmdY       ---------------------------------------CACTCGTTTCAGAACTTTGGG
MdmdII      ---------------------------------------CACTCGTTTCAGAACTTTGGG
MdmdV       ---------------------------------------CACTCGTTTCAGAACTTTGGG
                                                    *** *   ***    *  *

Md-ncm      AAATAACACCTCCACTTCGTCAAATTCTTCGGAGGATACTAATAACGACTCACGAAATGA
MdmdIII     TATTAAAAGACGGATTACTTACCATTGTTCTCTTAATGGCAATATAT------TTCTTAG
MdmdY       TATTAAAAGACGGATTACTTACCATTGTTCTCTTAATGGCAATATAT------TTCTTAG
MdmdII      TATTAAAAGACGGATTACTTACCATTGTTCTCTTAATGGCAATATAT------TTCTTAG
MdmdV       TATTAAAAGACGGATTACTTACCATTGTTCTCTTAATGGCAATATAT------TTCTTAG
             * *** *     * * * *   *** ***     **   ****             *

Md-ncm      AAGCGAAACAAATGCTGACAAAAATGTGGTAGAGACAAAAACAACGT---CTGACAACAA
MdmdIII     ATTCG--------------AAACCTTTTGTCGATATATTAAGAAAATAACCTGGCGGTTA
MdmdY       ATTCG--------------AAACCTTTTGTCGATATATTAAGAAAATAACCTGGCGGTTA
MdmdII      ATTCG--------------AAACCTTTTGTCGATATATTAAGAAAATAACCTGGCGGTTA
MdmdV       ATTCG--------------AAACCTTTTGTCGATATATTAAGAAAATAACCTGGCGGTTA
            *  **              ***  * * ** ** * *  ** **  *   *** *    *

Md-ncm      -CAACAAAATGATGAACGCCGCCGAGACCGCTGCCAATGAGGAGAAGAATGGACGCCAGA
MdmdIII     ATCTGCACGTAACACCCGCAGTTTATCATATTCTCATAGGATTTCAATAACGACTCTCGA
MdmdY       ATCTGGACGTAACACCCGCAGTTTATCATATTCTCATAGGATTTCAATAACGACTCTCGA
MdmdII      ATCTGGACGTAACACCCGCAGTTTATCATATTCTCATAGGATTTCAATAACGACTCTCGA
MdmdV       ATCTGGACGTAACACCCGCAGTTTATCATATTCTCATAGGATTTCAATAACGACTCTCGA
                  *  * *    *** *   *      *  **  *      *  *  *** *  **

Md-ncm      AAAAGGAAAAATCAAAAACACCTTCAAAAGAGGACAAGAAATCACGGAAAAAGAAAAGTG
MdmdIII     ATAAA---CAATAAACAGATCCTGACAACAACAAAAATATGAATGCCACCGACGCCGAAT
MdmdY       ATAAA---CAATAAACAGATCCTGTCAACAACAAAAATATGAATGCCACCGACGCCGAAT
MdmdII      ATAAA---CAATAAACAGATCCTGTCAACAACAAAAATATGAATGCCACCGACGCCGAAT
MdmdV       ATAAA---CAATAAACAGATCCTGTCAACAACAAAAATATGAATGCCACCGACGCCGAAT
            * **     *** ** *   ***   **  *  * ** *        *   *

Md-ncm      AGTCCAGTTCTGAGTCTTCATCGTCTTCTGATTCCGACTCGTCGGAGTCCTCAAGTAGTT
MdmdIII     CTCGAAAACCGGAAAATAAACCTAGTTCTGAGTCTTCGTCGTCGGGGTCCACAAGTGGAT
MdmdY       CTCGAAAACCGGAAAATAAACCTAGTTCTGAGTCTTCGTCGTCGGGGTCCACAAGTGGAT
MdmdII      CTCGAAAACCGGAAAATAAACCTAGTTCTGAGTCTTCGTCGTCGGGGTCCACAAGTGGAT
MdmdV       CTCGAAAACCGGAAAATAAACCTAGTTCTGAGTCTTCGTCGTCGGGGTCCACAAGTGGAT
                 *   * **   *  * *   ****** **    ******* **** ***** * *

Md-ncm      CAAGTGATGGCGAAGTATCTACCAGCTCCGGTTCATCGTCAGACAGTGAGAAGGTTAAAA
MdmdIII     CAAGTGATGGAGAAGTGTCTTCCAAAACATATTTCAAA----------------------
MdmdY       CAAGTGATGGAGAAGTGTCTTCCAAAACATATTTCAAA----------------------
MdmdII      CAAGTGATGGAGAAGTGTCTTCCAAAACATATTTCAAA----------------------
MdmdV       CAAGTGATGGAGAAGTGTCTTCCAAAACATATTTCAAA----------------------
            ********** ***** *** ***   *   **

Md-ncm      GTAAAGTCAAAAACAAATCCAAAAGTCCATCAGCACAAAGGGAAGTAGTCCAAAAGGAAA
MdmdIII     --------AATAACAAATCAAAAGTTCTATCAGGGCAAAGGGAAG---TCGTATTGGAAG
MdmdY       --------AATAACAAATCAAAAGTTCTATCAGGGCAAAGGGAAG---TCGTATTGGAAG
```

```
MdmdII      --------AATAACAAATCAAAAGTTCTATCAGGGCAAAGGGAAG---TCGTATTGGAAG
MdmdV       --------AATAACAAATCAAAAGTTCTATCAGGGCAAAGGGAAG---TCGTATTGGAAG
                    ** ******** ***  ** *****  **********   **  *  ****

Md-ncm      CAGTTACAGATACGCCTGATAATATATCCAAGGAAACACAAGAGAAGTTGGTTGAAGATA
MdmdIII     TAGTACGTGATCTATCTTATACTATATGCAAGGAAGCAGAAGAAAAATTAGTTGAAAGAT
MdmdY       TAGTACGTGATCTATCTTATACTATATGCAAGGAAGCAGAAGAAAAATTAGTTGAAAGAT
MdmdII      TAGTACGTGATCTATCTTATACTATATGCAAGGAAGCAGAAGAAAAATTAGTTGAAAGAT
MdmdV       TAGTACGTGATCTATCTTATACTATATGCAAGGAAGCAGAAGAAAAATTAGTTGAAAGAT
             ***    ***    ** *** ***** ******* ** **** ** ** ******

Md-ncm      TTCCTAAGGAGAATGAATTGAATGTAAATTTGGCAAAGGAAGATAGCGTTAATGCCAAGC
MdmdIII     TTCCTAGAAAGGATGGATCGAATGAAATGCTTCCCAAGGAAGATAGCATAAATACTAACC
MdmdY       TTCCTAGAAAGGATGGATCGAATGAAATGCTTCCCAAGGAAGATAGCATAAATACTAACC
MdmdII      TTCCTAGAAAGGATGGATCGAATGAAATGCTTCCCAAGGAAGATAGCATAAATACTAACC
MdmdV       TTCCTAGAAAGGATGGATCGAATGAAATGCTTCCCAAGGAAGATAGCATAAATACTAACC
            ******   ** *** ** ***** **    *   * *********** * *** * ** *

Md-ncm      AAAATGACACAGAGGCTTCAAATGAGAACGTAGCGGAAGAAATAACAAAACCGCCTCAGA
MdmdIII     ATAATTACACAACGGATTCAAATGAACACCCAGTAGAACTAACGA---------------
MdmdY       ATAATTACACAACGGATTCAGATGAACACCCAGTAGAACTAACGA---------------
MdmdII      ATAATTACACAACGGATTCAAATGAACACCCAGTAGAACTAACGA---------------
MdmdV       ATAATTACACAACGGATTCAAATGAACACCCAGTAGAACTAACGA---------------
            * *** *****  ** **** ****  **  **  ***  **  *

Md-ncm      AGCAACCTGACACTGAAGAAGGTGAAATTACAGAAAATAATGAAAAAAATTCTTCAGCGA
MdmdIII     ------CAAAGACAGAAGAATGTAAAAATACAGAGAAGACTAAAAAAAAATCTTTTGTGC
MdmdY       ------CAAAGACAGAAGAATGTAAAAATACAGAGAAGACTAAAAAAAAATCTTTTGTGC
MdmdII      ------CAAAGACAGAAGAATGTAAAAATACAGAGAAGACTAAAAAAAAATCTTTTGTGC
MdmdV       ------CAAAGACAGAAGAATGTAAAAATACAGAGAAGACTAAAAAAAAATCTTTTGTGC
                  *  * ** ****** ** *** ****** ** * * ******* ****  * *

Md-ncm      GATCTCCTTCCAAAGAAAAACAACTGTCCGCCAATAGAAGCCGTTCTAGAGAAAGAAGAA
MdmdIII     GAGCTCTCTCCAAAGATAAACAACTATCCGCCTATAGAAGCCGTTCTAGAAGCACACGTT
MdmdY       GAGCTCTCTCCAAAGATAAACAACTATCCGCCTATAGAAGCCGTTCTAGAAGCACACGTT
MdmdII      GAGCTCTCTCCAAAGATAAACAACTATCCGCCTATAGAAGCCGTTCTAGAAGCACACGTT
MdmdV       GAGCTCTCTCCAAAGATAAACAACTATCCGCCTATAGAAGCCGTTCTAGAAGCACACGTT
            ** ***  ******** ******** ****** ****************   * * *

Md-ncm      GTC---ATTCTGGAGACAGTAGAGGAGCTCATTCAGGAGACAAAGGAAGTCCTTCAAGAC
MdmdIII     TAAGTTATTCTGGACACATTAGCCGAACTCATTCAGTAGAA-------------------
MdmdY       TAAGTTATTCTGGACACATTAGCCGAACTCATTCAGTAGAA-------------------
MdmdII      TAAGTTATTCTGGACACATTAGCCGAACTCATTCAGTAGAA-------------------
MdmdV       TAAGTTATTCTGGACACATTAGCCGAACTCATTCAGTAGAA-------------------
                  ******** *** ***  ** ********* ***

Md-ncm      TTAAAAGAAGTCCTTCAAGATCTAAAAAAAGTCCATCACGTTCTAAAAGGAGTGTATCAA
MdmdIII     -----AAAAGTCTATCAAGATATAAAAAAAGT------------------GTATTAAGAA
MdmdY       -----AAAAGTCTATCAAGATATAAAAAAAGT------------------GTATTAAGAA
MdmdII      -----AAAAGTCTATCAAGATATAAAACAAGT------------------GTATTAAGAA
MdmdV       -----AAAAGTCTATCAAGATATAAAAAAAGT------------------GTATTAAGAA
                 * *****  ******* ***** ****                       **  **

Md-ncm      GAAACAGAAGTTCCTCCAGACACGGCCGAGATTCATCTAGGAACAGACGAAGTGTTTCCG
MdmdIII     ATCGAAGAACTTCCTTTGGACACGGCCGAGATTCTTCTACGACCAAAAGGTCTGTCTCCA
MdmdY       ATCGAAGAACTTCCTTTGGACACGGCCGAGATTCTTCTACGACCAAAAGGTCTGTCTCCA
MdmdII      ATCGAAGAACTTCCTTTGGACACGGCCGAGATTCTTCTACGACCAAAAGGTCTGTCTCCA
MdmdV       GTCGAAGAACTTCCTTTGGACACGGCCGAGATTCTTCTACGACCAAAAGGTCTGTCTCCA
                 **** *****   *************** **** ** ** * *   *** ***

Md-ncm      GCGATAAGGAGAACCAACTAAGACGGAAATCAAGGTCAAGATCTAGCAGAAGTCGTAGTA
MdmdIII     GGGACAAGGACAATCGACTAAGACGAAGAATA------GGATCTAGCAGAAGCCATACTA
MdmdY       GGGACAAGGACAATCGACTAAGACGAAGAATA------GGATCTAGCAGAAGCCATACTA
MdmdII      GGGACAAGGACAATCGACTAAGACGAAGAATA------GGATCTAGCAGAAGCCATACTA
MdmdV       GGGACAAGGACAATCGACTAAGACGAAGAATA------GGATCTAGCAGAAGCCATACTA
            * ** ***** ** * ********* * *  *       ************ * ** **

Md-ncm      GATCTCGCAGTCGCAGTCGTTCTCGAAGACCAGAACGTAAACACCGATCAAGAACACCAC
MdmdIII     GATCT------CACAGTCGATTTCGAAGGTCGGAAAAGAAGCTTCCATCAAGATCTCCAC
```

```
MdmdY       GATCT------CACAGTCGATTTCGAAGGTCGGAAAAGAAGCTTCCATCAAGATCTCCAC
MdmdII      GATCT------CACAGTCGATTTCGAAGGTCGGAAAAGAAGCTTCCATCAAGATCTCCAC
MdmdV       GATCT------CACAGTCGATTTCGAAGGTCGGAAAAGAAGCTTCCATCAAGATCTCCAC
            *****      * ****** * ******  * ***   ** *  * ******* * ****

Md-ncm      GACGATCACGTTCACGAGAAAGGCGACATGAACGACGTCGTTCGGTGTCATCAGATTATG
MdmdIII     GACGAATTCGTTCACAAGAAAGACGACATGAACGACGTCGTTCGATGTCGTCAGATTATG
MdmdY       GACGAATTCGTTCACAAGAAAGACGACATGAACGACGTCGTTCGATGTCGTCAGATTATG
MdmdII      GACGAATTCGTTCACAAGAAAGACGACATGAACGACGTCGTTCGATGTCGTCAGATTATG
MdmdV       GACGAATTCGTTCACAAGAAAGACGACATGAACGACGTCGTTCGATGTCGTCAGATTATG
            *****   ******* ****** ********************* **** **********

Md-ncm      ATGCGAGAAGACGATCTCGTCGCTCTGAGTCAATCGAAAGAAGACGTGAGGAACGTAAGA
MdmdIII     AAAGGATAGCGTTAC---GCCGATCTGAGCCAATCAAAAGAAGAGATAAAG---------
MdmdY       AAAGGATAGCGTTAC---GCCGATCTGAGCCAATCAAAAGAAGAGATAAAG---------
MdmdII      AAAGGATAGCGTTAC---GCCGATCTGAGCCAATCAAAAGAAGAGATAAAG---------
MdmdV       AAAGGATAGCGTTAC---GCCGATCTGAGCCAATCAAAAGAAGAGATAAAG---------
            *   ** *     *    * ** ****** ***** ********  * * *

Md-ncm      GACGTCATGCAGAACGTGATGAAAGGGAAAAATCAAAACGATCCAGACGAGATGAAGATG
MdmdIII     ------------------------------------------------------ATGAAT
MdmdY       ------------------------------------------------------ATGAAT
MdmdII      ------------------------------------------------------ATGAAT
MdmdV       ------------------------------------------------------ATGAAT
                                                                  * **

Md-ncm      ATTCGTTCAAGACAAATAAAGTATCTGCTGAAATGGAAAAAGATAAAGAGAATGATAATG
MdmdIII     TCTTCAAAAACAATAAGAAAGTATCTGGCGATATAAAAAAGGGAAAGGGGAACGATAATG
MdmdY       TCTTCAAAAACAATAAGAAAGTATCTGGCGATATAAAAAAGGGAAAGGGGAACGATAATG
MdmdII      TCTTCAAAAACAATAAGAAAGTATCTGGCGATATAAAAAAGGGAAAGGGGAACGATAATG
MdmdV       TCTTCAAAAACAATAAGAAAGTATCTGGCGATATAAAAAAGGGAAAGGGGAACGATAATG
              *     ** *  ** **********  ** **  **** *  ** * *** *******

Md-ncm      CCACGGTAACTGATCCCAAGGCCAAGATTACAGAACGTCAAAGGAAAACTGTAGACATAC
MdmdIII     GTACTGTAGCAGAACTCGAGGCCAAGATAACAGAACGTCAAAGGAAAAGTCTAGATATAC
MdmdY       GTACTGTAGCAGAACTCGAGGCCAAGATAACAGAACGTCAAAGGAAAAGTCTAGATATAC
MdmdII      GTACTGTAGCAGAACTCGAGGCCAAGATAACAGAACGTCAAAGGAAAAGTCTAGATATAC
MdmdV       GTACTGTAGCAGAACTCGAGGCCAAGATAACAGAACGTCAAAGGAAAAGTCTAGATATAC
              ** *** * ** * * ********** ******************* * **** ****

Md-ncm      TTACATCACGTACCGGTGGAGCCTATATACCACCAGCTAAGTTGCGTATGATGCAAGCAG
MdmdIII     TAACATCACGTACAGGCGGTGCTTGTTTAACACCCGATAAATTGCGTATGATACAAGCAG
MdmdY       TAACATCACGTACAGGCGGTGCTTGTTTAACACCCGATAAATTGCGTATGATACAAGCAG
MdmdII      TAACATCACGTACAGGCGGTGCTTGTTTAACACCCGATAAATTGCGTATGATACAAGCAG
MdmdV       TAACATCACGTACAGGCGGTGCTTGTTTAACACCCGATAAATTGCGTATGATACAAGCAG
            * *********** ** ** ** * * ** **** * *** *********** *******

Md-ncm      AAATTACGGACAAAGCATCGGCTGCATATCAACGTATTGCTTGGGAAGCTCTTAAGAAAT
MdmdIII     AAATTACAGACAAATCATCGGCTGCATATCAAAGTATAGCTCGGGAAGCTCTTAAGAAAT
MdmdY       AAATTACAGACAAATCATCGGCTGCATATCAGAGTATAGCTCGGGAAGCTCTTAAGAAAT
MdmdII      AAATTACAGACAAATCATCGGCTGCATATCAGAGTATAGCTCGGGAAGCTCTTAAGAAAT
MdmdV       AAATTACAGACAAATCATCGGCTGCATATCAGAGTATAGCTCGGGAAGCTCTTAAGAAAT
            ******* ****** ****************  **** *** ******************

Md-ncm      CCATACATGGTTACATTAATAAAGTCAATGTGGATAATATTGCAATTATTACCCGAGAAC
MdmdIII     ACATACATGGTTACATCAATAAAGTCAATGTGGATAGTGTTGCAGTTATCACCCGAAAAT
MdmdY       ACATACATGGTTACATCAATAAAGTCAATGTGGATAGTGTTGCAGTTATCACCCGAAAAT
MdmdII      ACATACATGGTTACATCAATAAAGTCAATGTGGATAGTGTTGCAGTTATCACCCGAAAAT
MdmdV       ACATACATGGTTACATCAATAAAGTCAATGTGGATAGTGTTGCAGTTATCACCCGAAAAT
             *************** ******************* * ***** **** ****** **

Md-ncm      TGCTAAAGGAAAATATAGTACGTGGTAGAGGCTTGCTTTGCCGTTCCATAATACAGGCTC
MdmdIII     TGCTTAAGGATAATATAGTACGTGGTAGAGGCGTGCTTTGCCATTCCATAATACAAGCTC
MdmdY       TGCTTAAGGATAATATAGTACGTGGTAGAGGCGTGCTTTGCCATTCCATAATACAAGCTC
MdmdII      TGCTTAAGGATAATATAGTACGTGGTAGAGGCGTGCTTTGCCATTCCATAATACAAGCTC
MdmdV       TGCTTAAGGATAATATAGTACGTGGTAGAGGCGTGCTTTGCCATTCCATAATACAAGCTC
            **** ***** ********************* ********* ************ ****

Md-ncm      AAGCTGCATCGCCGACATTTACCCATGTTTATGCCGCCTTGGTGGCTATTATTAATTCGA
```

```
MdmdIII     AAGCTACATCGCCAACATTTACCCATGTTTATGCCGCCATGGTGGCCATTATTAATTCAA
MdmdY       AAGCTACATCGCCAAAATTTACCCATGTTTATGCCGCCATGGTGGCCATTATTAATTCAA
MdmdII      AAGCTACATCGCCAACATTTACCCATGTTTATGCCGCCATGGTGGCCATTATTAATTCAA
MdmdV       AAGCTACATCGCCAACATTTACCCATGTTTATGCCGCCATGGTGGCCATTATTAATTCAA
            ***** ******* * ********************** ******* *********** *

Md-ncm      AATTTCCCAATATTGGAGAATTGCTGTTGAAACGTTTGGTAATACAATTTAAACGAGCCT
MdmdIII     AATTTCCCAATATTGGGGAATTGCTGTTGAAACGTTTGGTCATACAGTTCAAACGAGCAT
MdmdY       AATTTCCCAATATTGGGGAATTGCTGTTGAAACGTTTGGTCATACAGTTCAAACGAGCAT
MdmdII      AATTTCCCAATATTGGGGAATTGCTGTTGAAACGTTTGGTCATACAGTTCAAACGAGCAT
MdmdV       AATTTCCCAATATTGGGGAATTGCTGTTGAAACGTTTGGTCATACAGTTCAAACGAGCAT
            **************** *********************** ***** ** ******** *

Md-ncm      TTCGACGTAACGATAAGACTGTGTGTTTGTCATCTTCACGTTTCATAGCCCACTTGGTCA
MdmdIII     TTGGGTGTAACGATAAGACCGTTTGTTTGACTTCTTCACATTTTATTGCCCACTTGGTTA
MdmdY       TTGGGTGTAACGATAAGACCGTTTGTTTGACTTCTTCACATTTTATTGCCCACTTGGTTA
MdmdII      TTGGGTGTAACGATAAGACCGTTTGTTTGACTTCTTCACATTTTATTGCCCACTTGGTTA
MdmdV       TTGGGTGTAACGATAAGACCGTTTGTTTGACTTCTTCACATTTTATTGCCCACTTGGTTA
            ** *  ************* ** ****** * ******* *** ** *********** *

Md-ncm      ATCAAAGGGTGGCCCATGAAATCTTGGCCTTGGAAATTCTAACGCTTTTGGTTGAATCGC
MdmdIII     ATCAAAGGGTGGCCCATGAAATTTTGGCCCTAGAAATTCTAACACTCTTAATTGAATCGC
MdmdY       ATCAAAGGGTGGCCCATGAAATTTTGGCCCTAGAAATTCTAACACTCTTAATTGAATCGC
MdmdII      ATCAAAGGGTGGCCCATGAAATTTTGGCCCTAGAAATTCTAACACTCTTAATTGAATCGC
MdmdV       ATCAAAGGGTGGCCCATGAAATTTTGGCCCTAGAAATTCTAACACTCTTAATTGAATCGC
            ********************** ****** * *********** ** **  *********

Md-ncm      CGACAGACGATAGTGTTGAAGTGGCCATAGCATTTCTTAAGGAATGTGGTATGAAGTTGA
MdmdIII     CGACTGATGATAATGTTGAAGTGGCCATAACATTTCTTAAGGAATGTGGTATGAAATTGA
MdmdY       CGACTGATGATAATGTTGAAGTGGCCATAACATTTCTTAAGGAATGTGGTATGAAATTGA
MdmdII      CGACTGATGATAATGTTGAAGTGGCCATAACATTTCTTAAGGAATGTGGTATGAAATTGA
MdmdV       CGACTGATGATAATGTTGAAGTGGCCATAACATTTCTTAAGGAATGTGGTATGAAATTGA
            **** ** **** **************** ************************* ****

Md-ncm      CAGAGGTCTCATCGAAGGGCATTGGAGCCATTTTTGAAATGCTCAAAAATATCTTGCATG
MdmdIII     CAGAGGTATCATCGGATAGGGTTGGAGGCATTTTTGAATTGCTTAAAAATATCTTGCATC
MdmdY       CAGAGGTATCATCGGATAGGGTTGGAGGCATTTTTGAATTGCTTAAAAATATCTTGCATC
MdmdII      CAGAGGTATCATCGGATAGGGTTGGAGGCATTTTTGAATTGCTTAAAAATATCTTGCATC
MdmdV       CAGAGGTATCATCGGATAGGGTTGGAGGCATTTTTGAATTGCTTAAAAATATCTTGCATC
            ******* ****** *  *  ****** ********** **** *************** 

Md-ncm      AGGGCAAGTTAGATAAACGTGTGCAGTATATGATTGAAGTTGTTTTCCAAGTACGCAAAG
MdmdIII     AAGGCAAGTTGGATAAACGTGTACAATATATGATTAAAGTTTTATTTCAAGTACGCAGAG
MdmdY       AAGGCAAGTTGGATAAACGTGTACAATATATGATTAAAGTTTTATTTCAAGTACGCAGAG
MdmdII      AAGGCAAGTTGGATAAACGTGTACAATATATGATTAAAGTTTTATTTCAAGTACGCAGAG
MdmdV       AAGGCAAGTTGGATAAACGTGTACAATATATGATTAAAGTTTTATTTCAAGTACGCAGAG
            * ******** *********** ** ********* ***** * ** ********** **

Md-ncm      ATGGTTTTAAGGATCATCAATCGGTCATAGAGTCCTTGGAATTGGTAGAAGAGGATGATC
MdmdIII     ATGGCTTCAAGGACCATCAATCGATTATTGAGTCATTAGAATTGGTAGAAGAATATGCTC
MdmdY       ATGGCTTTAAGGACCATCAATCGATTATTGAGTCATTAGAATTGGTAGAAGAATATGCTC
MdmdII      ATGGCTTTAAGGACCATCAATCGATTATTGAGTCATTAGAATTGGTAGAAGAATATGCTC
MdmdV       ATGGCTTTAAGGACCATCAATCGATTATTGAGTCATTAGAATTGGTAGAAGAATATGCTC
            **** ** ***** ********* * ** ***** ** **************  *** **
```

```
Md-ncm        AATTTACACATTTACTAATGTTGGATGATGCCACATCCCCAGAAGATGTACTGAGTAAGT
MdmdIII       AATTCACTCATTTGCTATTGTTGGAAGATGTCACATACCCGAAAGATATATTGAGTAAGT
MdmdY         AATTCACTCATTTGCTATTGTTGGAAGATGTCACATACCCGAAAGATATATTGAGTAAGT
MdmdII        AATTCACTCATTTGCTATTGTTGGAAGATGTCACATACCCGAAAGATATATTGAGTAAGT
MdmdV         AATTCACTCATTTGCTATTGTTGGAAGATGTCACATACCCGAAAGATATATTGAGTAAGT
              **** ** ***** *** ******* **** ***** ***  ***** ** *********
                                                INTRON
Md-ncm        TTCCATGCAATTTATAAAAAAAGGTGTTATTTATTGATAAATCTTTTTCCTTTTTAGATG
MdmdIII       TAATAGAAATTAAT-----AAATTATTTTTTTTGTCATTAAATTTGTTAATTTTTAGACG
MdmdY         TAATAGAAATTAAT-----AAATTATT--TTTTGTCATTAAATTTGTTAATTTTTAGACG
MdmdII        TAATAGAAATTAAT-----AAATTATT--TTTTGTCATTAAATTTGTTAATTTTTAGACG
MdmdV         TAATAGAAATTAAT-----AAATTATT--TTTTGTCATTAAATTTGTTAATTTTTAGACG
              *   *   * *        ***    *  ***  * ** **  ** **  ******** *

Md-ncm        CTTTTAAATTTGATGAACAATACGAGGCCAATGAAGAAAAATACAAAGGGCTTAGTAAAG
MdmdIII       AATTTAAATTCGACGACCAGTACGAGACCAATGAAGAGAAATATAAAGCACTTAGTAAGG
MdmdY         AATTTAAATTCGACGATCAGTACGAGACCAATGAAGAGAAATATAAAGCACTTAGTAAGA
MdmdII        AATTTAGATTCGACGATCAGTACGAGACCAATGAAGAGAAATATAAAGCACTTAGTAAGA
MdmdV         AATTTAAATTCGACGATCAGTACGAGACCAATGAAGAGAAATATAAAGCACTTAGTAAGA
                **** *** ** ** ** ****** ********** ***** ****  ********

Md-ncm        AGATTTTGGGTAGTGATGCCTCAGATTCGGATGGCTCCTCCGGTTCAGGTAGTGATTCCG
MdmdIII       ATATTTTGGGTAGCCATGCTTCAGATTCGGATGGCTCCTTCGGTTCGGGTAGTAATTCCG
MdmdY         ATATTTTGGGTAGCCATGCTTCAGATTCGGATGGCTCCTTCGGTTCGGGTAGTAATTCCG
MdmdII        ATATTTTGGGTAGCCATGCTTCAGATTCGGATGGCTCCTTCGGTTCGGGTAGTAATTCCG
MdmdV         ATATTTTGGGTAGCCATGCTTCAGATTCGGATGGCTCCTTCGGTTCGGGTAGTAATTCCG
              * ***********  **** ******************* ****** ****** ******

Md-ncm        ACTCTGAATCATCTGACTCGGATGAGGACAAAAATGAGGGGGATGAGAAACCCACAGCTG
MdmdIII       AAACTGCATTATCTGACTGTGATAAGGTCAAAAATGAGGTTAATGATAAATACACGAGTG
MdmdY         AAACTGCATTATCTGACTGTGATAAGGGCAAAAATGAGGTTAATGATAAATACACGAGTG
MdmdII        AAACTGCATTATCTGACTGTGATAAGGGCAAAAATGAGGTTAATGATAAATACACGAGTG
MdmdV         AAACTGCATTATCTGACTGTGATAAGGGCAAAAATGAGGTTAATGATAAATACACGAGTG
              *  *** ** ********  *** *** ***********   **** ***  ***   **

Md-ncm        GCGATATTATTGATAATACGGAAACCAATCTGATTGCGTTAAGGCGAACTATATATCTGA
MdmdIII       GTGACATTATTGATGAAACGAAACCAAATCTGATTGCGTTAAGAAGAACAATATATCTTA
MdmdY         GTGACATTATTGATGAAACGAAACCAAATCTGATTGCGTTAAGAAGAACAATATATCTTA
MdmdII        GTGACATTATTGATGAAACGAAACCAAATCTGATTGCGTTAAGAAGAACAATATATCTTA
MdmdV         GTGACATTATTGATGAAACGAAACCAAATCTGATTGCGTTAAGAAGAACAATATATCTTA
              * ** ********* * *** ** * *****************  **** ******** *

Md-ncm        CCATTAACTCCAGTTTGGATTATGAGGAGTGCGCCCATAAGCTTATGAAAATGCAATTGA
MdmdIII       CCTTAAATTCCTGTTTGGATTATGAGGAATGTGCCCAAAAATTAATGAAAACGCAATTGA
MdmdY         CCTTAAATTCCTGTTTGGATTATGAGGAATGTGCCCAAAAATTAATGAAAATGCAATTGA
MdmdII        CCTTAAATTCCTGTTTGGATTATGAGGAATGTGCCCAAAAATTAATGAAAATGCAATTGA
MdmdV         CCTTAAATTCCTGTTTGGATTATGAGGAATGTGCCCAAAAATTAATGAAAATGCAATTGA
              ** * ** *** **************** ** ***** **  * ******* ********

Md-ncm        AACCCGGCCAAGAAATAGAGCTTTGCCATATGTTTTTAGATTGCTGCGCTGAACAACGGA
MdmdIII       AAACTTGTCAACAAAATGAGTTTTGCCAAATATTGTTAGATTGCTGCGCTGAACAAAGAA
MdmdY         AAACTTGTCAACAAAATGAGTTTTGCCAAATATTGTTAGATTGCTGCGCTGAACAAAGAA
MdmdII        AAACTTGTCAACAAAATGAGTTTTGCCAAATATTGTTAGATTGCTGCGCTGAACAAAGAA
MdmdV         AAACTTGTCAACAAAATGAGTTTTGCCAAATATTGTTAGATTGCTGCGCTGAACAAAGAA
              ** *  * *** ***  *** ******* ** ** ********************* * *

Md-ncm        CTTATGAGAAGTTCTATGGTCTGTTGGCTCAGAGATTTTGTAACATCAACAAATCGTATA
MdmdIII       CCTATGAGAAGTTCTATGGCCTTTTAACTCACCGAATTTGTAAAATGAACAAGTCTTTTA
MdmdY         CCTATGAGAAGTTCTATGGCCTTTTAACTCACCGAATTTGTAAAATGAACAAGTCTTTTA
MdmdII        CCTATGAGAAGTTCTATGGCCTTTTAACTCACCGAATTTGTAAAATGAACAAGTCTTTTA
MdmdV         CCTATGAGAAGTTCTATGGCCTTTTAACTCACCGAATTTGTAAAATGAACAAGTCTTTTA
              * ***************** ** **  ****  ** ******* ** ***** ** * **

Md-ncm        TAGAACCGTTTGAGGAAATCTTCAAGGATACCTATCAGACCACACATCGTTTGGATACAA
MdmdIII       TAGAACCATTCAAAGAAATCTTCAAGGATATCTGTCAGACTACGCATTGTTTAGATACAA
MdmdY         TAGAACCATTCAAAGAAATCTTCAAGGATATCTGTCAGACTACGCATTGTTTAGATACAA
MdmdII        TAGAACCATTCAAAGAAATCTTCAAGGATATCTGTCAGACTACGCATTGTTTAGATACAA
MdmdV         TAGAACCATTCAAAGAAATCTTCAAGGATATCTGTCAGACTACGCATTGTTTAGATACAA
              ******* **  * **************** ** ****** ** *** **** *******
```

```
Md-ncm      ATCGTTTGCGGAATGTAAGCAAATTCTTTGCTCATTTGTTGTTTACCGATGCCATAAGTT
MdmdIII     ATCGTTTGCGGAATATAAGCAAATTCTTTGCTCATTTGTTGTTTACCGATGCAATTAGTT
MdmdY       ATCGTTTGCGGAATATAAGCAAATTCTTTGCTCATTTGTTGTTTACCGATGCAATTAGTT
MdmdII      ATCGTTTGCGGAATATAAGCAAATTCTTTGCTCATTTGTTGTTTACCGATGCAATTAGTT
MdmdV       ATCGTTTGCGGAATATAAGCAAATTCTTTGCTCATTTGTTGTTTACCGATGCAATTAGTT
            ************** ************************************ ** ****

Md-ncm      GGGATGTTTTGGACTGCATAAAGCTGAATGAAGACGACACAACATCATCAAGTCGTATTT
MdmdIII     GGGATGTTTTAGACTGCATAAAGCTAACTGAAGATGAGGCAATAACATCACGTTGTATTT
MdmdY       GGGATGTTTTAGACTGCATAAAGCTAACTGAAGATGAGGCAATAACATCACGTTGTATTT
MdmdII      GGGATGTTTTAGACTGCATAAAGCTAACTGAAGATGAGGCAATAACATCACGTTGTATTT
MdmdV       GGGATGTTTTAGACTGCATAAAGCTAACTGAAGATGAGGCAATAACATCACGTTGTATTT
            ********** ************** * ****** **  *** * ***** ** ******

Md-ncm      TCATAAAAATTCTTTTTCAAGAGTTGGCCGAGTACATGGGTCTGGGCCAATTGAATAAAA
MdmdIII     TTATAAAAAGTTTCTTTCAAGAATTGGTCGAATACATGGGTCTGTATCACTTTAATAAAA
MdmdY       TTATAAAAAGTTTCTTTCAAGAATTGGTCGAATACATGGGTCTGTATCACTTTAATAAAA
MdmdII      TTATAAAAAGTTTCTTTCAAGAATTGGTCGAATACATGGGTCTGTATCACTTTAATAAAA
MdmdV       TTATAAAAAGTTTCTTTCAAGAATTGGTCGAATACATGGGTCTGTATCACTTTAATAAAA
            * ******* * * ******** **** *** ************   ** ** *******

Md-ncm      AACTTAAGGATGAAGTTTTAGCAGAGAGTTTGGCCGGGCTATTTCCCAAAGATAACCCCA
MdmdIII     AACTTAAGACTGAAGTTTTAGCTGGAACTTTGGCGGGACTATTCCCCAAAGATAATCCCA
MdmdY       AACTTAAGACTGAAGTTTTAGCTGGAACTTTGGCGGGACTATTCCCCAAAGATAATCCCA
MdmdII      AACTTAAGACTGAAGTTTTAGCTGGAACTTTGGCGGGACTATTCCCCAAAGATAATCCCA
MdmdV       AACTTAAGACTGAAGTTTTAGCTGGAACTTTGGCGGGACTATTCCCCAAAGATAATCCCA
            ********  ************ *  * ****** ** ***** *********** ****

Md-ncm      GAAATACCCGATTTTCCATTAACTTTTTCACATCCATTGGTTTGGGTGGCCTAACTGATG
MdmdIII     GAAATATACGCTTCTCCATTAACTTTTTCACATCTATTGGTTTGGGCGGAATAACTAATG
MdmdY       GAAATATACGCTTCTCCATTAACTTTTTCACATCTATTGGTTTGGGCGGAATAACTAATG
MdmdII      GAAATATACGCTTCTCCATTAACTTTTTCACATCTATTGGTTTGGGCGGAATAACTAATG
MdmdV       GAAATATACGCTTCTCCATTAACTTTTTCACATCTATTGGTTTGGGCGGAATAACTAATG
            ******  ** ** ******************** *********** **  ***** ***

Md-ncm      AATTGCGTCAATTCCTAAAGAATGCCCCGAAATCTGTACCCGCCATAAATGCTGAAATTC
MdmdIII     AATTGTGTCAACTCCTAAAGATTGCTCCGAAATCTGCACCCTCGTCCTCATCATCAAGTT
MdmdY       AATTGTGTCAACTCCTAAAGATTGCTCCGAAATCTGCACCCTCGTCCTCATCATCAAGTT
MdmdII      AATTGTGTCAACTCCTAAAGATTGCTCCGAAATCTGCACCCTCGTCCTCATCATCAAGTT
MdmdV       AATTGTGTCAACTCCTAAAGATTGCTCCGAAATCTGCACCCTCGTCCTCATCATCAAGTT
            ***** ***** ********* *** ********** **** *        *   ** * 

Md-ncm      TGGCCAATAAACCTGTAGATGTCTCAAATTCGTCGTC-GTCCTCCTCGTCATC------A
MdmdIII     CTTTATCATCGGAATTGTCTGCACC--ATCCGACGACGATTCTTCAAGTGATTCGGAAAA
MdmdY       CTTTATCATCGGAATTGTCTGCACC--ATCCGACGACGATTCTTCAAGTGATTCGGAAAA
MdmdII      CTTTATCATCGGAATTGTCTGCACC--ATCCGACGACGATTCTTCAAGTGATTCGGAAAA
MdmdV       CTTTATCATCGGAATTGTCTGCACC--ATCCGACGACGATTCTTCAAGTGATTCGGAAAA
                           *   **   *  ** ** ** *  * ** *  ** **       *

Md-ncm      TCATCTTCAACCTCGTCTTCATCATCATCAAGTTCTTCATCTTCGGAATCGTC-------
MdmdIII     TAAGAAAAAACATAAAGGCAAAAATAAGAAAATGACCAAGAAGAAAAATCCTTCGAAAAA
MdmdY       TAAGAAAAAACATAAAGGCAAAAATAAGAAAATGACCAAGAAGAAAAATCCTTCGAAAAA
MdmdII      TAAGAAAAAACATAAAGGCAAAAATAAGAAAATGACCAAGAAGAAAAATCCTTCGAAAAA
MdmdV       TAAGAAAAAACATAAAGGCAAAAATAAGAAAATGACCAAGAAGAAAAATCCTTCGAAAAA
            * *     *** *       *  ** *  ** *     *       **** *        

Md-ncm      ------------T-----------GCAT--------------------------------
MdmdIII     AAAGGAAAAAACTAAAAAATTTGTAGGTAAAAATAAAATAGCCGCTAAGAATAAAACTAT
MdmdY       AAAGGAAGAAACTAAAAAAATTGTAGGTAAAAATAAAATAGCCGCTAAGAATAAAACTAT
MdmdII      AAAGGAAAAAACTAAAAAAATTGTAGGTAAAAATAAAATAGCCGCTAAGAATAAAACTAT
MdmdV       AAAGGAAAAAACTAAAAAAATTGTAGGTAAAAATAAAATAGCCGCTAAGAATAAAACTAT
                        *              *                                
```

```
Md-ncm      -----------------------CATCATCTTCTAGT-----------------------
MdmdIII     AAAGCGAAGGACTGACAAAGACAACTCTTCTTCAAAAGATAATTTTTTGAAAAGCGAAAG
MdmdY       AAAGCGAAGGACTGACAAAGACAACTCTTCTTCAAAAGATAATTTTTTGAAAAGCGAAAG
MdmdII      AAAGCGAAGGACTGACAAAGACAACTCTTCTTCAAAAGATAATTTTTTGAAAAGCGAAAG
MdmdV       AAAGCGAAGGACTGACAAAGACAACTCTTCTTCAAAAGATAATTTTTTGAAAAGCGAAAG
                                     ** ***** *

Md-ncm      ------------------------------------------------------------
MdmdIII     TAGCAGCAACGAATCAATATCCTTAGATAGTTTATCTTCGGAATTGTTTGCACCATCGTC
MdmdY       TAGCAGCAACGAATCAATATCCTTAGATAGTTTATCTTCGGAATTGTTTGCACCATCGTC
MdmdII      TAGCAGCAACGAATCAATATCCTTAGATAGTTTATCTTCGGAATTGTTTGCACCATCGTC
MdmdV       TAGCAGCAACGAATCAATATCCTTAGATAGTTTATCTTCGGAATTGTTTGCACCATCGTC


Md-ncm      -GATTCGTCGTCGGATTCTTCAAGTGATTCAGAAGTGGATAAGAAAAAACGTAAAGGCAA
MdmdIII     TTATTCGTCGTCGGAATCGTCAAACGATTCAGAAAGTA---AGGAAAAACATAAAGGCAA
MdmdY       TTATTCGTCGTCGGAATCGTCAAACGATTCAGAAAGTA---AGGAAAAACATAAAGGCAA
MdmdII      TTATTCGTCGTCGGAATCGTCAAACGATTCAGAAAGTA---AGGAAAAACATAAAGGCAA
MdmdV       TTATTCGTCGTCGGAATCGTCAAACGATTCAGAAAGTA---AGGAAAAACATAAAGGCAA
              ************* ** ****  *********       ** ****** *********

Md-ncm      GGTTAA---GAAGACCAAGAAGAAGAAAACTTCAAATAAAAAGGAGAAAACTAAAAAATC
MdmdIII     AAATAAGAAGATGACCAAAAAGAAAAATCCTTCGAATAAAAAGGAAAAAACAAAAAAAAA
MdmdY       AAATAAGAAGATGACCAAAAAGAAAAATCCTTCGAATAAAAGGGAAAAAACAAAAAAAAA
MdmdII      AAATAAGAAGATGACCAAAAAGAAAAATCCTTCGAATAAAAGGGAAAAAACAAAAAAAAA
MdmdV       AAATAAGAAGATGACCAAAAAGAAAAATCCTTCGAATAAAAGGGAAAAAACAAAAAAAAA
               ***   ** ****** ***** **  **** ******* *** ***** ******

Md-ncm      GAAAAGTAAAAATAAAAAAGACGCCAAGAAAAAGGCCAAGAAGCGAAAGTCCAAAAAAGG
MdmdIII     ATTAAGTAAAAATAAAAAAGCCCCAAACAAAAATACTAAGA-------------------
MdmdY       ATTAAGTAAAAATAAAAAAGCCCCAAACAAAAATACTAAGA-------------------
MdmdII      ATTAAGTAAAAATAAAAAAGCCCCAAACAAAAATACTAAGA-------------------
MdmdV       ATTAAGTAAAAATAAAAAAGCCCCAAACAAAAATACTAAGA-------------------
               ***************** * * ** *****  * ****

Md-ncm      CAGCACTTCTTCAGAAGATGATTCTTCGAAAAGCGAAAGTAGTAGCAGTGAATCAAAATC
MdmdIII     --AGCGAATGACTGAAAAGGACATTTCTTCTTCTGAAAGTAGCATCAGTGAATCAAAATC
MdmdY       --AGCGAATGACTGAAAAGGACATTTCTTCTTCTGAAAGTAGCATCAGTGAATCAAAATC
MdmdII      --AGCGAATGACTGAAAAGGACATTTCTTCTTCTGAAAGTAGCATCAGTGAATCAAAATC
MdmdV       --AGCGAATGACTGAAAAGGACATTTCTTCTTCTGAAAGTAGCATCAGTGAATCAAAATC
                       * *** * **   ***       ******** * ***************

Md-ncm      TTCAGACAGTTCGGACTCGGAGCAAAGTGGAAATGATAAATCGAAGAAGAAGACTAAATC
MdmdIII     TTTAAATTGTTCGGCCTCGAACCAAAATGAGAATGAAAAACGGAAGAAAAGGGTTACATC
MdmdY       TTTAAATTGTTCGGCCTCGAACCAAAATGAGAATGAAAAACGGAAGAAAAGGGTTACATC
MdmdII      TTTAAATTGTTCGGCCTCGAACCAAAATGAGAACGAAAAACGGAAGAAAAGGGTTACATC
MdmdV       TTTAAATTGTTCGGCCTCGAACCAAAATGAGAATGAAAAACGGAAGAAAAGGGTTACATC
            ** * *  ****** **** * **** **  ** ** ***  ****** * *  ** ***

Md-ncm      GAAATCAAAAACAAAACCGAATAAGAAGTCCAAACGATCTGACTCAGAAGACGTTGAAAA
MdmdIII     GAAATCAAGAACAAAACGGGTAAAGATGTTTAAACAATGTCAATGGGTTGACGCGGACAA
MdmdY       GAAATCAAGAACAAAACGGGTAAAGATGTTTAAACAATGTCAATGGGTTGACGCGGACAA
MdmdII      GAAATCAAGAACAAAACGGGTAAAGATGTTTAAACAATGTCAATGGGTTGACGCGGACAA
MdmdV       GAAATCAAGAACAAAACGGGTAAAGATGTTTAAACAATGTCAATGGGTTGACGCGGACAA
            ******** ******** *   **** **  **** ** * * *  *  ****  ** **

Md-ncm      TGAACGAGATATTAAACGTAAGAAAAGGGAAGATTTCGGAAATTACATTCATGAAGATAG
MdmdIII     TCAACGAGATATTAAACGCAAGAAAAGGGCAGAATATAGATATGAACCTCTTGTATATAG
MdmdY       TCAACGAGATATTAAACGCAAGAAAAGGGCAGAATATAGATATGAACCTCTTGTATATAG
MdmdII      TCAACGAGATATTAAACGCAAGAAAAGGGCAGAATATAGATATGAACCTCTTGTATATAG
MdmdV       TCAACGAGATATTAAACGCAAGAAAAGGGCAGAATATAGATATGAACCTCTTGTATATAG
            * **************** ********** *** *   ** ** *   ** ** * ****

Md-ncm      AAGAAGAGACGAGGAATATTCAAAGAATGGTGGAAGAGATCGCAAAAAGGATAATCACGG
MdmdIII     AAAAAGAAATGAGGAATGTTTAAAGAAGGGCGCACCAAATTGCAGSAAAGATAACTATGG
MdmdY       AAAAAGAAATGAGGAATATTTAAAGAAGGGCGGACCAAATTGCAGGAAAGATAACTATGG
MdmdII      AAAAAGAAATGAGGAATGTTTAAAGAAGGGCGCACCAAATTGTAGGAAAGATAACTATGG
MdmdV       AAAAAGAAATGAGGAATGTTTAAAGAAGGGCGCACCAAATTGCAGGAAAGATAACTATGG
            ** **** * ******* ** ****** ** * *  * ** * *  ** *****  * **
```

```
Md-ncm      TAATAATAAGGAGAAACGGGAAATACCACAACGTCGCGATAGCGAAATTGAAAGGCGTCG
MdmdIII     ---TAATAGGCAGAATCATGAAATATCACAACGTCATGATAGTGAAATTAAAAGACGCCG
MdmdY       ---TAATAGGCAGAATCATGAAATATCACAACGTCATGATAGTGAAATTAAAAGACGCCG
MdmdII      ---TAATAGGCAGAATCATGAAATATCACAACGTCATGATAGTGAAATTAAAAGACGCCG
MdmdV       ---TAATAGGCAGAATCATGAAATATCACAACGTCATGATAGTGAAATTAAAAGACGCCG
               ***** * **** *  ****** *********  ***** ****** **** ** **

Md-ncm      AGAGGAGCGCGAAAAACGCCATCGTGAAAGAGAAAGAAATTTCTCACGTTCCAGATCAAG
MdmdIII     GGAAGAGCGAAAAAAACGCCACCATGAAAAAAATCACTCACGTGAATATAAACGTTCTAA
MdmdY       GGAAGAGCGAAAAAAACGCCACCATGAAAAAAATCACTCACGTGAATATAAACGTTCTAA
MdmdII      GGAAGAGCGAAAAAAACGCCACCATGAAAAAAATCACTCACGTGAATATAAACGTTCTAA
MdmdV       GGAAGAGCGAAAAAAACGCCACCATGAAAAAAATCACTCACGTGAATATAAACGTTCTAA
             ** *****  ********** * ***** * *            *  *    * ** *

Md-ncm      ATCTC-------------------------------------------------------
MdmdIII     ATTAGGGCTCTGCCAGAGGGAATATTTTTTATATATGTGCTGTCAGTTTTATTATCCATG
MdmdY       ATTAGGGCCCTGCCAGAGGGAATATTTTTTATATATGTGCTGTCAGTTTTATTATCCATG
MdmdII      ATTAGGGCTCTGCCAGAGGGAATATTTTTTATATATGTGCTGTCAGTTTTATTATCCATG
MdmdV       ATTAGGGCTCTGCCAGAGGGAATATTTTTTATATATGTGCTGTCAGTTTTATTATCCATG
            **

Md-ncm      ------------------------------------------------------------
MdmdIII     TACATTTCAATGTTTATGTCAAAATTGTCATTTTACATTCTACTCCTAATTCTTGCTATC
MdmdY       TACATTTCAATGTTTATGTCAAAATTGTCATTTTACATTCTACTCCTAATTCTTGCTATC
MdmdII      TACATTTCAATGTTTATGTCAAAATTGTCATTTTACATTCTACTCCTAATTCTTGCTATC
MdmdV       TACATTTCAATGTTTATGTCAAAATTGTCATTTTACATTCTACTCCTAATTCTTGCTATC


Md-ncm      ----------------
MdmdIII     AAACACTTACATATGT
MdmdY       AAACAC----------
MdmdII      AAACAC----------
MdmdV       AAACAC----------
```

Chapter 4

Functional analysis of *Mdmd* in *M. domestica* by embryonic RNAi

This chapter is published as part of:
Sharma, A., Heinze, S.D., Wu, Y., Kohlbrenner, T., Morilla, I., Brunner, C., Wimmer, E.A., van de Zande, L., Robinson, M.D., Beukeboom, L.W., Bopp, D. (2017). Male sex in houseflies is determined by *Mdmd*, a paralog of the generic splice factor gene *CWC22*. *Science* ***356***, 642–645.

ABSTRACT

Insect sex determination pathways consist of a hierarchical cascade of genes, instructed by a bewildering diversity of primary signals. Houseflies have a unique polymorphic sex determination system in which the primary signal is a dominant male-determining factor, *M*, that either resides on the Y chromosome or on different autosomes. In the previous chapters, we described *Mdmd* as a possible candidate for male development in *Musca domestica*. Here, we report on embryonic RNAi (eRNAi) experiments to validate the male-determining role of *Mdmd*. Transient silencing of *Mdmd* in embryos leads to partial feminization in males of the M^{II}, M^{III}, M^{V} and M^{Y} strains consistent with a male-determining role of *Mdmd*. Embryonic silencing of its paralog *Md-ncm* causes lethality in both male and female embryos consistent with its predicted role in general splicing regulation. Our study reveals that duplication and neofunctionalization of an essential splicing regulating gene, *Md-ncm*, can adopt the role of the instructive signal in the sex determination hierarchy.

INTRODUCTION

Generation of males and females is a common feature of unisexual organisms as part of their reproduction (West and Sheldon, 2002). Genetic mechanisms for the determination of the two sexes are however not conserved among organismal groups. Insects illustrate this diversity; sex determination systems vary at the chromosomal, gene and gene regulation level between species (Pease and Hahn, 2012; Bachtrog *et al*., 2014; Vicoso and Bachtrog, 2015). For example, some groups have male heterogamety (XX-XY system) in which sex can be determined by a dominant male-determining gene on the Y chromosome or by X chromosome dose, as in the case of *Drosophila* (Erickson and Quintero, 2007). Other groups are characterized by female heterogamety or haplodiploidy, or yet other more specific mechanisms (Normark, 2003; Ellegren, 2011; Beukeboom and Perrin, 2014; Bachtrog *et al*., 2014; Blackmon *et al.*, 2015).

The general pattern of insect sex determination consists of a primary genetic signal (instruction) that is distinct in males and females and acts via a binary key genetic switch (transduction) that responds to global effectors (execution) to achieve the selected genetic program (Bopp, Saccone, and Beye, 2014). The insect sex determination pathway is conserved at the level of the key transductory axis, formed by the *transformer* (*tra*) and *doublesex* (*dsx*) genes, but highly diverse at the level of the upstream instructive signals (Sánchez, 2008; Verhulst *et al.*, 2010). The polymorphic sex determination system of the housefly, *M. domestica*, reflects this diversity in regulation and genes (Franco *et al.*, 1982; Dübendorfer *et al.*, 2002; Hediger *et al.*, 2010; Hamm *et al.*, 2015).

In the previous chapters, we identified a paralog of splicing factor *CWC22* as a putative male determiner in *M. domestica*, which we named *Mdmd* (for *Musca domestica male determiner*). We isolated *Mdmd* from different *M*-carrying strains (M^{II}, M^{III}, M^{V} and M^{Y}) and found, by sequence comparison, these strains to carry the same *Mdmd*. We also identified the paralog of the male-specific *Mdmd* in the female genome of *M. domestica*. This paralog is known as *nucampholin* by homology to *Drosophila* (CWC22 in mammals), so we named this *Md-ncm* (*Musca domestica-nucampholin*). *Md-ncm* is present in both males and females.

A functional analysis is needed to verify further that the male-specific *Mdmd* is indeed the gene for determining male sex in the housefly. Here, we perform functional tests with embryonic RNAi in an M^{I}, M^{II}, M^{III}, M^{V} and M^{Y} strain of *M. domestica*. We investigate whether suppression of *Mdmd* by embryonic RNAi results in expression of female-specific splicing of its downstream target gene *Md-tra* and conversion of males to females. We also test whether silencing of *Mdmd* by eRNAi affects the expression of its paralog *Md-ncm*.

MATERIALS AND METHODS

Musca domestica strains and culturing

All housefly strains and culture methods used in this study are described in chapters 2 and 3.

Embryonic RNAi of *Mdmd* and *Md-ncm*

Two dsRNAs templates, 33840_c0 and 22793_c0, were designed based on the predicted *Mdmd* sequence coding for the amino-terminal coding region and carboxy-terminal coding region of the MDMD protein, respectively, and synthesized by *in vitro* transcription of DNA templates comprising the T7 promoter sequences on both ends. Total RNA was extracted from a single M^{I}, M^{II}, M^{III}, M^{V} and M^{Y} male with the NucleoSpin RNAII Kit (Machery-Nagel) and cDNA was isolated with the Transcription High Fidelity cDNA Synthesis Kit (Roche). PCR was performed under the following conditions: an initial denaturation step at 94°C for 2 min was followed by 35 amplification cycles (denaturation at 92°C for 30 s, annealing at 59°C for 30 s and elongation at 72°C for 30 s) and a final elongation step at 72°C for 3 min. The 25 µl reaction mixtures contained 10 µl 5×GoTaq reaction Buffer, GoTaq reaction Buffer (Promega), 10 pmol of each primer, 10 mM dNTP mix and 25 mM MgCl2. PCR products were purified (JETquick, PCR Product Purification Spin Kit/ 250) and 1 µg of each was used as a template for *in vitro* transcription of dsRNAs (MEGAscript RNAiKit, High Yield Transcription Kit do dsRNA, Ambion).

Embryos were collected one hour after fertilization and injected with dsRNA 33840_c0 and 22793_c0 separately and in combination at 1 µg/µl and 500 ng/µl concentration. dsRNA M122, a male-specific sequence unrelated to *Mdmd,* was used as a control for unspecific RNAi effects. M122 eRNAi has no effects in both males and females. After microinjection embryos were transferred to a sieve, dechorionated with 3% NaOCl (Sodium hypochlorite solution) for 3 min under a compound microscope after rinsing with tap water, and placed in Ringer's solution (7.2 g sodium chloride, 0.37 g potassium chloride and 0.17 g calcium chloride dissolved in one litre of distilled water). Embryos were aligned on a coverslip in vertical rows with the posterior end of the embryo at 0.5 mm from the coverslip border under a compound microscope. Coverslips were placed in a plastic container with silica gel beads for 5 min for dehydration and covered with 10S Voltalef Prolabo oil. dsRNAs solution was injected with a glass capillary and a micromanipulator

(Femtojet-Eppendorf) under a dissecting microscope. After injection, coverslips with embryos were incubated on an apple-juice agar plate at 25°C for 24 h. Surviving embryos were collected and placed into a beaker filled with pig dung at 25°C until hatching. For silencing of *Md-ncm*, we designed dsRNAs template Mdncm5_7 at the 5' end of *Md-ncm* at a similar position as for *Mdmd*. Synthesis of dsRNA solution and procedure for microinjections were same as described above.

dsRNA injected flies were dissected under a microscope 3 to 5 days after hatching. The gonads were placed into the well of a microtiter plate filled with phosphate buffered-solution (PBS, pH 7.4), fixed by 3.6% paraformaldehyde for 30 min at RT, rinsed with PBST (PBS+ 0.1% Triton X-100) and permeabilized in PBS containing 1% Triton X-100 and 10 mg/ml BSA (Bovine serum albumin) for 1 h at RT. The samples were then incubated in a 1:500 dilution of DAPI (4',6-diamidino-2-phenylindole, nuclear counterstain) overnight in the dark and transferred onto a slide containing 50% glycerol. The tissues were observed under fluorescence microscopy and stored at 4°C.

cDNA amplifications

To check the expression of *Mdmd*, *Md-tra*^{F} and *Md-ncm* in *Mdmd* dsRNA (33840_c0 and 22793_c0) injected embryos of different *M* strains; we conducted the following procedure; total RNA was extracted and cDNA prepared from single flies as described in chapter 2. For cDNA amplifications primer pairs Mdmd_7s/Mdmd_8as were used for *Mdmd*, and Md-ncm_1s/ Md-ncm_2as and Md-ncm_9s/Md-ncm_10as for *Md-ncm*. PCR was performed in 50 µl volume containing 10 µl 5x GoTaq Reaction Buffer, 1.5 µl 25 mM $MgCl_2$, 1.5 µl 10 mM of each dNTP, 0.3 µl GoTaq DNA Polymerase (Promega), 1.5 µl 10 µM of each primer and 2 µl genomic DNA. PCR thermal cycling consisted of 2 min initial denaturation at 94°C, followed by 35 cycles of 30 s at 92°C, 30 s at 59°C and 30 s at 72°C, and a final elongation of 3 min at 72°C. For expression analysis of *Md-tra*, amplifications were performed on the cDNA with primers Md-tra9/Md-tra20 for *Md-tra*^{F} transcripts. Cytochrome P450 (Cyp) primer pair, CYP6D3-1/ CYP6D3-2, was used as an internal standard.

Quantitative *Mdmd* expression studies

Quantitative real-time PCR measured the relative expression levels of *Mdmd* in 10h and 20h old male embryos collected from homozygous M^{III} males crossed to homozygous *pw, bwb* females. RNA isolation was performed with the NucleoSpin RNAII Kit (Machery-Nagel) and was subsequently reverse transcribed with the Transcription High Fidelity cDNA Synthesis Kit (Roche). Subsequent qPCR was done with a 1:3 cDNA dilution and PerfeCTaTM SYBR®Green (300 nM) mix (Quanta Biosciences, Gaithersburg, USA) on an

Applied Biosystems 7300 Real-Time PCR System with 250 nM *Mdmd* qPCR primers 22793_c0_qF1 and 22793_c0_qR1. β-Actin was used as internal control for relative gene quantification with 250 nM of primers β-Actin _F1 and β-Actin_R1. All primers sets were developed with Primer3Plus. The qPCR profile was as follows: 95°C for 3 min, 40 cycles of 95°C for 15 s, 56°C for 30 s and 72°C for 30 s followed by a standard ABI7300 dissociation curve to control for nonspecific amplification.

LinRegPCR 11.0 (Ramakers *et al.*, 2003) was used for base-line correction and calculation of N0 value with PCR efficiencies per amplicon from raw fluorescence data generated by 7300 System SDS Software (Applied Biosystems, CA, USA). Relative expression levels of *Mdmd* in *M112*-dsRNA injected, and *Mdmd*-dsRNA injected male embryos were determined by dividing *Mdmd* N0 values by β-Actin N0. Three biological samples were used and three technical replicates from each of the 10h control, 10h *Mdmd*-dsRNA and 20h *Mdmd*-dsRNA samples. Two biological samples and three technical replicates from each were used for the 20h control samples injected with *M112*-dsRNA. The average of the relative expression (RE) values and their standard errors were calculated from *Mdmd* mRNA levels, 10h and 20h after injections. *Mdmd* expression differences between the *Mdmd*-dsRNA injected and control *M112*-dsRNA samples were compared with the Mann-Whitney U test.

RESULTS

Embryonic silencing of *Mdmd* in the M^{III} strain causes sex transformation of the gonads

We designed two dsRNAs templates, 33840_c0 and 22793_c0, in the amino-terminal coding and carboxy-terminal coding region of *Mdmd*, respectively. Both dsRNA templates are unique to *Mdmd* and not present in *Md-ncm* to avoid any cross-hybridization effects (Fig. 4.1).

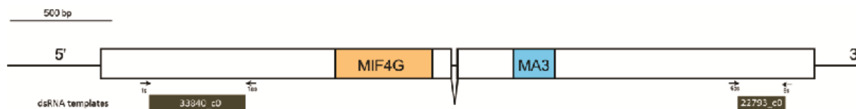

Figure 4.1: Design of *Mdmd* dsRNA from for eRNAi. Two cDNA fragments of *Mdmd,* 33840_c0, and 22793_c0 are used as templates for in vitro transcription for dsRNAs preparation and amplified with the primer pairs 1s/1as and 6s/6as. They correspond to two of the 5 fragments which were in the initial screen identified as male-specific.

For the first RNAi experiments, we used the M^{III} strain in which *M* is present on the third autosome and linked with dominant phenotypic markers pw^+ and bwb^+. We injected *Mdmd* dsRNA fragments 33840_c0 and 22793_c0 separately or together in 1 h old embryos. As a control, we injected dsRNA of a male-specific sequence M122 which is unrelated to *Mdmd* and did not have any effect on male and female development. All of the injected *Mdmd* individuals developed externally as normal males, but upon dissection appeared to have fully differentiated ovaries instead of testes (Fig. 4.2). Both the single and combined injections of 33840_c0 and 22793_c0 dsRNAs in *M*-carrying individuals induced differentiation of female gonads (ovaries) at a high percentage (Table 4.1). Thus, RNAi silencing of *Mdmd* causes partially feminized males which implies that *Mdmd* is required for differentiation of male gonads in the M^{III} strain.

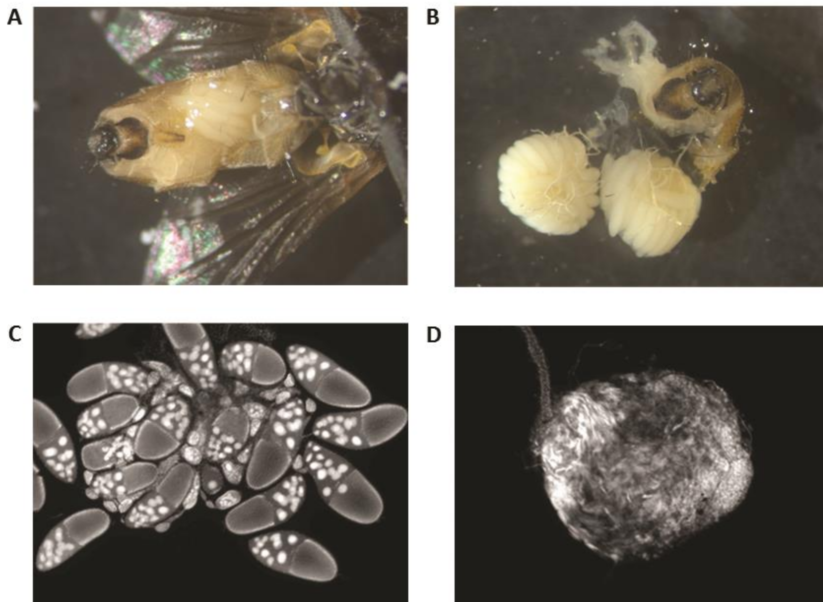

Figure 4.2: Embryonic silencing of *Mdmd* is transient and leads to ovarian differentiation in M^{III} males. M^{III}/+ individuals injected with dsRNA against *Mdmd*. (A) Adult male abdomen with fully differentiated eggs. (B) Dissected ovaries from the same male. (C) Ovaries are containing normal cysts composed of nurse cells and egg chambers. (D) M^{III}/+ male with normal testis. All pictures are DAPI stained.

Table 4.1: Results of injection of dsRNA 33840_c0 and 22793_c0 separately and in combination in the M^{III} strain. 56 to 83% of M^{III}/+ males had ovaries, +/+ females always have normal ovaries. Injection of dsRNA of *M122* as control yielded only males and females with normal testes and ovaries respectively. Total numbers and percentage (in brackets) are indicated.

strain	dsRNA	concentration	M^{III}/+ with testes	M^{III}/+ with ovaries	+/+ with ovaries
M^{III}	33840_c0	1 µg/µl	68 (0.32)	147 (0.68)	180
M^{III}	33840_c0	2 µg/µl	12 (0.44)	15 (0.56)	36
M^{III}	22793_c0	1 µg/µl	54 (0.33)	106 (0.67)	170
M^{III}	33840_c0+22793_c0	1 µg/µl	7 (0.17)	34 (0.83)	33
M^{III}	M122	1 µg/µl	48	0	45

Expression levels of *Mdmd* after eRNAi in male-only embryos by real time-qPCR

To examine the RNAi effects on *Mdmd* expression, we measured the relative transcript levels of *Mdmd* 10h and 20h after dsRNA injections. To this end, we injected *Mdmd* dsRNA templates 33840_c0 and 22793_c0 in 1h old all male embryos that were collected from a cross between homozygous M^{III} males and homozygous *pw, bwb* females. A significant 70% reduction of *Mdmd* transcript levels is observed in M^{III}/+ embryos 10h after dsRNA injection, while 20h post-injection *Mdmd* expression levels have returned to that of the control injected embryos (Fig. 4.3). This result indicates that the observed partial feminization in males after eRNAi (Fig. 4.2 and Fig. 4.5) is due to a transient reduction in the expression of *Mdmd*, and that expression is restored after 20h to a comparable level to control samples.

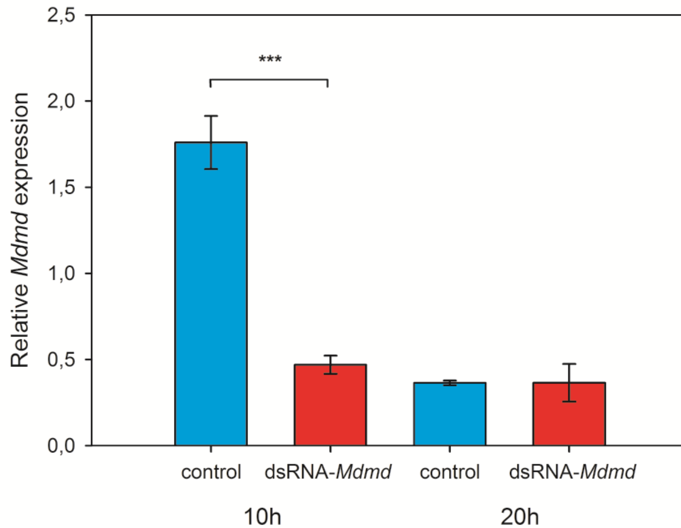

Figure 4.3: Relative levels of *Mdmd* mRNA 10h and 20h after injection with dsRNA against *Mdmd*. *Mdmd* expression is reduced after dsRNA injection in 10h old M^{III}/+ male embryos, but not in 20h old embryos compared to the *M112* control.

Expression analysis of *Mdmd* and *Md-tra*^{F} after eRNAi in the M^{III} strain

To investigate the effects of embryonic silencing of *Mdmd* on splicing of *Md-tra* in adult individuals of the M^{III} strain, we measured relative levels of *Md-tra* male and female splice variants in gonads and in the gonadectomized body. *Md-tra*^{F} transcripts are present in the ovaries of adult females and males that were injected with *Mdmd* dsRNA at early embryonic stage. In addition, substantial levels of *Mdmd* transcripts were detected in non-gonadal tissues of male adults with ovaries suggesting that restored activity of *Mdmd* at late stages prevented systemic female differentiation (Fig. 4.4).

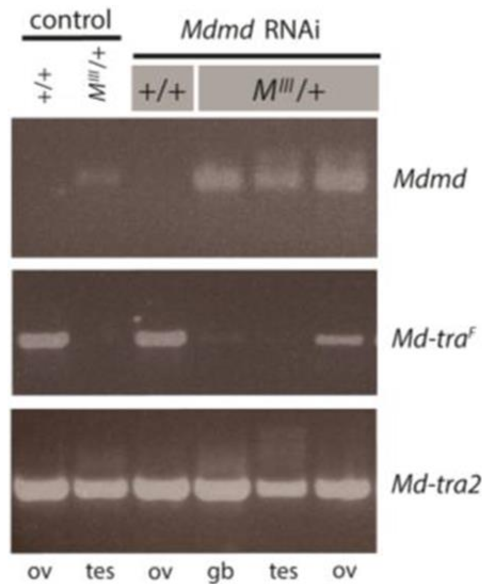

Figure 4.4: RT-PCR analysis of *Mdmd* and *Md-tra*^{F} in eRNAi treated individuals of the M^{III} strain. *Mdmd* transcripts and female transcripts of *Md-tra* (*Md-tra*^{F}) in control wild-type (+/+) ovaries (ov) and wild-type (M^{III}/+) testes (tes). *Mdmd* transcripts can be detected in *Mdmd* dsRNA-injected (M^{III}/+) gonadectomized bodies (gb), testes and ovaries. *Md-tra*^{F} transcripts are only present in *Mdmd* dsRNA-injected (+/+) and (M^{III}/+) ovaries. Amplification with *Md-tra2* primers served as a control for input levels of cDNA.

Embryonic RNAi of *Mdmd* in other *M* strains

dsRNAs 33840_c0 and 22793_c0 were injected separately or together in syncytial embryos of the M^I, M^{II}, M^V and M^Y strains. In all cases, except M^I, the majority of surviving *M/+* individuals are phenotypically normal males with differentiated ovaries (Fig. 4.5) similar to the results of the M^{III} strain (Fig. 4.2). Injections of both dsRNAs resulted in a higher penetrance of the male ovary phenotype. As expected no-*M* females are not affected by *Mdmd* silencing, they have normally differentiated ovaries and are fertile.

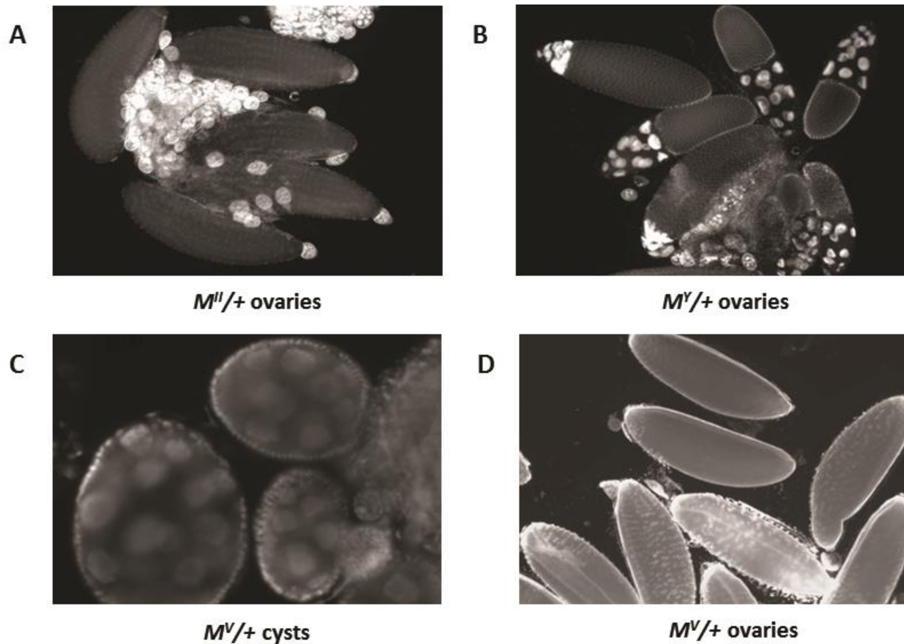

Figure 4.5: Embryonic silencing of *Mdmd* in M^{II}, M^V and M^Y males. DAPI (4',6-diamidino-2-phenylindole) stained ovaries are containing normal cysts composed of nurse cells and egg chambers. (A) Normal eggs in males of M^{II} strain. (B) Fully differentiated eggs in males of M^Y strain. (C) Differentiated cysts in males of M^V strain. (D) Fully developed eggs in males of M^V strain. All pictures were taken at 5x magnification.

Table 4.2. Embryonic RNAi in M^{II}, M^{V}, and M^{Y} males led to a high proportion of males with ovaries, with the exception of M^{I} strain where males had normal differentiated testes. Total numbers and percentages (brackets) are indicated.

strain	dsRNA	concentration	*M*/+ with testes	*M*/+ with ovaries	+ /+ with ovaries
M^{II}	22793_c0	0.5 µg/µl	3 (0.15)	17 (0.85)	14
M^{V}	33840_c0+22793_c0	1 µg/µl	4 (0.26)	11 (0.73)	5
M^{Y}	33840_c0+22793_c0	1.2 µg/µl	5 (0.23)	16 (0.76)	12
M^{I}	33840_c0+22793_c0	1 µg/µl	20	0	26

In conclusion, silencing of *Mdmd* by injecting dsRNA in syncytial embryos of different *M* strains (M^{II}, M^{III}, M^{V} and M^{Y}) resulted in flies that developed externally as males, but the majority of these males contained fully differentiated ovaries instead of testes, except for M^{I} males that all developed into normal fertile males. From this result, we infer that *Mdmd* is essential for specifying the male gonadal fate in these strains, except for M^{I}. Incomplete feminization may be explained by the transient nature of embryonic RNAi.

Expression of *Mdmd, Md-tra*^{F} and *Md-ncm* after eRNAi in other strains

To test the knock down effect of *Mdmd*, we measured the expression of mRNAs of *Mdmd*, *Md-tra*^{F} and *Md-ncm* in various tissues of injected adult flies. *Md-tra*^{F} is expressed in the ovaries and body of the injected female (+/+) and male (M^{II}/+). *Md-ncm* is expressed in both male (M^{II}/+) and female (+/+) of all injected adult flies. The two bands amplified from *Md-ncm* cDNA are likely due to the presence of processed and unprocessed RNA (or genomic DNA) as the difference in size corresponds to the size of the intron. *Md-tra2*, used here as a control for input levels of cDNA, is expressed in all tissues of both males and females of the M^{III} strain. The results are similar for the M^{II} (Fig. 4.6), M^{III} (Fig. 4.6) and M^{Y} (Fig. 4.8) strain.

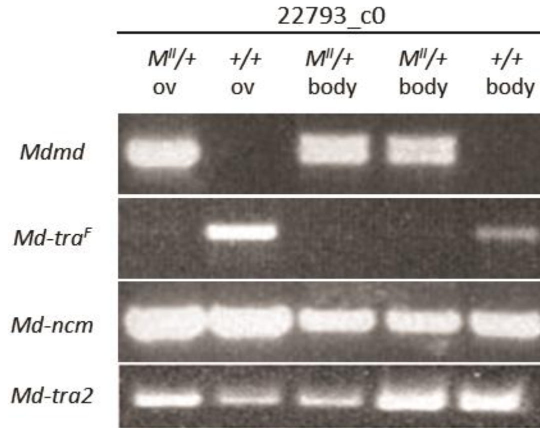

Figure 4.6: Expression of *Mdmd*, *Md-tra*^{F} and *Md-ncm* in adults of M^{II} strain after injection of dsRNA of template 22793_c0 at early embryonic stages. Different from the results obtained with injected M^{III}/+ animals, *Md-tra*^{F} is not present in the ovaries or other tissues of ¨pregnant¨ males.

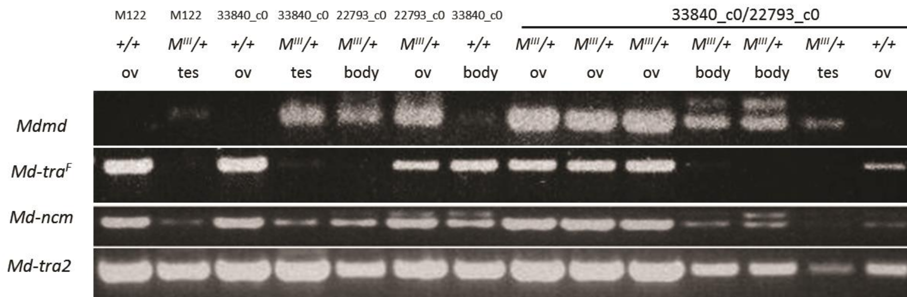

Figure 4.7: Expression of *Mdmd*, *Md-tra*^{F} and *Md-ncm* in adults of M^{III} strain after injection of dsRNA at early embryonic stages. dsRNA templates 33840_c0 and 22793_c0, solely and in combination injected in 1h old embryos of the M^{III} strain.

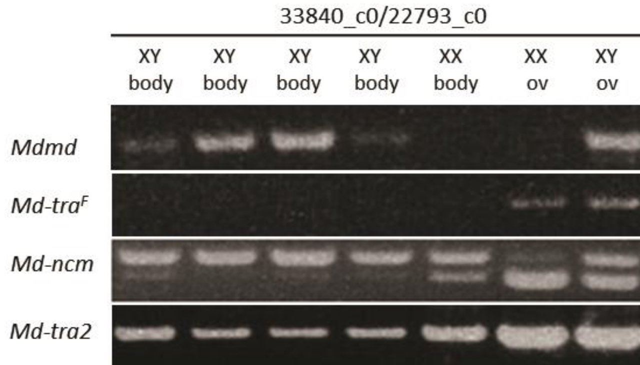

Figure 4.8: Expression of *Mdmd*, *Md-tra*^{F} and *Md-ncm* in adults of M^Y strain after injection of dsRNA at early embryonic stages. A combination of dsRNA templates 33840_c0 and 22793_c0, injected in 1h old embryos of the M^Y strain.

Based on these eRNAi expression analysis data we can make three conclusions; (1) The expression pattern of *Mdmd* is the same in the body, testis and ovary of M^{II}, M^{III} and M^Y strain. (2) High levels of *Md-tra*^{F} transcripts are found in ovaries of M^{III} and M^Y strain suggesting that *Md-tra*^{F} transcripts are particularly expressed in tissues in which *Mdmd* expression was repressed. (3) *Mdmd* silencing does not affect the expression of *Md-ncm* in both male and female injected adult flies.

Embryonic RNAi of *Md-ncm* causes lethality in both males and females

We designed dsRNA template Mdncm5_7 against *Md-ncm* (Fig. 4.9) and injected it in the syncytial embryos of the M^{III} and M^{II} strain. A normal concentration (1 µg/µl) of *Md-ncm* dsRNA resulted in lethality of both males and females, so we reduced the concentration gradually. Gradually increasing the concentration of *Md-ncm* dsRNA led to decreasing percentage of surviving male and female individuals (Table 4.3). Control injections with dsRNA of *M122* resulted in normal survival rates similar to those observed in *Mdmd* silencing experiments. We conclude that the *Md-ncm* is essential for the development of both males and females.

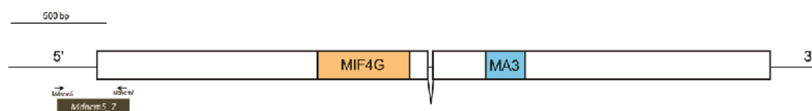

Figure 4.9: Design of *Md-ncm* dsRNA for RNAi. dsRNA template Mdncm5_7 was prepared from a region at the 5' end of *Md-ncm* which has low similarity to *Mdmd*.

Table 4.3. Results of embryonic RNAi of *Md-ncm*. The highest concentration of 1 µg/µl caused 100% lethality during embryonic development in both males and females of the M^{III} and M^{II} strain. A gradual reduction in the concentration of dsRNA Mdncm5_7 increases the number of surviving males and females of the M^{III} strain.

strain	dsRNA	concentration	no. injected embryos	no. hatching larvae	no. adult females	no. adult males	% larvae surviving to adulthood
M^{III}	Mdncm5_7	1 µg/µl	341	0	0	0	0
M^{III}	Mdncm5_7	0.20 µg/µl	880	154	0	3	0.01
M^{III}	Mdncm5_7	0.10 µg/µl	590	89	5	5	0.11
M^{III}	Mdncm5_7	0.05 µg/µl	506	54	4	6	0.18
M^{III}	Mdncm5_7	0.025µg/µl	857	72	21	16	0.51
M^{II}	Mdncm5_7	1 µg/µl	56	0	0	0	0
M^{III}	M122	1 µg/µl	507	149	45	48	0.62

DISCUSSION

The aim of the present study was to validate the role of *Mdmd* in male development. Embryonic RNAi of *Mdmd* led to the feminization of the gonads in *M* bearing individuals of the M^{II}, M^{III}, M^{V} and M^{Y} strains. Injected embryos developed as normal males externally, with male somatic tissues and genitalia, but, instead of testes, carried fully differentiated ovaries. These findings demonstrate that *Mdmd* is required for differentiation of male gonads in these *M. domestica* strains. Interestingly, eRNAi of *Mdmd* in the M^{I} strain did not lead to any feminization in males, suggesting that M^{I} males have either (a) different gene for male differentiation or the targeted sequences of *Mdmd* have diverged in the M^{I} strain to such an extent that RNAi was not effective.

We also have molecular confirmation of feminization following eRNAi of *Mdmd* from the expression analysis of *Mdmd*. Following repression of *Mdmd*, *Md-tra* splicing is irreversibly shifted to the female mode, *Md-tra*^{F}, in the gonads which produces a pair of normal looking ovaries. This result is consistent with the proposed function of *Mdmd* in shifting *Md-tra* splicing from female into the male-specific mode and suggests that *Mdmd* acts upstream of *Md-tra*.

How can we explain that only the gonads were feminized in *M* bearing individuals following RNAi of *Mdmd*? The most likely reason for incomplete feminization is that eRNAi in *M. domestica* is not long-lasting. Silencing can only be achieved in *M. domestica* by injecting dsRNA into the non-cellular environment of embryos when they are at the syncytial stage. Comparison of transcripts levels in 10h and 20h old embryos indicated that it has only a transient effect and expression increases to normal levels in subsequent development stages when dsRNA levels become diluted. Embryonic gonads may be the earliest targets of the sex-determining pathway, so once their sex is specified in the embryo, this is irreversible and cannot be altered at later developmental stages.

From this, we conclude that restoration of *Mdmd* activity at late developmental stages (e.g., during metamorphosis) can still repress the female-promoting activity of zygotic *Md-tra* in dimorphic tissues resulting in male rather than female differentiation. Based on these findings, we can conclude that silencing of *Mdmd* during early development irreversibly fixes the female fate of dimorphic gonadal tissue resulting in males producing eggs ("pregnant" males) instead of sperm because germline follows the sexual fate of the surrounding gonadal soma. In other words, the activity of *Mdmd* is essential and required for specification of male gonads.

Another important insight of this RNAi based study is that embryonic silencing of *Md-ncm* in early stage embryos (0-1h old) causes lethality of both males and females in the M^{III} and M^{II} strain consistent with its predicted vital role in general RNA splicing. Nevertheless, it is conceivable that *Md-ncm* is specifically needed to promote female splicing of *Md-tra* and that *Mdmd* acts as a dominant negative regulator of this specific function of *Md-ncm*. If *Md-ncm* is indeed involved in upholding the self-propagating loop of *Md-tra,* then reducing or compromising the activity of *Md-ncm* may shift splicing to the male mode and cause female-male transformation. When gradually decreasing the level of *Md-ncm* by eRNAi, we observed a decreasing percentage of surviving male and female individuals but no sex transformations in the surviving individuals with a female genotype.

It is possible that the essential functions of *Md-ncm* mask its possible role in sex determination. As an alternative approach, we suggest to overexpress *Md-ncm* in *Mdmd* bearing embryos to examine whether this can override the loop breaking activity of *Mdmd* and restore female splicing.

ACKNOWLEDGEMENTS

We thank Claudia Brunner for technical assistance with microinjections and Anna Rensink and Elzemiek Geuverink for technical assistance with qPCR.

Box 1.

Functional analysis of *Mdmd* by CRISPR/Cas9 approach

Here I describe my attempts to generate loss-of-function alleles of *Mdmd* with the CRISPR/Cas9 system. Embryonic RNAi treatment of *Mdmd* in very early embryos (0-1h) resulted in incomplete feminization of adults with male genotypes in the M^{II}, M^{III}, M^{V} and M^{Y} strains (chapter 4, Fig. 4.2 and 4.5). The most likely explanation for this partial feminization is that the knock-down effect is transient and that *Mdmd* expression returned to the normal level at later stages (after 20h) (chapter 4, Fig. 4.3). We only see effects in the ovaries as they are fixed during embryogenesis and cannot be altered even in the presence of *Mdmd* at later developmental stages. This interpretation was confirmed by measuring *Mdmd* levels of 10h and 20h after eRNAi treatment. To stably disrupt *Mdmd* functions throughout development we applied the CRISPR/Cas9 gene editing method to produce permanent loss-of-function alleles of *Mdmd*. Single crosses were set up with CRISPR/Cas9 cocktails (via plasmid, *in vitro transcribed* RNA and protein) injected G_0 adult black males which are heterozygous for the brown body mutation (*bwb*$^{+}$*/bwb*) linked with dominant phenotypic markers *pw*$^{+}$ and *bwb*$^{+}$ and untreated homozygous brown females (*bwb/bwb*) of the M^{III} strain. We reasoned that disrupting *Mdmd* in G_0 injected adults would give rise to black females (*bwb*$^{+}$*/bwb*) in the F_1 generation. Black females (*bwb*$^{+}$*/bwb*) would indicate a complete knockdown of *Mdmd* because of the germline mutation in the *Mdmd* allele which causes non-homologous end joining-mediated (NHEJ) disruption of *Mdmd* coding sequences (Fig. 4.10).

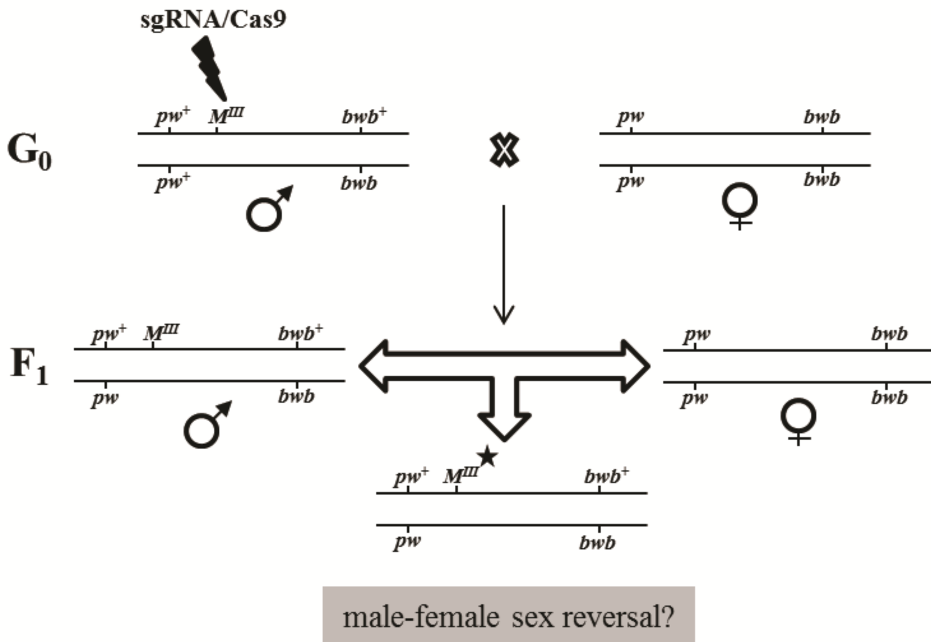

Figure 4.10: Experimental scheme for the CRISPR/Cas9 based knock-out studies. sgRNA/Cas9 complexes were injected into pw^{+} bwb^{+} males of the M^{III} strain, and these males were outcrossed with untreated females of the same M^{III} strain. F_1 progenies were screened for the occurrence of rare pw^{+} and bwb^{+} females.

All *M. domestica* strains and culture methods used for this study are described in chapter 2 and 3.

Design of *Mdmd* target disruption sites for CRISPR/Cas9 approach

The target sites for *Mdmd* disruption at the DNA level were chosen to be ~20 nucleotides (nt) long and followed by a PAM sequence (Protospacer adjacent motif, NGG) site which is required for the activity of Cas9 endonuclease (Anders *et al.*, 2014). The following primer pairs (5' to 3' direction) were used to generate single guide RNAs (sgRNAs): *Mdmd*-Cris1, *Mdmd*-Cris3, *Mdmd*-Cris5 and *Mdmd*-Cris6. The target-specificity of *Mdmd* sgRNAs should be 12 nt sequence preceding the PAM site to avoid off-targets effects on *Md-ncm*. Candidate target sequences for *Mdmd* were identified by the ZiFit Targeter software and tested for off-targets effects in the female genome of *M. domestica* by nucleotide blast

(BLASTN) program of NCBI. Forward and reverse primers corresponding to the selected target disruption sites of *Mdmd* were used to generate sgRNAs; forward primers were preceded (5') by CTTCG sequence and reverse primers followed (5') by CAAA sequence.

Cloning of the sgRNAs into the pU6-BbsI-chiRNA plasmid

The forward and reverse primers were diluted in TE buffer, aligned through thermo-cycling (95°C 5 min; then ramped to 25°C at a rate of -5°C in 26 cycles) and ligated with the pU6-BbsI-chiRNA plasmid previously digested with the BbsI (NBE) enzyme. Plasmids were transformed into *Escherichia coli* by standard cloning procedures using blue-white screening to pick up recombinant colonies for preparation (minipreps). Minipreps were performed with the GenElute plasmid miniprep Kit (Sigma-Aldrich). They were tested for plasmid insertion by PCR (2 min initial denaturation at 94°C, followed by 29, 32 and 35 cycles of 30 s at 92°C, 30 s at 59°C and 30 s at 72°C, and a final elongation of 3 min at 72°C). Primers used for PCR were respectively the reverse primer of each sgRNAs (*Mdmd*-Cris1, *Mdmd*-Cris3, *Mdmd*-Cris5 and *Mdmd*-Cris6) and a T7 primer. The transformed minipreps were used to generate midipreps with the GenElute HP Plasmid midiprep Kit (Sigma-Aldrich).

The injection-mix cocktail was prepared by co-precipitation of one or more transformed pU6-BbsI-chiRNA vectors and the phsp70-Cas9 plasmid (provided by Basler lab, IMLS, Zurich) and eluted into the 30 µl ddH20. Midipreps were diluted to a concentration of 250 ng/µl each, for injections of combined pU6-BbsI-chiRNAs. Single midipreps of pU6-BbsI-chiRNA were injected at a concentration of 500 ng/µl. The plasmid for Cas9 (phsp70Cas9) was diluted to a final concentration of 500 ng/µl for all injection-mix solutions. The co-precipitate of injection-mix was centrifuged for 2 min at RT preceding each micro-injection into 1h old embryos to eliminate possible crystal particles in the solution.

***In vitro* transcribed sgRNAs and sgRNAs/Cas9 constructs**

Target disruption-sites in the *Mdmd* sequence were chosen with a GG nucleotide pair at the start of the targeted sequence (5'), which is required to initiate the *in vitro* transcription of sgRNAs. The primer pairs used are Cr-Fa and Cr-Fb. For *in vitro* transcription of sgRNAs, a PCR (30 s initial denaturation at 98°C, followed by 35 cycles of 30 s at 98°C, 30 s at 60°C and 15 s at 72°C, and a final elongation of 10 min at 72°C) was performed with Phusion polymerase (New England Biolabs) and High-Fidelity buffer (New England Biolabs) in a final volume of 100 µl with only forward and reverse primer (no DNA template). The forward primer contained the T7 polymerase binding site followed by the sgRNA sequence. The T7 promoter sequence is located upstream of the forward primer to

yield the process of *in vitro* transcription. The reverse primer was a standard oligonucleotide incorporating the sgRNA sequence. PCR products were purified with PCR Product Purification Spin Kit (JETquick). sgRNAs were synthesized according to instructions of the Megatranscript T7 kit (Ambion) with 300 ng of target template and a 5' flanking T7 promoter as starting material. After RNA synthesis, template was removed by incubating with TurboDNase (mMESSAGE mMACHINE T7 Ultra Kit, Ambion) for 15 min at 37°C. The sgRNAs were purified by phenol/chloroform extraction and precipitated with isopropanol. In the final step sgRNAs were diluted to 10 µg/µl in RNAse free ddH20 and stored at -80°C. For microinjection, 8 ml injection solutions were made by mixing three separate sgRNAs aliquots. The injection mix was diluted to a final concentration of 65 ng/µl each and combined with the Cas9 mRNA (provided by Mosimann lab, IMLS, Zurich) diluted to a concentration of 500 ng/µl.

Microinjection of *in vitro* transcribed sgRNAs and Cas9 as a protein

SgRNAs (Cr-F1, Cr-F2, Cr-F3, Cr-F4, Cr-F5 and Cr-F6) were generated with the same protocol as described above. The Cas9 recombinant protein of lyophilised state (160 KD, *Streptococcus pyrogens*) was used from ToolGen genome engineering company (South Korea). The protein was reconstituted into 20 mM hepes, 150 mM kcl and 1% sucrose buffer solution and injected as 600 ng/ul for Cas9 protein and 250 ng/ul for each sgRNAs. Injection mix was incubated for 5 min at room temperature for the strong affinity of Cas9 and sgRNA complex before microinjection into 1h old embryos of the M^{III} strain.

Microinjection of sgRNA-Cas9 complexes into M^{III} embryos

Embryos of the M^{III} host strain were collected 1h after egg laying and the chorion membrane removed by incubating in 3% sodium hypochlorite solution (NaOCl) for 1.5 min. Dechorionated embryos were rinsed thoroughly with water and Ringer's solution. Embryos were aligned on a cover slip with posterior ends pointing to the injection site, dehydrated for 4 min in a silica gel chamber and covered with 3S/10S (1:4) Voltalef oil (Prolabo). A glass needle was filled with the preloaded sgRNA-Cas9 mix and injected into the posterior end of 0-1h old embryos. After injection, excess Voltalef oil was carefully removed and cover slips were placed on an agar plate overnight at 25°C. Surviving larvae were transferred to a small beaker with porcine manure. Males were collected shortly after eclosion and crossed with untreated virgin females of the M^{III} strain.

CRISPR/Cas 9 editing to generate *Mdmd* loss-of-function alleles

In our first strategy to target and disrupt the activity of *Mdmd* at the DNA level, I microinjected the combination of sgRNAs and Cas9 as a plasmid vector into syncytial stage

(0-1h) embryos of the M^{III} strain. To induce double strand breaks into *Mdmd*, I used well-tested CRISPR/Cas9 tools of *Drosophila*. For incorporating sgRNAs and tracrRNA (trans-activating crRNA) together, I used the pU6-chiRNA cloning vector under the control of *D. melanogaster* polymerase III derived U6 promoter. For expressing the Cas9 cDNA, I used a pHsp70-Cas9 vector under the control of a *D. melanogaster* heat shock promoter (Hsp70). Each combination of sgRNAs has two pu6-chiRNA plasmids and the one Cas9 expressing vector. This dual combination of sgRNAs was co-injected into 1h old embryos of the M^{III} strain (Fig. 4.11).

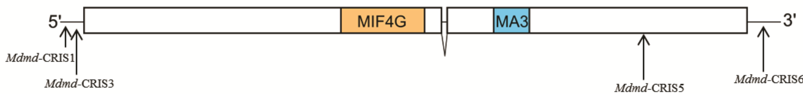

Figure 4.11: Co-injection of Cas9 and sgRNAs as a plasmid to target *Mdmd*. Four target sites (*Mdmd*-CRIS1, *Mdmd*-CRIS3, *Mdmd*-CRIS5 and *Mdmd*-CRIS6) were designed against *Mdmd* that are *Mdmd*-specific and not present in *Md-ncm* to avoid any off-target effects in the genome of *M. domestica*.

I screened 6142 adult males (bwb^{+}/bwb) of the F_1 generation, and all black individuals were normal looking fertile males. I dissected 50 adult males, and all had normal testes. Sexual dimorphic tissues (i.e., the distance between the eyes and the morphology of the genitalia) of injected males were also normal (Table 4.4). I did not find any feminized females among the F_1 adult male progeny (bwb^{+}/bwb), contrary to our expectations.

Table 4.4: Results of applying four distinct combinations of dual sgRNAs with the Cas9 expressing vector. The surviving rate of injected embryos is ~20%. All 6142 F_1 generation black individuals (*bwb⁺/bwb*) developed as normal males.

sgRNAs	injected embryos	larvae	G_0 adult males (*bwb⁺/bwb*)	G_0 adult females (*bwb/bwb*)	F_1 adult males (*bwb⁺/bwb*)	F_1 adult females (*bwb/bwb*)
Cr1 + Cr5	399	95 (24%)	19	18	1889	1502
Cr3 + Cr5	790	187 (23%)	33	35	1942	1662
Cr3 + Cr6	212	59 (22%)	14	9	1193	884
Cr1 + Cr6	393	86 (17%)	17	14	1118	1395

Next, I co-injected all four pU6-chiRNA vectors and provided heat shock treatment to injected embryos. After microinjection, I exposed the embryos to 42°C for 30 min in an incubator to enhance the transcription of the Hsp70 promoter driving Cas9 expression. This resulted in a reduction in the survival of injected embryos (Table 4.5). Again, I did not observe any feminization in the examined 5645 F_1 generation adults. They were all normal fertile males of *bwb⁺/bwb* genotype (Table 4.5). A likely cause for the decline in embryo survival is the higher efficiency of Hsp70 promoter driving Cas9 plasmid activity induced by heat shock treatment.

Table 4.5: Results of second attempt to combine four sgRNAs injections into 1h old embryos of M^{III} strain with heat shock treatment (42°C, 30 min). The survival rate of embryos is reduced to 10%, but all the 5645 F_1 individuals (*bwb⁺/bwb*) developed as males.

sgRNAs	injected embryos	larvae	G_0 adult males (*bwb⁺/bwb*)	G_0 adult females (*bwb/bwb*)	F_1 adult males (*bwb⁺/bwb*)	F_1 adult females (*bwb/bwb*)
Cr1+Cr3+ Cr5+Cr6	6840	742 (10%)	96	87	5645	4704

From our previous study, we know that *Mdmd* is present in multiple tandem repeated copies in the M^{III} strain (Sharma *et al.*, 2017). So in the second approach, we targeted multiple copies of *Mdmd* that might have a functional role. For this purpose, I designed the target sequence, *Mdmd*-CRI which is present in the highly conserved domains (MIF4G) of *Mdmd* and not similar to *Md-ncm* sequence (Fig. 4.12). This target sequence is expected to be expressed in all copies of the M^{III} strain. I used the same chiRNA (Chimeric RNA, SgRNA+Cas9) constructs as described in the first approach (Fig. 4.11) and microinjected again into 1h old embryos of the M^{III} strain. The experimental design was the same as described in Fig. 4.10. Again, I did not observe any male to female sex reversal in the 521 adult individuals of F_1 generation (Table 4.6).

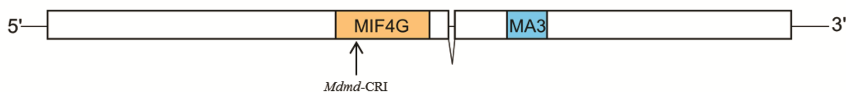

Figure 4.12: Co-injection of Cas9 and sgRNAs as a plasmid to target the *Mdmd* conserved domain MIF4G.

Table 4.6: Results of single sgRNA *Mdmd*-CRI injections into 1h old embryos of the M^{III} strain. The surviving rate of injected embryos is 16%. All the 521 F_1 individuals (*bwb*$^+$/*bwb*) developed as normal males.

sgRNAs	injected embryos	larvae	G_0 adult males (*bwb*$^+$/*bwb*)	G_0 adult females (*bwb/bwb*)	F_1 adult males (*bwb*$^+$/*bwb*)	F_1 adult females (*bwb/bwb*)
***Mdmd*-CRI**	1488	279 (16%)	35	33	521	449

In our third strategy to target *Mdmd*, I experimented with the sgRNAs, and Cas9 delivery as *in vitro* transcribed RNAs. I used this procedure to reduce the off-target effects and to enhance the efficiency of germ line transmission of mutated target sites (Fig. 4.13). I injected the combination of two sgRNAs (Cr-Fa and Cr-Fb) for disrupting *Mdmd* with capped and polyadenylated Cas9 mRNA into 1h old embryos. To check for *Mdmd* disruption, I used the same crossing procedure as described above (Fig. 4.10) but I did not observe any sex reverted females in the 2302 individuals of F_1 generation (Table 4.7).

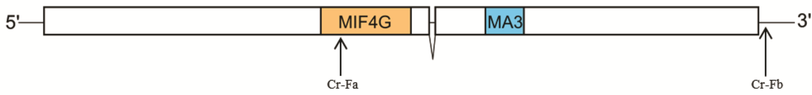

Figure 4.13: Co-injection of Cas9 and with sgRNAs (Cr-Fa and Cr-Fb) as *in vitro* transcribed RNAs to target *Mdmd*.

Table 4.7: Results of the combination of two sgRNAs (Cr-Fa and Cr-Fb) with Cas9 mRNA injected into 1h old embryos of M^{III} strain. All the 2302 F_1 adult individuals developed as normal males (bwb^+/bwb).

sgRNAs	injected embryos	larvae	G_0 adult males (bwb^+/bwb)	G_0 adult females (bwb/bwb,	F_1 adult males (bwb^+/bwb)	F_1 adult females (bwb/bwb)
Cr-Fa+Cr-Fb	1194	187 (16%)	45	38	2302	NA

As a fourth experimental approach to knock-out *Mdmd,* we designed six different *Mdmd*-specific target sites (Cr-F1, Cr-F2, Cr-F3, Cr-F4, Cr-F5 and Cr-F6) which were not present in *Md-ncm*. I injected the dual and combined sgRNAs of *Mdmd* with Cas9 protein into 1h old embryos of the M^{III} strain (Fig. 4.14). I did not observe any male to female sex reversal among 5472 F_1 individuals (Table 4.8).

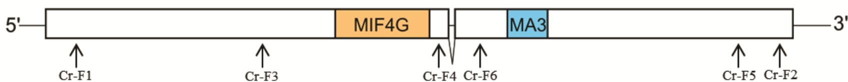

Figure 4.14: Co-injection of Cas9 as a protein and sgRNAs as *in vitro* transcribed RNAs.

Table 4.8: Results of injecting the dual (Cr-F1+Cr-F2, Cr-F3+Cr-F5, Cr-F4+Cr-F6) and combinations (Cr-F1, Cr-F2, Cr-F3, Cr-F4, Cr-F5 and Cr-F6) of sgRNAs into 1h old embryos of the M^{III} strain. All the 5472, F_1 individuals (*bwb*$^+$/*bwb*) developed as healthy males.

sgRNAs	injected embryos	larvae	G_0 adult males (*bwb*$^+$/*bwb*)	G_0 adult females (*bwb/bwb*)	F_1 adult males (*bwb*$^+$/*bwb*)	F_1 adult females (*bwb/bwb*)
Cr-F1+Cr-F2	1006	189 (19%)	24	31	1474	1468
Cr-F3+Cr-F5	1242	151 (12%)	26	17	2660	2501
Cr-F4+Cr-F6	1791	188 (11%)	22	20	716	552
(Cr-F1, Cr-F2, Cr-F3, Cr-F4, Cr-F5 and Cr-F6)	950	149 (16%)	40	21	622	511

In conclusion, we took four different approaches to using the CRISPR-Cas9 method for generating a loss of function alleles of *Mdmd* to disrupt the activity of *Mdmd*. Unfortunately, all our attempts, including the development of chiRNAs constructs (plasmid and *in vitro* transcribed RNA) and Cas9 (protein, *in vitro* transcribed RNA) failed to generate loss-of-function alleles of *Mdmd*. More recent attempts by other researchers were however successful to target *Mdmd* disruption using Cas9 protein which was expressed as an His-tagged protein and purified from bacteria (A. Meccariello and G. Saccone, personal communication) and resulted in complete sex reversal to fertile females, which is reported in our study (Sharma *et al.*, 2017).

APPENDIX 4.1

Primer sequences

PRIMER NAME	PRIMER SEQUENCE (5'-3')	PURPOSE
Md-ncm_9s	CTCATCGAAGGGCATTGGAGC	cDNA amplifications
Md-ncm_10as	AATAATATCGCCAGCTGTGGGTTT	cDNA amplifications
Mdmd_7s	ATCATCGGATAGGGTTGGAGG	cDNA amplifications
Mdmd_8as	ATAATGTCACCACTCGTGTATTTA	cDNA amplifications
Mdmd_6	GCTCTTCCCGGCGTCTTTTA	eRNAi
Mdmd_6as	GGTTGACGCGGACAATCAAC	eRNAi
Mdmd_1s	ATCAGGGCAAAGGGAAGTCG	eRNAi
Mdmd_1as	GATTGGCTCAGATCGGCGTA	eRNAi
Md-ncm5	TGAAAGCGAAACAAATGCTG	eRNAi
Md-ncm7	GTCAGGTTGCTTCTGAGGCG	eRNAi
22793_c0_qF1	TGGTGCGCCCTTCTTTAAAC	qPCR
22793_c0_qR1	GTTGACGCGGACAATCAACG	qPCR
β-Actin_F1	ATGGGTTGGTATGGGACA	qPCR
β-Actin_R1	CACGATTAGCCTTGGGAT	qPCR
Mdtra-24as	GATGCATTTTGTGCATCGCAA	cDNA amplifications
Md-tra2_F1	TTGCTTGAGTTGCCTGCTGC ATA	*Md-tra2*
Md-tra2_R1	CGTCCCCTGTAAACACCTGGG	*Md-tra2*
CYP6D3-1	GTTCGGTAATATTTGGCTTGG	control
CYP6D3-2	CCCGTATTCCGTAGTTGAATT	control

CRISPR primers (5'-3' direction):

Mdmd-Cris1	GCAGTCGCTAGCCTTCGCCCCG
Mdmd-Cris3	GTTTGAAAATAGCCGGCGTT
Mdmd-Cris5	GAAAATAAGAAAAAACATAA
Mdmd-Cris6	GGTATTTTTACCCTAATCCC
Cr-Fa	GGAATTACTACCCGAACCGA
Cr-Fb	GGTATTTTTACCCTAATCCC
Cr-F1	GCAAAGGGA AGTCGTAT
Cr-F2	GTTGAAAGATTTCCTAGAA
Cr-F3	CGATATAAA AAAGGGAAA
Cr-F4	TATCATCGGATAGGGTTGG
Cr-F5	ATGTTTAAACAATGTCAA
Cr-F6	GATAGTGAAATTAAAAGACGC

T7 polymerase
GAAATTAATAGCACTCACTATAGG18GTTTTAGAGCTAGAAATA
Common reverse primer
5’AAAAGCACCGACTGCGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATTT CTAGCTCTAAAAC3’

Chapter 5

Summarizing discussion

The aim of my research was to identify and characterize the *M*-factor of the housefly (*Musca domestica*) and to explain how *M*-factors have evolved on different locations in the housefly genome. Here, I briefly summarize the results and give suggestions for future studies on this intriguing topic.

Identification of the male-determining gene of *M. domestica*

The central core of the housefly sex determination system, as in many other insects, is the autoregulatory splicing loop of the *transformer (Md-tra)* gene. In this loop, the Md-TRA protein ensures its own synthesis by promoting the splicing of *Md-tra* transcripts into the female isoform ($Md\text{-}tra^{F}$). Md-TRA protein is essential for the female-specific splicing of transcripts of the downstream gene *doublesex (Md-dsx)*. If the autoregulatory loop is not started or interrupted, *Md-tra* transcripts will be spliced into the male isoform(s) ($Md\text{-}tra^{M}$), that contains a premature stop codon and consequently yield a non-functional truncated Md-TRA protein. As a consequence, *Md-dsx* transcripts will be spliced in the (default) male mode.

It was assumed that an *M*-factor is responsible for disrupting the autoregulatory *Md-tra* splicing loop. In order to do so, it was postulated that a male-determining gene should at least fulfill the following conditions: first, it should be present and expressed only in males, second, it should be expressed in a very early stage of embryonic development (before cellular blastoderm), and third, it should shift splicing of zygotic *Md-tra* transcripts into male isoforms (Hediger *et al.*, 2010).

By making use of the possibilities of *M. domestica* genetics, of which the details are described in chapter 2, two sets of early embryonic transcripts were generated, one male-specific and one female-specific. By comparing these sets, male-specific early embryonic transcripts were identified. The sequences of these transcripts were aligned with the available female housefly genome (Scott *et al.*, 2014). Obviously, male-specific *M*-factor sequences should be present in the early male transcript set, but absent from the female genome. The comparison yielded several possible candidates, described in chapter 2. One of these candidates stood out particularly, as it showed homology to the splicing factor CWC22. Because the *M*-factor is assumed to function as an interrupter of the autoregulatory splicing loop of *Md-tra* it was decided to consider this candidate as the most likely *M*-factor. The male-specific gene to have produced the transcripts was termed *Musca domestica male determiner* (*Mdmd*).

The male-specific sequences from the transcriptome analysis allowed the design of primer-combinations to compose the complete coding sequence of *Mdmd.* The gene codes

for a protein of 1190 amino acids and its coding sequence is interrupted by one intron. When this sequence was used in a BLASTN analysis of the female genome of *M. domestica*, a highly similar gene was detected. Comparison of *Mdmd* and this other gene revealed that both genes code for a protein with two conserved domains, MIF4G and MA3, but with diverged amino and carboxy terminal ends, and have an intron at the same position. The open reading frame of *Mdmd* is shorter owing to several small in-frame deletions. Phylogenetic analysis and synteny showed that this other gene is the ortholog of *nucampholin* in *Drosophila*, and therefore named *Md-ncm* (*Musca domestica-nucampholin*). This means that the male-determining gene *Mdmd*, that is present in the male genome only, shows high similarity to another gene, *Md-ncm*, that is present in the genomes of both males and females and deeply conserved from yeast to mammals. We were able to detect the presence of *Mdmd* in the genomes of male flies of which the *M*-factor has been mapped on the Y-chromosome or on chromosomes II, III, V, and Y respectively (M^{II}, M^{III}, M^{V} and M^{Y}). Further characterization of the molecular nature of *Mdmd* by sanger sequencing using divergent primer pairs between *Mdmd* copies revealed that *Mdmd* is present as tandemly repeated copies in the *M*-locus of different *M. domestica* strains (M^{II}, M^{III}, and M^{Y} strains) (Sharma *et al.*, 2017). Interestingly, however, no *Mdmd* gene could be identified for male flies with the *M*-factor mapped on chromosome I.

The high degree of similarity between *Mdmd* and *Md-ncm* suggests that *Mdmd* arose from *Md-ncm* by a duplication event. A comparison between sequences from *M. domestica* and other species, described in chapter 3, revealed a distinct divergence pattern of *Mdmd, Md-ncm* and other *ncm* sequences. The fact that the sequences of *Mdmd* are identical irrespective of the autosome from which they are derived, and are monophyletic with *Md-ncm* led to two conclusions: *Mdmd* likely arose as a duplication (paralog) of *Md-ncm* and subsequently translocated to different genomic sites.

Is *Mdmd* functioning as a male determiner?

To confirm the male-determining function of *Mdmd* a functional analysis was performed by transiently silencing *Mdmd* in early embryos of different *M*-strains. The results from embryonic silencing in chapter 4 confirmed that early *Mdmd* activity is required for differentiation of male gonads in M^{II}, M^{III}, M^{V} and M^{Y} strains. The fact that only gonadal tissues were sex reverted is most likely explained by the transient effect of injected dsRNA. Once *Mdmd* expression is restored at late embryonic stages, when dsRNA is no longer present, it will repress zygotic *Md-tra* during the larval/pupal stage transition when most somatic tissues are sexually specified. Quantitative *Mdmd* expression analysis indeed showed that *Mdmd* expression levels are back to the normal level after 20h post injection of *Mdmd*-dsRNA into only male-embryos. Based on these findings, it was concluded that

Mdmd is required for the development of male gonads and that the male fate of gonadal tissues becomes already irreversibly fixed at embryonic stages.

In contrast to effects in males from M^{II}, M^{III}, M^{V} and M^{Y} strains, no phenotypic effect was observed in M^{I} males, consistent with the apparent lack of an *Mdmd* gene in males of this strain. The finding that the male determiner in M^{I} is different from *Mdmd* opens an exciting possibility that even intraspecific variation exists regarding the nature of male determiners (Sharma *et al.*, 2017).

It was later shown that by generating loss-of-function alleles of *Mdmd* using the CRISPR/Cas9 approach, targeted *Mdmd* disruption results in complete sex reversal of males to fertile females. Such females with disrupted *Mdmd* exclusively express the female splice variants of *Md-tra* and *Md-dsx* which is consistent with the role of *Mdmd* in directing male splicing of *Md-tra* (Sharma *et al.*, 2017). Another convincing validation for *Mdmd* to be the male determiner would be to transiently misexpress *Mdmd* in genotypically female individuals and see if *Mdmd* is sufficient for male determination.

Is *Mdmd* a direct regulator of *Md-tra* splicing?

Mdmd is a paralog of *Md-ncm* which is required for pre-mRNA splicing in metazoans. In this section, I present two possible models to explain the evolution of the male-determining function of *Mdmd* after its duplication from *Md-ncm*. (I) *Mdmd* acquired a male-determining function by neofunctionalization adopting a new molecular function unrelated to *Md-ncm* or, (II) *Mdmd* behaves as a dominant negative regulator, interfering with those functions of *Md-ncm* which are required to promote female splicing of *Md-tra*. The latter presumes that *Md-ncm* is a direct positive regulator of *Md-tra*. If *Md-ncm* is indeed required for female *Md-tra* splicing, then a gradual decrease of its activity by eRNAi-based silencing in genotypically female individuals may shift *Md-tra* splicing to the male mode and, as a result, male fate may follow. I tested this hypothesis by microinjecting increasingly higher concentrations of *Md-ncm* dsRNA into early-stage embryos of the M^{III} strain. Increasing the concentration of dsRNA led to higher levels of lethality in both males and females (see chapter 4) and, thereby, possibly masking a role of *Md-ncm* in sex determination. So, as a future strategy to test whether *Md-ncm* is involved in this regulation and antagonized by *Mdmd*, a reciprocal experiment could be performed in which *Md-ncm* activity is increased in *Mdmd* containing embryos. Suppression by increasing the dose of wild-type activity is a commonly used strategy to validate antimorphic mutations. In this case, it is not the antimorphic activity against a wild-type allele but the antagonistic effect against the paralog (*Md-ncm*) which is tested. If our hypothesis of model II (Figure 5.1) is correct, the extra provision of *Md-ncm* activity, e.g., by injecting mRNA or genomic fragments of *Md-ncm*,

ill override the antagonistic activity of *Mdmd* and restore female splicing of *Md-tra*. We expect these *Mdmd* individuals to be feminized.

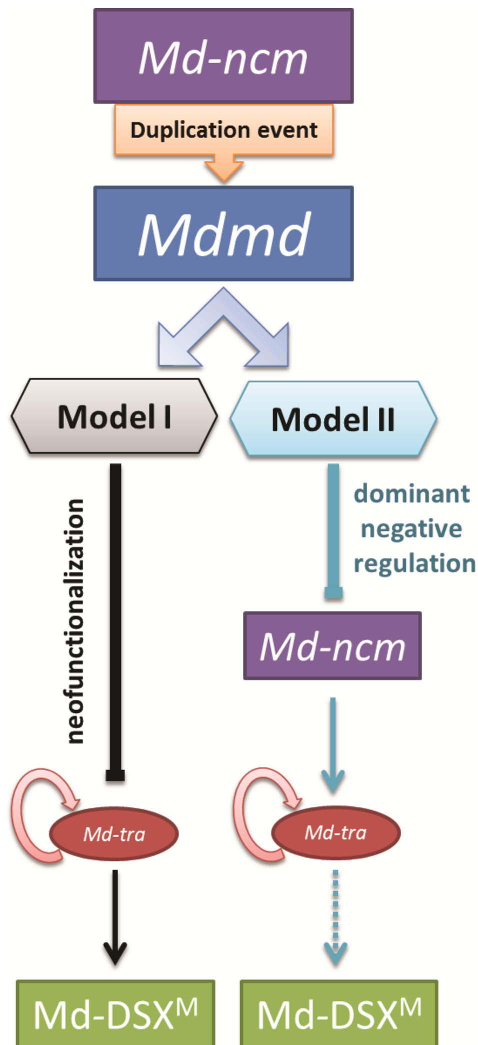

Figure 5.1. Model for *Mdmd* role in *Md-tra* splicing. In model I, we hypothesize that *Mdmd* acquired a novel male-determining function after duplication which blocks the selfregulatory loop of zygotic *Md-tra*. In model II, we hypothesize that *Md-ncm* is essential for promoting female splicing of *Md-tra* and that *Mdmd* became a dominant negative regulator of *Md-ncm*. In both models, the loop of *Md-tra* is shut down and, in absence of active Md-TRA, its target, the double-switch gene *doublesex* (*dsx*), produces male-specific DSX protein (Md-DSX^{M}) and male development ensues.

Pacific Bioscience (PacBio) sequencing of M^{III} strain males identified at least six copies of *Mdmd*. In the *M*-locus, these six copies of *Mdmd* exist as long reads of 12kb (M3.1 and M3.2), 17kb (M1, M2.1, and M2.2) and 20kb (M4, M5, and M6) region respectively and are not identical to each other. Interestingly among all *Mdmd* copies, three (M4, M5, and M6) are similar to each other and contain a full-length *Mdmd* genomic sequence, but we could not find an ORF because of the inaccuracy of PacBio sequencing or because these copies do not contain an ORF. These three copies are particularly interesting for further investigation because they are located one after another and separated by lncRNAs (~400bp) (Fig. 5.2). BLASTN search of these lncRNAs revealed that they are unique to *M. domestica*.

Mdmd has a long ORF, so it is particularly vulnerable to the accumulation of deleterious mutations. These multiple non-functional copies may have arisen from local amplification to preserve *Mdmd* functionality in a non-recombining region. Another explanation for having the multiple *Mdmd* copies in the *M*-locus is to enhance the quantitative expression of *Mdmd* in developing embryos for directing the downstream components of sex-determining pathways that ensure the sexual differentiation as male and female.

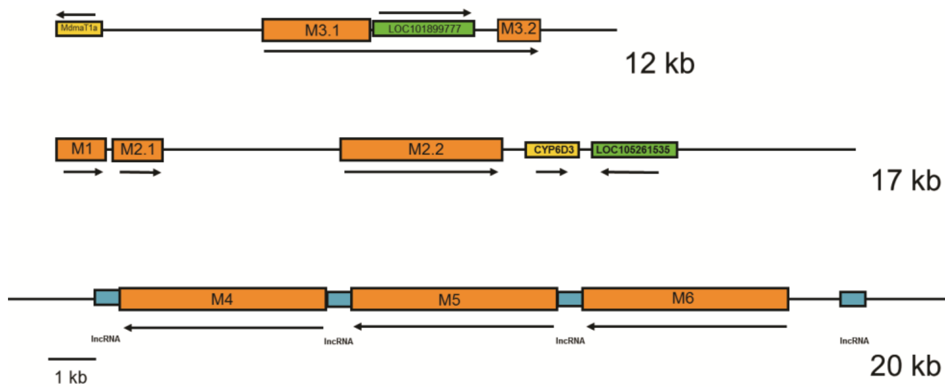

Figure 5.2. Number and organization of *Mdmd* copies in the *M*-locus as detected by PacBio sequencing of $M^{III}/+$ males. The 12 kb region has one copy of *Mdmd* (M3.1 and M3.2, yellow color) which flanked by retrotransposons (yellow and green color). The 17 kb region has two copies of *Mdmd* (M1, M2.1, and M2.2, yellow color) and retrotransposons (yellow and green color). The 20 Kb region has three copies of *Mdmd* (M4, M5, and M6) flanked by lncRNAs (~400bp, blue color) (Pac-Bio derived data sets were analysed by Mark Robinson and Stephan Schmeing, UZH).

Future perspectives

There is inadequate knowledge about the nature and action of male-determining genes in insects. Until now only two male-determining genes *Nix* and Y-linked *Yob* in mosquitoes have been identified. Ectopic expression of *Nix* in *Aedes aegypti* partially masculinizes female genitalia (Hall *et al.*, 2015) and in *Anopheles gambiae*, ectopically expressed *Yob* transcripts causes lethality in XX individuals most likely due to disruption of dosage compensation (Krzywinska *et al.*, 2016). One promising approach to prove the male-determining function of *Mdmd* would be misexpression analysis. Injection of *Mdmd* capped mRNA containing an intact ORF into early stage housefly embryos may be sufficient to cause masculinization of genotypic female individuals. In a different approach, a gDNA fragment or plasmid clone with an intact *Mdmd*, and all the promoter elements necessary for accurate expression, could be microinjected into early-stage embryos and tested for masculinization. This approach is similar to Koopman *et al*. (1991) who showed that a 14-kilobase genomic DNA fragment of *Sry* is sufficient to induce testis differentiation and subsequent male development when introduced into chromosomally female (XX) mouse embryos.

Nix encodes a 288–amino acid polypeptide with two RNA recognition motifs, and *Yob* encodes a short 56–amino acid protein. The housefly male determiner MDMD, on the other hand, has 1190-amino residues with two conserved domains, MIF4G and MA3. The lack of sequence similarity among *Nix, Yob*, and *Mdmd* is likely due to differences in the targets that are being regulated by these male determiners. *A. aegypti* and *A. gambiae* both appear to lack a *tra* ortholog, but both use a *dsx* ortholog as the final effector at the bottom of the pathway. It is believed that *dsx* is not the direct target of *Yob* and *Nix*, but instead an intermediate transducer equivalent to *tra* exists, but that has not been identified yet. It is thus conceivable that mosquitoes use a transduction system different from *tra* to regulate *dsx*.

In conclusion, our study demonstrates how a novel gene can arise by duplication and adopt a critical role in an essential developmental process such as sex determination. It gives further proof to a general principle in evolution that duplications of existing genes are used as play dough for creating new functions and thus increasing the genomic repertoire for expressing new adaptive traits. Our study further provides insights into the causes of the remarkable diversity in sex-determining pathways. There is also an applied significance of our study. Houseflies are vectors of many diseases in humans causing thousands of deaths each year worldwide. Understanding the molecular basis of sex determination can help to improve existing or develop novel genetic strategies, like CRISPR-Cas9 gene drive systems and Sterile Insect Technique (SIT) in which large numbers of sterilised males are released

into natural populations to outcompete the fertile males and restrict reproduction. These approaches will help to control houseflies and other insect pests.

References

Alexandrov A, Colognori D, Shu M-D, Steitz JA (2012). Human spliceosomal protein CWC22 plays a role in coupling splicing to exon junction complex deposition and nonsense-mediated decay. *Proc Natl Acad Sci U S A* **109**: 21313–8.

Anders C, Niewoehner O, Duerst A, Jinek M (2014). Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513**: 569–73.

Anderson JL, Marí A, Braasch I, Amores A, Hohenlohe P, Batzel P, *et al.* (2012). Multiple sex-associated regions and a putative sex chromosome in zebrafish revealed by RAD mapping and population genomics. *PLoS One* **7**: e40701.

Bachtrog D et al (2014). Sex determination: why so many ways of doing it? *PLoS Biol* **12**: e1001899.

Barbosa I, Haque N, Fiorini F, Barrandon C, Tomasetto C, Blanchette M, *et al.* (2012). Human CWC22 escorts the helicase eIF4AIII to spliceosomes and promotes exon junction complex assembly. *Nat Struct Mol Biol* **19**: 983–29.

Beukeboom LW and Perrin N (2014). The evolution of sex determination. Oxford Univ Press.

Beverley SM, Wilson AC (1984). Molecular evolution in *Drosophila* and the higher Diptera. *J Mol Evol* **21**: 1–13.

Beye M, Hasselmann M, Fondrk MK, Page RE, Omholt SW (2003). The gene *csd* is the primary signal for sexual development in the honeybee and encodes an SR-type protein. *Cell* **114**: 419–429.

Blackmon H, Hardy NB, Ross L (2015). The evolutionary dynamics of haplodiploidy: Genome architecture and haploid viability. *Evolution* (*N Y*) **69**: 2971–2978.

Bopp D (2010). About females and males: Continuity and discontinuity in flies. *J Genet* **89**: 315–323.

Bopp D, Saccone G, Beye M (2014). Sex determination in insects: Variations on a common theme. *Sex Dev* **8**: 20–28.

Bradley KM, Breyer JP, Melville DB, Broman KW, Knapik EW, Smith JR (2011). An SNP-based linkage map for zebrafish reveals sex determination loci. *G3 Genes|Genomes|Genetics* **1**: 3–9.

Burghardt G, Hediger M, Siegenthaler C, Moser M, Dübendorfer A, Bopp D (2005). The *transformer2* gene in *Musca domestica* is required for selecting and maintaining the

female pathway of development. *Dev Genes Evol* **215**: 165–176.

Burtis KC, Baker BS (1989). *Drosophila doublesex* gene controls somatic sexual differentiation by producing alternatively spliced mRNAs encoding related sex-specific polypeptides. *Cell* **56**: 997–1010.

Çakır Ş, Kence A (1996). The distribution of males having XY and XX chromosomes in housefly populations (Diptera: Muscidae) of Turkey. *Genetica* **98**: 205–210.

Carroll SB, Prudhomme B, Gompel N (2008). Regulating evolution. *Sci Am* **298**: 60–67.

Cline TW (1983). The interaction between *daughterless* and *sex-lethal* in triploids: A lethal sex-transforming maternal effect linking sex determination and dosage compensation in *Drosophila melanogaster*. *Dev Biol* **95**: 260–274.

Coelho CMA, Kolevski B, Walker CD, Lavagi I, Shaw T, Ebert A, *et al.* (2005). A genetic screen for dominant modifiers of a small-wing phenotype in *Drosophila melanogaster* identifies proteins involved in splicing and translation. *Genetics* **171**: 597–614.

Crozier RH (1971). Heterozygosity and sex determination in haplodiploidy. *American Naturalist* **105**: 399-412.

Denholm I, Franco MG, Rubini PG, Vecchi M (1983). Identification of a male determinant on the X chromosome of housefly (*Musca domestica* L.) populations in south-east England. *Genet Res Cambridge* **42**: 311–322.

Dübendorfer A, Hediger M, Burghardt G, Bopp D (2002). *Musca domestica*, a window on the evolution of sex-determining mechanisms in insects. *Int J Dev Biol* **46**: 75–79.

Ellegren H (2011). Sex-chromosome evolution: recent progress and the influence of male and female heterogamety. *Genetics* **12**: 157–66.

Erickson JW, Quintero JJ (2007). Indirect effects of ploidy suggest X chromosome dose, not the X:A ratio, signals sex in *Drosophila*. *PLoS Biol* **5**: 2821–2830.

Feldmeyer B, Kozielska M, Weissing FJ, Beukeboom LW, Pen I (2008). Climatic variation and the geographical distribution of sex determination mechanisms in the housefly. *Evol Ecol Res* **10**: 797–809.

Fotedar R, Baberjee U, Singh S, Shriniwas, Verma AK (1992). The housefly (*Musca domestica*) as a carrier of pathogenic microorganisms in a hospital environment. *J Hosp Infect* **20**: 209–215.

Franco MG, Rubini PG, Vecchi M (1982a). Sex determinants and their distribution in various populations of *Musca domestica* L. of W. Europe. *GenetRes* **40**: 279–294.

Gempe T, Beye M (2011). Function and evolution of sex determination mechanisms, genes and pathways in insects. *BioEssays* **33**: 52–60.

Geuverink E, Beukeboom LW (2014). Phylogenetic distribution and evolutionary dynamics of the sex determination genes *doublesex* and *transformer* in insects. *Sex Dev* **8**: 38–49.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, *et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**: 644–52.

Greenberg, B (1965). Flies and disease. *Sci. Am.* **213**: 92-99.

Gubbay J, Collignon J, Koopman P, Capel B, Economou A, Münsterberg A, *et al.* (1990). A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. *J Colloid Interface Sci* **374**: 685–689.

Hall AB, Basu S, Jiang X, Qi Y, Vladimir A, Biedler JK, *et al.* (2015). A male-determining factor in the mosquito *Aedes aegypti*. *Science* **348**: 1268–1270.

Hamm RL, Meisel RP, Scott JG (2015). The evolving puzzle of autosomal versus Y-linked male determination in *Musca domestica*. *G3 Genes|Genomes|Genetics* **5**: 371–384.

Hamm RL, Shono T, Scott JG (2005). A cline in frequency of autosomal males is not associated with insecticide resistance in house fly (Diptera: Muscidae). *J Econ Entomol* **98**: 171–176.

Hediger M, Henggeler C, Meier N, Perez R, Saccone G, Bopp D (2010). Molecular characterization of the key switch *F* provides a basis for understanding the rapid divergence of the sex-determining pathway in the housefly. *Genetics* **184**: 155–170.

Hediger M, Minet AD, Niessen M, Schmidt R, Hilfiker-Kleiner D, Çakir Ş, *et al.* (1998). The male-determining activity on the Y chromosome of the housefly (*Musca domestica* L.) consists of separable elements. *Genetics* **150**: 651–661.

Herpin A, Schartl M (2015). Plasticity of gene-regulatory networks controlling sex

determination: of masters, slaves, usual suspects, newcomers, and usurpators. *EMBO Rep* **16**: 1–15.

Hilfiker-kleiner D, Dübendorfer A, Hilfiker A, Nöthiger R (1994). Genetic control of sex determination in the germ line and soma of the housefly, *Musca domestica*. *Development* **2538**: 2531–2538.

Hilfiker-Kleiner D, Dübendorfer A, Hilfiker A, Nöthiger R (1993). Developmental analysis of two sex-determining genes, *M* and *F*, in the housefly, *Musca domestica*. *Genetics* **134**: 1187–1194.

Hiroyoshi (1961). The linkage map of the housefly, *Musca domestica* L. *Genetics* **46**: 1373–1380.

Hiroyoshi T (1964). Sex-limited inheritance and abnormal sex ratio in strains of the Housefly. *Genetics* **50**: 373–385.

Inoue H, Fukumori Y, Hiroyoshi T (1983). Mapping of autosomal male-determining factors of the housefly, *Musca domestica* L., by means of sex-reversal. *Japanese J Genet* **58**: 451–461.

Inoue H, Hiroyoshi T (1986). A maternal-effect sex-transformation mutant of the housefly, *Musca domestica* L. *Genetics* **112**: 469–482.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, *et al.* (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**: 1647–1649.

Keyes LN, Cline TW, Schedl P (1992). The primary sex determination signal of *Drosophila* acts at the level of transcription. *Cell* **68**: 933–943.

Kiuchi T, Koga H, Kawamoto M, Shoji K, Sakai H, Arai Y, *et al.* (2014). A single female-specific piRNA is the primary determiner of sex in the silkworm. *Nature* **509**: 4–6.

Kozielska M, Feldmeyer B, Pen I, Weissing FJ, Beukeboom LW (2008). Are autosomal sex-determining factors of the housefly (*Musca domestica*) spreading north? *Genet Res Cambridge* **90**: 157–165.

Krzywinska E, Dennison NJ, Lycett GJ, Krzywinski J (2016). A maleness gene in the malaria mosquito *Anopheles gambiae*. *Science* **2927**: 1431–1434.

Li B, Dewey CN (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**: 323.

McDonald IC, Evenson P, Nickel CA, Johnson, OA (1978). House fly genetics: isolation of a female determining factor on chromosome 4. *Ann. Ent. Soc. Am.*, **75**: 692–694

Manolakou P, Lavranos G, Angelopoulou R (2006). Molecular patterns of sex determination in the animal kingdom: a comparative study of the biology of reproduction. *Reprod Biol Endocrinol* **4**: 59.

Marchler-Bauer A, Bryant SH (2004). CD-Search: Protein domain annotations on the fly. *Nucleic Acids Res* **32**: 327–331.

Marin I, Baker SB (1998). The evolutionary dynamics of sex determination. *Science* **281**: 1990–1994.

Mawaribuchi S, Yoshimoto S, Ohashi S, Takamatsu N, Ito M (2012). Molecular evolution of vertebrate sex-determining genes. *Chromosom Res* **20**: 139–151.

McCarthy DJ, Chen Y, Smyth GK (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* **40**: 4288–4297.

McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, *et al.* (2013). Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res* **41**: 597–600.

Nickel CA and Wagoner DE (1974). Mutants on linkage gropus 3 and 4 of the house fly. *Ann Entomol Soc Am* **67**: 775-776.

Normark BB (2003). The evolution of alternative genetic systems in insects. *Annu Rev Phys Chem* **48**: 397–423.

Ohbayashi F, Suzuki MG, Mita K, Okano K, Shimada T (2001). A homologue of the *Drosophila* doublesex gene is transcribed into sex-specific mRNA isoforms in the silkworm, *Bombyx mori*. *Comp Biochem Physiol-B* **128**: 145–158.

Pane A, Salvemini M, Delli Bovi P, Polito C, Saccone G (2002). The *transformer* gene in *Ceratitis capitata* provides a genetic basis for selecting and remembering the sexual fate. *Development* **129**: 3715–3725.

Parkhurst SM and Meneely PM (1994). Sex determination and dosage compensation: lessons from flies and worms. *Science* 264: 924–932.

Pease JB, Hahn MW (2012). Sex chromosomes evolved from independent ancestral linkage groups in winged insects. *Mol Biol Evol* **29**: 1645–1653.

Ramakers C, Ruijter JM, Lekanne Deprez RH, Moorman AFM (2003). Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neurosci Lett* **339**: 62–66.

Robinson MD, McCarthy DJ, Smyth GK (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140.

Saccone G, Pane A, Polito LC (2002). Sex determination in flies, fruitflies and butterflies. *Genetica* **116**: 15–23.

Salvemini M, D'Amato R, Petrella V, Aceto S, Nimmo D, Neira M, *et al.* (2013). The orthologue of the fruitfly sex behaviour gene fruitless in the mosquito *Aedes aegypti*: evolution of genomic organisation and alternative splicing. *PLoS One* **8**: e48554

Sánchez L (2008). Sex-determining mechanisms in insects. *Int J Dev Biol* **52**: 837–56.

Sarre SD, Georges A, Quinn A (2004). The ends of a continuum: Genetic and temperature-dependent sex determination in reptiles. *BioEssays* **26**: 639–645.

Schmidt R, Hediger M, Nöthiger R, Dübendorfer A (1997). The mutation *masculinizer* (*man*) defines a sex-determining gene with maternal and zygotic functions in *Musca domestica* L. *Genetics* **145**: 173–183.

Schmidt R, Hediger M, Roth S, Nöthiger R, Dübendorfer A (1997). The Y-chromosomal and autosomal male-determining *M* factors of *Musca domestica* are equivalent. *Genetics* **147**: 271–280.

Schütt C, Nöthiger R (2000). Structure, function and evolution of sex-determining systems in Dipteran insects. *Development* **127**: 667–677.

Scott JG, Liu N, Kristensen M, Clark AG (2009). A case for sequencing the genome of *Musca domestica* (Diptera: Muscidae). *J Med Entomol* **46**: 175–182.

Scott JG, Warren WC, Beukeboom LW, Bopp D, Clark AG, Giers SD, *et al.* (2014). Genome of the house fly, *Musca domestica* L., a global vector of diseases with adaptations to a septic environment. *Genome Biol* **15**: 466.

Sharma A, Heinze SD, Wu Y, Kohlbrenner T, Morilla I, Brunner C, *et al.* (2017). Male sex in houseflies is determined by *Mdmd*, a paralog of the generic splice factor gene *CWC22*. *Science* **356**: 642–645.

Shiimori M, Inoue K, Sakamoto H (2013). A specific set of exon junction complex subunits

is required for the nuclear retention of unspliced RNAs in *Caenorhabditis elegans*. *Mol Cell Biol* **33**: 444–456.

Steckelberg AL, Altmueller J, Dieterich C, Gehring NH (2015). CWC22-dependent pre-mRNA splicing and eIF4A3 binding enables global deposition of exon junction complexes. *Nucleic Acids Res* **43**: 4687–4700.

Tomita T, Wada Y (1989). Multifactorial sex determination in natural populations of the housefly (*Musca domestica*) in Japan. *Japanese J Genet* **64**: 373–382.

Traut W, Niimi T, Ikeo K, Sahara K (2006). Phylogeny of the sex-determining gene sex-lethal in insects. *Genome* **49**: 254–262.

Traut W, Sahara K, Marec F (2008). Sex chromosomes and sex determination in Lepidoptera. *Sex Dev* **1**: 332–346.

Traut W, Willhoeft U (1990). A jumping sex determining factor in the fly *Megaselia scalaris*. *Chromosoma* **99**: 407–412.

Valenzuela N, Adams DC, Janzen FJ (2003). Pattern does not equal process: exactly when is sex environmentally determined? *Am Nat* **161**: 676–683.

Verhulst EC, van de Zande L, Beukeboom LW (2010a). Insect sex determination: It all evolves around *transformer*. *Curr Opin Genet Dev* **20**: 376–383.

Vicoso B, Bachtrog D (2015). Numerous transitions of sex chromosomes in Diptera. *PLoS Biol* **13**: 1–22.

Werren JH, Beukeboom LW (1998). Sex determination, sex ratios, and genetic conflict. *Annu Rev Ecol Syst* **29**: 233–261.

West SA, Sheldon BC (2002). Constraints in the evolution of sex ratio adjustment. *Science* **295**: 1685–1688.

Wilkins AS (1995). Moving up the hierarchy: A hypothesis on the evolution of a genetic sex determination pathway. *BioEssays* **17**: 71–77.

Wu TD, Watanabe CK (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**: 1859–1875.

English summary

Proper assignment of the sexual fate is an essential developmental process. In insects, a variety of sex determination mechanisms exists with different chromosomal systems and sex determination genes. Sex is determined by a hierarchically arranged cascade of gene regulatory events that consists of a primary instructive signal at the top followed by a binary transducing switch and a global effector at the bottom (Bopp *et al.*, 2014). This sex determination gene cascade is activated in early embryos and splices downstream genes sex-specifically. The sex-specific products of the most downstream gene, *doublesex*, then direct differentiation of the developing embryo into either male or female. The primary genetic signals are highly diverse, but the downstream genes more conserved (Gempe and Beye, 2011; Bopp *et al.*, 2014). According to the "bottom-up hypothesis" proposed by Wilkins (1995), these cascades evolve from the bottom upwards, by consecutive recruitment of genes at the top.

The housefly, *Musca domestica*, is a remarkable example of diversity in sex-determining mechanisms and is, therefore, an ideal model to study evolutionary diversification of sex determination pathways (Dübendorfer *et al.*, 2002). The polymorphic sex determination system of houseflies has been known for more than five decades and is unique in the existence of natural populations with dominant male-determiners (*M*-factors) present on the Y chromosome or any of the five autosomes (Hiroyoshi, 1964; Franco *et al.*, 1982; Denholm *et al.*, 1983). It has been a longstanding question whether these elusive *M*-factors represent different genes or copies of the same gene which translocated to different sites in the genome.

The *M*-factor is the primary signal in the sex-determination cascade and prevents the maternal activation of the zygotic *Md-tra* self-regulatory loop (Inoue and Hiroyoshi, 1986; Hilfiker-Kleiner *et al.*, 1993). It results in a non-functional splice form of the *transformer* gene and subsequent splicing of *doublesex* into the (default) male form (Hediger *et al.*, 2010). The main aim of my research was to identify and characterize the *M*-factor(s) of the housefly. I also addressed an important evolutionary question, i.e., how *M*-factors can evolve on different locations in the genome.

For identification of the *M*-factor, it was assumed that *M* is expressed only in males and shifts splicing of zygotic *Md-tra* transcripts into male isoforms already at the early pre-blastoderm stages (Hediger *et al.*, 2010). Based on these premises we performed a differential expression analysis (based on RNAseq) of only male and only female embryos that were collected at the syncytial embryonic stage. This yielded several mRNAs that were present in male but not female embryos. They were then compared to the female genome sequences of *M. domestica* (Scott *et al.*, 2014) which revealed several male-biased orphan reads that are present in males and absent in females. Amongst these male-biased orphan reads, five male-specific contigs were found with a high level of sequence similarity to the

pre-mRNA splicing factor coding gene *CWC22/nucampholin*. Different sets of primers were designed to amplify overlapping parts of these five male-specific contigs and confirmed they are part of the same transcriptional unit. We named this putative male determiner of the housefly *Mdmd* (for *Musca domestica male determiner*) (chapter 2).

Based on high sequence similarity of *Mdmd* with the generic splice factor gene *CWC22*, we inferred that *Mdmd* is a paralog of *CWC22* (Sharma *et al*., 2017). CWC22 has a general pre-mRNA splicing function and is required for exon junction complex (EJC) assembly (Barbosa *et al*., 2012; Steckelberg *et al*., 2012; Alexandrov *et al*., 2012). BLASTN searches of *Mdmd* sequences in the female genome of *M. domestica* identified its paralogous gene *Md-ncm* (for *Musca domestica-nucampholin*) which is an ortholog of *ncm* in *D. melanogaster*. The high sequence similarity between *Mdmd* and *Md-ncm* indicated that *Mdmd* originated from *Md-ncm* by a gene duplication event. Two domains of MDMD, MIF4G, and MA3 are conserved among metazoans and involved in mRNA surveillance, translation and non-sense decay (NMD) in many species (Alexandrov *et al*., 2012).

Mdmd was initially identified from the M^{III} strain of *M. domestica* in which males carry the *M*-factor on the third chromosome. One immediate follow-up question was whether *Mdmd* is also present on the Y chromosome and any of the other autosomes. The finding and sequence identity of *Mdmd* in males of different *M*-strains (M^{II}, M^{III}, M^{V}, and M^{Y}) suggests that *Mdmd* is one and the same gene in all these *M*-strains. In contrast to *Mdmd*, and as expected, the generic spliceosomal factor *Md-ncm* is present in the genomes of both males and females in these different *M*-strains. A phylogenetic analysis of *Mdmd* in different *M*-strains revealed that they cluster tightly together, indicating that Y-linked and autosomal *M*-factors have a common origin and translocated to different genomic sites in the genome of *M. domestica* (chapter 3). Phylogenetic analysis with MDMD, MD-NCM and NCM protein sequences among the different phyla of metazoans indicate a distinct divergence pattern of *Mdmd* from *Md-ncm*, whereby *ncm* follows the established phylogenetic pattern among metazoans. In other higher dipterans, paralogs of *ncm* have not been found suggesting that the recruitment of *Mdmd* as a male determiner is unique to houseflies (chapter 3). The other interesting point is that *Mdmd* is present as tandemly repeated functional and non-functional copies in the *M*-locus of different *M. domestica* strains (Sharma *et al*., 2017). These multiple *Mdmd* copies may have arisen by local amplification on the Y chromosome or an autosome, before and/or after translocation to preserve the functionality of *Mdmd* (chapter 5).

To functionally test the male-determining function of *Mdmd*, embryonic RNAi (eRNAi) of *Mdmd* was performed in early embryos of different *M*-strains. eRNAi based silencing of *Mdmd* confirmed that early *Mdmd* activity is required for differentiation of male gonads in M^{II}, M^{III}, M^{V} and M^{Y} strains (chapter 4). In contrast, no sex-reversed

phenotype was observed in M^I males which suggests that males of the M^I strain carry a different kind of *M*-locus. The partial feminization of males, i.e., only gonads were sex-reversed, is most likely due to the transient nature of the eRNAi. Nevertheless, once *Mdmd* expression is restored at late embryonic stages, it appears to repress zygotic *Md-tra* during the larval/pupal transition stage, and male development ensues. The results of quantitative *Mdmd* expression analysis further strengthen this argumentation (chapter 4).

It was also tested whether suppression of *Mdmd* by eRNAi affects the female-specific splicing of *Md-tra*. After *Mdmd* knock-down splicing of *Md-tra* transcripts was irreversibly shifted to the female mode, *Md-tra*^{F}, in the ovaries of the sex-reversed males. These results suggest that *Mdmd* acts as an upstream repressor of female splicing of *Md-tra*. Based on these findings, it was concluded that *Mdmd* is required for differentiation of male gonads at the syncytial embryonic stage. To permanently disrupt *Mdmd* function during all development stages, a CRISPR/Cas9 genome editing approach can be applied.

In chapter 5, we presented two possible models for how *Mdmd* acquired its male-determining function after duplication from *Md-ncm*. In model I, we hypothesized that *Mdmd* acquired a novel function unrelated to *Md-ncm* which specifically inhibits the autoregulatory loop of zygotic *Md-tra*. In model II, we hypothesized that *Mdmd* acts as a dominant negative regulator of *Md-ncm* and interferes with those functions of *Md-ncm* that are necessary for promoting female splicing of *Md-tra*.

There is a limited understanding of the molecular nature and mechanisms of male-determining genes in insects. Only three male-determining genes *Nix*, *Yob* (both in mosquitoes, Hall *et al.*, 2015; Krzywinska *et al.*, 2016) and *Mdmd* (in *M. domestica*, Sharma *et al.*, 2017) have been identified to date. Interestingly, these three male-determiners have no sequence similarity with each other, illustrating the potentially enormous diversity of primary sex determination signals. Further identification of male-determining and female-determining genes in other insect species will broaden our knowledge about regulation and evolution of sex determination pathways.

In conclusion, our study reveals how an existing gene can arise by duplication and acquire a sex determination role by neofunctionalization. It shows how heterogeneity of sex determination can be present within an insect species and the genetic program responsible for development of males and females can rapidly change in the course of evolution. Overall, this knowledge not only helps us better understand the evolutionary biology of sex determination but it is also instructive for developing novel, sustainable strategies to control agricultural pests or vectors of diseases, like fruit flies and mosquitoes.

Dutch summary

In organismen die zich geslachtelijk voortplanten, is het correct totstandkomen van de geslachtelijke identiteit een essentieel proces. In insecten bestaat er een breed scala aan mechanismen die ten grondslag liggen aan deze geslachtsbepaling; er komen verschillende chromosomale systemen en verschillende geslachtsbepalende genen voor. Het geslacht wordt bepaald door een cascade van genen met een hiërarchische organisatie. Bovenaan (aan het begin van) deze cascade bevindt zich een primair geslachtsbepalend signaal, gevolgd door een "aan/uit" schakelaar die het primaire signaal interpreteert en verder leidt naar een globale effector onderaan (aan het einde van) de cascade. De geslachtsbepalende cascade wordt in het vroeg-embryonaal stadium geactiveerd, en loopt via geslachtsspecifieke splicing van het pre-mRNA van de "aan/uit" schakelaar *transformer* (*tra*) naar *doublesex* (*dsx*) aan het einde van deze cascade. Afhankelijk van het primaire signaalwordt de ontwikkeling van het embryo door *dsx* ofwel als mannetje ofwel als vrouwtje ingezet. De primaire genetische signalen in de cascade zijn zeer divers, terwijl de genen onderaan de cascade juist geconserveerd zijn (Gembe en Beye, 2011; Bopp *et al.,* 2014). Volgens de bottom-up hypothese, voorgesteld door Wilkins (1995), evolueren deze cascades van onder naar boven, waarbij er achtereenvolgens steeds nieuwe genen bovenaan de cascade worden opgenomen.

De huisvlieg, *Musca domestica*, is een opmerkelijk voorbeeld van de grote diversiteit van geslachtsbepalende mechanismen en is daarom een ideaal modelorganisme om de evolutionaire diversificatie van geslachtsbepalingscascades te bestuderen (Dübendorfer et al., 2002). Het polymorfe geslachtsbepalingssysteem van huisvliegen is al meer dan vijftig jaar bekend, en is uniek in het feit dat er verschillende natuurlijke populaties bestaan waar een dominante genetische factor die de mannelijke ontwikkeling veroorzaakt (de *M*-factor) op verschillende chromosomen voor kan komen (op het Y-chromosoom, of op een van de vijf autosomen) (Hiroyoshi, 1964; Franco *et al.*, 1982; Denholm *et al.*, 1983). De vraag of deze *M*-factoren verschillende genen zijn of kopieën van hetzelfde gen dat een translocatie heeft ondergaan naar verschillende posities in het genoom, is tot nu toe nog niet beantwoord.

De *M*-factor is het primaire signaal in de geslachtsbepalingscascade en verhindert de maternale activatie van de *transformer* auto-regulatie (Inoue en Hiroyoshi, 1986; Hilfiker-Kleiner *et al.*, 1993). De *M*-factor zorgt ervoor dat het pre-mRNA van *Md-tra* in een niet-functionele variant wordt gespliced, wat vervolgens leidt tot de splicing van het pre-mRNA van *Md-dsx* in de mannelijke variant (Hediger *et al.,* 2010). Het doel van mijn onderzoek was om de *M*-factor(en) in de huisvlieg te identificeren en te karakteriseren. Daarnaast heb ik mij gericht op de belangrijke evolutionaire vraag hoe *M*-factoren kunnen evolueren op verschillende plaatsen in het genoom.

Om de *M*-factor te identificeren werd aangenomen dat *M* alleen aanwezig is in mannetjes, en dat het reeds in het pre-blastodermale stadium tot expressie komt en zorgt

voor een overgang van zygotisch *Md-tra* naar de mannelijke isoformen (Hediger *et al.,* 2010). Op basis van deze aannames is een differentiële expressieanalyse uitgevoerd tussen RNAseq-data van uitsluitend mannelijke of vrouwelijke embryo's uit het syncytium stadium. Dit leidde tot de identificatie van een aantal mRNA's die alleen in mannelijke embryo's aanwezig waren. Deze zijn vervolgens vergeleken met vrouwelijke *M. domestica*-genoomsequenties (Scott *et al.,* 2014), waarbij een aantal sequenties werd gevonden, die aanwezig zijn in mannetjes, maar niet in vrouwtjes. Tussen deze sequenties, hebben we vijf man-specifieke contigs gevonden waarvan de sequentie in grote mate bleek overeen te komen met die van het gen *CWC22*/*nucampholin*, dat codeert voor een pre-mRNA splicing factor. Vervolgens zijn met verschillende primerparen de overlappende delen van deze vijf man-specifieke contigs geamplificeerd en daarmee is bevestigd dat deze deel uitmaken van één transcriptionele eenheid. Dit betekent dat de vermoedelijke initiator van de mannelijke ontwikkeling in de huisvlieg is geïdentificeerd. Dit gen is *Mdmd* genoemd (afkorting van *Musca domestica male determiner*) (hoofdstuk 2).

Op basis van een hoge mate van overeenkomst in de sequentie van de twee genen, is geconcludeerd dat *Mdmd* een paraloog is van *CWC22* (Sharma *et al.,* 2017). CWC22 heeft een algemene functie in het splicen van pre-mRNA, en is vereist voor de formatie van het exon junction complex (EJC) (Barbosa *et al.,* 2012; Steckelberg *et al.,* 2012; Alexandrov *et al.,* 2012). Een BLASTN-zoekopdracht van *Mdmd*-sequenties in het vrouwelijke *M. domestica*-genoom leidde tot de identificatie van de paraloog *Md-ncm* (afkorting van *Musca domestica nucampholin*), een ortholoog van *ncm* in *Drosophila melanogaster*. De hoge mate van overeenstemming tussen de sequenties van *Mdmd* en *Md-ncm* is een aanwijzing dat *Mdmd* is ontstaan door een genduplicatie. Het eiwit MDMD bevat twee domeinen, MIF4G en MA3, die geconserveerd zijn binnen de Metazoa; en deze zijn in meerdere soorten betrokken bij het bewaken van de integriteit van het mRNA, translatie, en *non-sense* decay (NMD) (Alexandrov *et al.,* 2012).

Mdmd is als eerste geïdentificeerd in de M^{IIII}-stam van *M. domestica*, waar de *M*-factor op het derde chromosoom ligt. De eerste vraag die vervolgens werd gesteld was of *Mdmd* ook aanwezig is op het Y-chromosoom, en op een van de andere autosomen. De vondst van *Mdmd* in verschillende *M*-stammen (M^{II}, M^{IIII}, M^{V} en M^{Y}), en de grote mate van overeenkomst in de sequentie van deze kopieën, suggereert dat *Mdmd* hetzelfde gen is in deze verschillende stammen. Naar verwachting bleek de spliceosoomfactor *Md-ncm*, in tegenstelling tot *Mdmd*, aanwezig te zijn in de genomen van zowel mannetjes als vrouwtjes van de verschillende *M*-stammen. In een fylogenetische analyse clusterden de *Mdmd*-versies van de verschillende *M*-stammen sterk samen, wat aangeeft dat de Y-gebonden en autosomale *M*-factoren een gemeenschappelijke oorsprong hebben, en door een translocatie op verschillende plekken in het *M. domestica*-genoom terecht zijn gekomen (hoofdstuk 3). Een fylogenetische analyse van de MDMD-, MD-NCM-, en NCM-eiwitsequenties in

verschillende metazoa fyla geven inzicht in hoe *Mdmd* van *Md-ncm* is gedivergeerd, terwijl *ncm* het patroon van de geaccepteerde verwantschappen binnen de Metazoa volgt. In nauw verwante vliegensoorten zijn er geen paralogenvan *ncm* gevonden, hetgeen suggereert dat de opname van *Mdmd* in de geslachtsbepalende cascade uniek is voor de huisvlieg (hoofdstuk 3). Een andere interessante vondst is dat *Mdmd* aanwezig is in meerdere kopieën waarvan een aantal functioneel en andere niet-functioneel lijken te zijnen dat dit verschilt voor*M*-loci van verschillende *M. domestica*-stammen (Sharma *et al.,* 2017). Deze kopieën van *Mdmd* zijn wellicht ontstaan doordat ze voor of na translocatie lokaal meerdere malen zijn gedupliceerd op het Y-chromosoom of op de verschillende autosomen, om de functionaliteit van *Mdmd* te behouden (hoofdstuk 5).

Om de functie van *Mdmd* als bepalende factor voor mannelijke ontwikkeling te toetsen, werd embryonale RNAi (eRNAi) uitgevoerd op *Mdmd* in vroege embryo's van verschillende *M*-stammen. Blokkade van *Mdmd* door middel van eRNAi resulteerde in een omkering van het geslacht: mannetjes vertoonden vrouwelijke ontwikkeling met volledig gedifferentieerde ovaria. Hiermee kon bevestigd worden dat *Mdmd*-activiteit in een vroeg embryonaal stadium noodzakelijk is voor de differentiatie van mannelijke gonaden in de M^{III}-, M^{IIII}-, M^{V}- *en* M^{Y}-stammen (hoofdstuk 4). In M^{I}-mannetjes werd echter geen geslachtsomkering waargenomen, hetgeen suggereert dat deze stam een ander *M*-locus heeft. Daarnaast is getoetst of de suppressie van *Mdmd* door eRNAi invloed had op de vrouw-specifieke *splicing* van *Md-tra*. Na een knock-down van *Mdmd* bleken de *Md-tra*-transcripten in de ovaria van de mannetjes met omgekeerd geslacht, alleen in de vrouwspecifieke splicevariant (*Md-tra*^{F}) voor te komen, en deze verandering leek onomkeerbaar. Dit suggereert dat *Mdmd* de vrouw-specifieke splicing van *Md-tra* onderdrukt. Op basis van deze resultaten konden worden geconcludeerd dat *Mdmd* vereist is voor de differentiatie van mannelijke gonaden in het syncytiale embryonale stadium. Om in de toekomst de functie van *Mdmd* in alle stadia van de ontwikkeling te onderdrukken, zou CRISPR/Cas9 toegepast kunnen worden.

In hoofdstuk 5 worden twee modellen geopperd om te beschrijven hoe *Mdmd* een functie heeft gekregen in het bepalen van mannelijke ontwikkeling na duplicatie van *Md-ncm*. De hypothese in model I is dat *Mdmd* een nieuwe functie heeft gekregen die geen verband houdt met de functie van *Md-ncm*, waarbij *Mdmd* specifiek de auto-regulatie van zygotisch *Md-tra* inhibeert. De hypothese in model II is dat *Mdmd* een dominante negatieve regulator van *Md-ncm* is, en dat *Mdmd* interfereert met de functies van *Md-ncm* die noodzakelijk zijn voor de vrouw-specifieke splicing van *Md-tra*.

De moleculaire mechanismen en genen die leiden tot mannelijke ontwikkeling in insecten zijn onvoldoende bekend. Er zijn op het moment slechts drie genen beschreven die leiden tot mannelijke ontwikkeling: *Nix*, *Yob* (beide in steekmuggen, Hall *et al.,* 2015;

Krzywinska *et al.,* 2016), en *Mdmd* (in *M. domestica*, Sharma *et al.,* 2017). Deze drie genen vertonen onderling geen overeenkomst in sequentie, wat laat zien dat de diversiteit aan primaire geslachtsbepalende signalen mogelijk zeer groot is. Verdere identificatie van genen die leiden tot mannelijke en vrouwelijke ontwikkeling in andere insectensoorten, zou het inzicht in de regulatie en evolutie van geslachtsbepaling aanzienlijk kunnen verbreden.

Tot slot laatdit onderzoek zien hoe een gen dat door duplicatie ontstaat een geheel nieuwe functie kan krijgen in het bepalen van geslachtsspecifieke ontwikkeling. Daarnaast laat het zien hoe heterogeen geslachtsbepaling kan zijn binnen een insectensoort, en dat het genetische mechanisme dat leidt tot deze geslachtsbepaling snel kan evolueren. Deze opgedane kennis leidt niet alleen tot een breder begrip van de evolutie van geslachtsbepaling, maar kan ook bijdragen aan de ontwikkeling van nieuwe, duurzame bestrijdingssystemen tegen insecten die schadelijk zijn voor de landbouw, of die ziekten overdragen, zoals fruitvliegen en steekmuggen.

ACKNOWLEDGEMENTS

It's time to bring to an end! Many people have contributed to this beautiful journey in one way or the other, and now I would like to take this moment to thank every one of them.

First and foremost, I extend my warm gratitude to my promoter Prof. Leo W. Beukeboom, your support, motivation, guidance and encouragements were conducive towards the completion of this thesis. I admire your passion and interest in science and research. I respect your caring, friendly and easygoing attitude towards your students. Thank you very much for providing me with an opportunity to work as your student.

I am deeply grateful to my co-promotor Dr. Daniel Bopp for giving me the opportunity to carry out my PhD work in his lab at the University of Zurich. I have always appreciated your views towards the subjects. Indeed I learned a lot from you. I would like to thank my other co-promoter Dr. Louis van de Zande for all fruitful discussions that we made.

I would like to acknowledge Prof. Mark Robinson for his very enthusiastic efforts for *Musca* project. Your support and prompt replies to my e-mails and queries were beneficial. I feel honoured and fortunate to have had a chance to work with you. Prof. Ernst Wimmer, your suggestions during the project was always of great help and compelled me to think differently. I especially appreciate your arrangements for the half early Gottingen brain stormy meetings.

I would also like to thank the members of the assessment committee, Prof. L. Rabinow, Prof. L. Sánchez and Prof. I. R. Pen for taking the time to read my thesis.

My sincere thanks to Andreas to whom I lived in Bassersdorf, Zurich for a long time. You taught me a lot and cared like a guardian. The moments I spent with you will always be precious to me. Many times I revised those days so that I could restore those memories throughout my life. Thanks for teaching me the basics of many things (especially skiing and tracking). I will always remember the world that you showed to me. I consider myself very fortunate to have you as a part of my life. Thanks for everything!

Big thanks to my colleague and good friend Claudia. It was a great time with you in Zurich. I appreciate your help and support during work and providing an amicable working environment. I admire your super friendly nature and "always ready to help other's" attitude. I miss your company. Leonidas, it was a pleasant time with you in Zurich and thank you for managed a trip to India with my family.

Beate, Sandra and Yvonne, thank you very much for a home feeling during all of my Swiss time. It was a great memorable time with all of you. I enjoyed a lot.

Linda and Sjors, Many thanks, for your lovely company on several occasions.

Thank you, Christine, to invite me twice to organise the Zumikerlauf and sharing some beautiful memories with your family.

Monika, thanks for your help and discussions on *Musca* project. It was an honour to meet a person like you.

Raquel, thank you very much for the delightful time. I learned a lot from you about the happiness of Life.

I would also like to thank number of students to whom I worked; Jan Otto, Christoph, Dominik, Nikola, Svenia, Nicolas, Tamara, Salome, Seraina, Tea, Michael, Arne, and Femke.

Thank you, Kirsten, for advising me to rent the place of your brother when I first arrived in Groningen. Roelof, Dirk Jan and Mark thank you for sharing the apartment. You guys helped me a lot during the first three months in Groningen before moving to Zurich. I can remember that with your help within one hour after reaching to the Groningen, I bought a bike from the shop and after that my Groningen life becomes more comfortable.

Ljubinka, thank you for being a paranymph. We spent around one year (6+6 months with coffee and laugh) in Groningen, and both time luckily we shared the same office (since you left Groningen I missed those moments sometimes). It was a great time to work with you (without any expectations to each other). I hope that in the near future we will do more collaborative research!

Thank you, Raj for preparing Indian food (when I have no place to cook in Groningen) and helping to find an apartment in Groningen during the wandering time of my PhD. Also thanks to Shubham, Siddhartha, Ketan, Dhwani and Hemant for having a friendly company. Thanks for all cooking, fun, travels and discussions over the years. Outside the lab, it was a memorable time for festival celebrations and dinners. I truly appreciate your time and efforts to my parents when they visited me in Groningen.

Big thanks to my Zurich company (Sunil, Vibhu, Jimit, Preetam, Kapil, and Arnamoy). It was a fabulous time with all of you, especially weekend party and many road trips to different places in Europe.

Thank you, Remo, Daniela, Anna, Daniel (junior) and Hanna for a delightful time and chat at IMLS, Zurich. Jacqueline, thanks for taking care of my questions regarding visa and residence permit in Switzerland.

Thank you WIRE for organising many international student trips in Netherlands. It was a fantastic time.

Rogier and Peter, I appreciate your help for the housefly culture in Groningen.

Anna, Many thanks for suggesting several group activities and taking care of the promotion arrangements. Elzemiek, Thank you, for organising many kinds of scientific activities (especially writing camp at Schiermonnikoog).

I would like to pass my sincere thanks to Peter for translating the thesis summary in Dutch.

Thanks, Xuan, Yuan and Fangying for all the (naive) jokes and versatile discussions that we had together during my last year in Groningen. Fangying, thanks for being my paranymph.

There are also so many people whose names are not mentioned. Big thanks for your help in many ways during the past years like finding and sharing an apartment, visa stuffs, short out the local things etc.

I would like to extend my gratitude to all of the Alumni and lab mates for scientific discussions, fun and parties (Borrels) over the years in Groningen and Zurich. I am glad to have a company like you. I am thankful to all to make my time memorable. As an Alumnus, I would like to connect with all of you shortly.

I would like to express my most profound gratitude to my family. I would not have reached this far without your constant support, love, trust, affection and motivation. Whatever I am today is because of you and for you. At last, I would like to dedicate my thesis to the thousands of houseflies that sacrificed for *Science*.

आकाश